SIAM J. APPL. MATH.
Vol. 67, No. 1, pp. 1–23

PERFECTLY MATCHED LAYERS FOR HYPERBOLIC SYSTEMS: GENERAL FORMULATION, WELL-POSEDNESS, AND STABILITY*

DANIEL APPELÖ†, THOMAS HAGSTROM‡, AND GUNILLA KREISS†

Abstract. Since its introduction the perfectly matched layer (PML) has proven to be an accurate and robust method for domain truncation in computational electromagnetics. However, the mathematical analysis of PMLs has been limited to special cases. In particular, the basic question of whether or not a stable PML exists for arbitrary wave propagation problems remains unanswered. In this work we develop general tools for constructing PMLs for first order hyperbolic systems. We present a model with many parameters, which is applicable to all hyperbolic systems and which we prove is well-posed and perfectly matched. We also introduce an automatic method for analyzing the stability of the model and establishing energy inequalities. We illustrate our techniques with applications to Maxwell's equations, the linearized Euler equations, and arbitrary 2×2 systems in $(2 + 1)$ dimensions.

Key words. perfectly matched layers, stability

AMS subject classifications. 35L45, 35B35

DOI. 10.1137/050639107

1. Introduction.

Many important wave propagation problems are posed on unbounded or large domains. Such problems must be solved on a truncated domain if numerical methods are to be used. There exist many techniques for truncating the original domain (see the review papers [15, 16, 23]), but one that has proved both efficient and accurate is the perfectly matched layer (PML) technique. The PML technique surrounds the domain where the solution is desired (the computational domain) by an artificial layer. The layer is constructed so that waves traveling across the interface between the layer and the computational domain are not reflected; that is, the layer is perfectly matched. Moreover, the layer is constructed so that, inside the layer, the solution decays exponentially in the direction normal to the interface. Hence, if the layer is sufficiently wide, the solution will be close to zero at the outer boundary, and therefore any stable boundary condition can be used there.

Besides the perfect matching and damping properties of the layer it is also desirable that the equations governing the PML be well-posed. This is especially important if a PML derived for a linear problem is to be applied to a nonlinear problem or a problem with variable coefficients. If the linearized problem is only weakly well-posed the corresponding nonlinear or variable coefficient problem can be ill-posed; see [20]. Well-posedness, by definition, allows the solution to grow exponentially in time, and therefore, for a PML to be practically useful, it must also be stable (in time). To

*Received by the editors August 29, 2005; accepted for publication (in revised form) June 16, 2006; published electronically November 3, 2006. Any conclusions or recommendations expressed in this paper are those of the author, and do not necessarily reflect the views of NASA or LLNL.
http://www.siam.org/journals/siap/67-1/63910.html

†Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, S-100 44, Sweden (appelo@nada.kth.se, gunillak@nada.kth.se). The first author was supported in part by the Swedish research council grant VR2004-2371, NASA contract NAG3-2692 and Lawrence Livermore National Laboratory. The third author was supported in part by the Swedish research council grant VR2004-2371.

‡Department of Mathematics and Statistics, The University of New Mexico, Albuquerque, NM 87131 (hagstrom@math.unm.edu). The author was supported in part by NSF grant DMS-0306285, ARO grant DAAD19-0301-0146, NASA contract NAG3-2692, and Lawrence Livermore National Laboratory subcontract B547968.

summarize, the key properties of a PML are perfect matching, well-posedness, and stability.

PMLs were originally introduced for Maxwell's equations by Bérenger [8]. Well-posedness and stability of the Bérenger PML has been the topic of numerous works. For example, Abarbanel and Gottlieb [2] showed that Bérenger's "split-field" PML was only weakly well-posed and that it supported linearly growing modes. Similar results were also obtained via Fourier and energy techniques by Bécache and Joly in [6]. The issue of weak well-posedness led to the development of various well-posed "physical" or "un-split" PMLs for Maxwell's equations; see [3, 13, 24]. These "un-split" PMLs were further improved by the inclusion of the so-called complex frequency shift (CFS), which has been used by Bécache, Petropoulos, and Gedney [7] to remove late-time linear growth.

For other applications such as the linearized Euler equations [18], the linearized shallow water equations [22], and anisotropic elasticity [10], there have been reports of exponentially growing solutions. In [1] Abarbanel, Gottlieb, and Hesthaven found that a stable PML could be derived for the linearized Euler equations by transforming the equations into a system whose dispersion relation resembled the dispersion relation of Maxwell's equations. The same transform was later used again to develop a stable PML for the linearized Euler equations [19, 11] and for the linearized shallow water equations [22].

Today there exist stable PML models for many important problems, but there are also problems, e.g., anisotropic elasticity and linearized Magneto Hydro Dynamics (MHD), for which stable PMLs have not yet been found. An open issue, then, is whether stable PMLs can be constructed in general. Also, stability and well-posedness for general hyperbolic systems has received less attention than particular cases. One exception is the paper [5] where Bécache, Fauqueux, and Joly give necessary conditions for stability of the split-field PML in terms of the geometrical properties of the dispersion relation. Also, in [4] we construct stable PMLs for arbitrary 2×2 symmetric hyperbolic systems in $(2 + 1)$ dimensions.

In this work we generalize the formulation of PML models for hyperbolic systems introduced in [17]. To make the model suitable for future applications, we introduce a very general formulation including many free parameters. One of these parameters adds a parabolic term in the tangential directions. By including this parameter, we can show that the equations of the PML are well-posed as long as the original hyperbolic system is well-posed. In addition, we give a proof that the layer is perfectly matched.

We also study the stability of our PML model. The question of stability is not trivial, and in general it has to be investigated separately for each new application. To simplify these investigations we introduce a technique, based on criteria for the number of zeros of a polynomial in a half-plane, that can be used to derive necessary and sufficient conditions for stability of any first order constant coefficient Cauchy problem. Moreover, if these conditions are fulfilled, there is also a local energy density that decays with time (see [14]). This energy density is automatically generated from the necessary and sufficient conditions. We use the technique to derive stability results for three interesting applications of our general model.

The rest of this paper will be organized as follows. In section 2 we present the general PML model for symmetric hyperbolic systems and show that it is perfectly matched and well-posed. In section 3 we introduce techniques from [14] used to determine the stability of a first order system with constant coefficients. If the system is stable, the technique will yield an energy with a local density that decays with time. In section 4 we analyze the stability of a PML model for Maxwell's equation in

two dimensions. The PML is constructed by using the general PML model described in section 2. We use the techniques from section 3 to establish the stability of the PML and list two associated energies. In section 5 we analyze a PML model for the linearized Euler equations and show that it is stable. In section 6 we consider the specialization of our general PML formulation to 2×2 symmetric hyperbolic systems in $(2+1)$ dimensions. In [4] we demonstrated how to choose the layer parameters as functions of the coefficient matrices. Here we prove that these choices will lead to a stable PML. In section 7 we conclude and discuss some possible extensions of the presented work.

2. A general PML. We consider the symmetric hyperbolic system in d dimensions:

$$\frac{\partial u}{\partial t} + A_x \frac{\partial u}{\partial x} + \sum_{l=1}^{d-1} A_{y_l} \frac{\partial u}{\partial y_l} + Cu = 0, \tag{1}$$

with initial data, u_0, supported in $-H < x < -h$, $h > 0$. Here $A_x = A_x^T$ and $A_{y_l} = A_{y_l}^T$. For simplicity we assume that A_x is invertible; if A_x is singular, we apply the PML only to the equations involving x derivatives.

Our construction of the layer equations matched to (1) follows the ideas suggested in [17]. It is based on a modification of the eigenvalues of the eigenvalue problem (equation (9) in the next section) obtained after Fourier and Laplace transformation of (1). Then one can consider a general transformation of the eigenvalues which is rationally dependent on the transform parameters, where the restriction to rationally dependent transformations leads to localizable layer equations. Considering a fairly general transformation, we are led to consider the following general PML model:

$$\frac{\partial u}{\partial t} + A_x \left((1+\sigma\eta)\frac{\partial u}{\partial x} + \sigma \left(\sum_{l=1}^{d-1} \xi_l \frac{\partial u}{\partial y_l} + \mu u \right) + \sum_j \phi_j \right) + \sum_{l=1}^{d-1} A_{y_l} \frac{\partial u}{\partial y_l} + Cu = 0, \tag{2}$$

$$\frac{\partial \phi_j}{\partial t} + \sigma\phi_j + \alpha_j \phi_j + \sum_{l=1}^{d-1} \beta_{jl} \frac{\partial \phi_j}{\partial y_l} - \sum_{l=1}^{d-1} \varepsilon_{jl} \frac{\partial^2 \phi_j}{\partial y_l^2} = \sigma \left(\gamma_j \frac{\partial u}{\partial x} + \sum_{l=1}^{d-1} \delta_{jl} \frac{\partial u}{\partial y_l} + \nu_j u \right). \tag{3}$$

Here all the additional parameters are real, and we also assume

$$1 + \sigma\eta > 0, \qquad \varepsilon_{jl} \geq 0. \tag{4}$$

To obtain spatial decay of waves propagating through the layer it is necessary that the real parts of the modified eigenvalues be bounded away from zero. As yet we have no general method for constructing stable layers, and it is conceivable that problems exist which require many more parameters than have been required in the examples treated so far. Thus we will analyze (2)–(3) in its full complexity when feasible. The effect of many of the parameters is not yet understood, but we know from the example in section 6 that acceptable parameter values depend on the matrices in system (1).

2.1. Perfect matching. To investigate the perfect matching of the layer we consider two problems. In the first, whose solution we denote u_1, (1) holds in $R^d \times R$, and in the second, whose solution is denoted u_2, we suppose that (1) holds in $x < 0$ and that (2) and (3) hold in $x > 0$. We also insist that u_2 be continuous. Our goal is to show that the restrictions of each solution to $x < 0$ are identical; that is, the layer is perfectly matched.

We begin by performing a Fourier–Laplace transformation in the tangential directions and in time. The duals of $y = [y_1, \ldots, y_{d-1}]$ are denoted by $k = [k_1, \ldots, k_{d-1}]$ and the dual of t by s. This leads to the problems

$$A_x \frac{\partial \hat{u}_1}{\partial x} + \left(sI + \sum_l ik_l A_{y_l} + C \right) \hat{u}_1 = \hat{u}_0, \quad x \in R, \tag{5}$$

and in the second case, for $x < 0$,

$$A_x \frac{\partial \hat{u}_2^L}{\partial x} + \left(sI + \sum_l ik_l A_{y_l} + C \right) \hat{u}_2^L = \hat{u}_0, \tag{6}$$

and for $x > 0$,

$$\begin{aligned} A_x &\left((1 + \sigma\eta) \frac{\partial \hat{u}_2^R}{\partial x} + \sigma \left(\sum_l ik_l \xi_l + \mu \right) \hat{u}_2^R + \sum_j \hat{\phi}_j \right) \\ &+ \left(sI + \sum_l ik_l A_{y_l} + C \right) \hat{u}_2^R = 0, \end{aligned} \tag{7}$$

$$\left(s + \sigma + \alpha_j + \sum_l ik_l \beta_{jl} + \sum_l \varepsilon_{jl} k_l^2 \right) \hat{\phi}_j = \sigma \left(\gamma_j \frac{\partial \hat{u}_2^R}{\partial x} + \left(\sum_l ik_l \delta_{jl} + \nu_j \right) \hat{u}_2^R \right). \tag{8}$$

The solution of (5) follows from the solution of the eigenvalue problem

$$\lambda A_x w + \left(sI + \sum_l ik_l A_{y_l} + C \right) w = 0. \tag{9}$$

We note that for $\Re s > |C|_2$ the eigenvalues, λ, cannot be purely imaginary. In particular if we normalize w to have length one, a straightforward computation yields

$$\Re\lambda = -\frac{\Re s + \Re w^* C w}{w^* A_x w}, \tag{10}$$

which implies

$$|\Re\lambda| > (\rho(A_x))^{-1} (\Re s - |C|_2). \tag{11}$$

Thus, taking $\Re s$ sufficiently large, we may assume that solutions of (9) fall into two sets labeled by the sign of the real parts of the eigenvalues:

$$\Re\lambda_1, \ldots, \Re\lambda_r < 0, \tag{12}$$

$$\Re\lambda_{r+1}, \ldots, \Re\lambda_n > 0. \tag{13}$$

Moreover, the matrix

$$M(s,k) = -A_x^{-1} \left(sI + \sum_l ik_l A_{y_l} + C \right), \tag{14}$$

can be block diagonalized:

$$QMQ^{-1} = \begin{pmatrix} S^- & 0 \\ 0 & S^+ \end{pmatrix}, \tag{15}$$

where the eigenvalues (12) are the eigenvalues of S^- and the eigenvalues (13) are the eigenvalues of S^+. Now the bounded solution of (5) is easy to write down as

$$\hat{u}_1 = Q^{-1} \begin{pmatrix} \int_{-\infty}^{x} e^{S^-(x-y)} f^-(y)dy \\ -\int_{x}^{\infty} e^{S^+(x-y)} f^+(y)dy \end{pmatrix}, \tag{16}$$

where

$$QA_x^{-1}\hat{u}_0 = \begin{pmatrix} f^- \\ f^+ \end{pmatrix}. \tag{17}$$

In particular the support properties of $\hat{u}_0$ and thus $f^{\pm}$ guarantee the existence of the integrals in (16). We note that at $x = 0$,

$$\hat{u}_1 = Q^{-1} \begin{pmatrix} \int_{-H}^{-h} e^{-S^- y} f^-(y)dy \\ 0 \end{pmatrix}. \tag{18}$$

We now compute $\hat{u}_2$ in each region. We first note that (8) can be solved directly:

$$\hat{\phi}_j = \frac{\sigma\left(\gamma_j \frac{\partial \hat{u}_2^R}{\partial x} + (\sum_l ik_l\delta_{jl} + \nu_j)\hat{u}_2^R\right)}{s + \sigma + \alpha_j + \sum_l ik_l\beta_{jl} + \sum_l \varepsilon_{jl}k_l^2}. \tag{19}$$

Now for $x > 0$ we transform the solution using the same transformation Q, which block diagonalizes the problem for $x < 0$. Setting $v = Q\hat{u}_2^R$, we find

$$v_x = \frac{1}{r(s,k) + \sigma p(s,k)} \begin{pmatrix} r(s,k)S^- - \sigma q(s,k)I & 0 \\ 0 & r(s,k)S^+ - \sigma q(s,k)I \end{pmatrix} v, \tag{20}$$

where the polynomials r, p, and q are defined up to a constant multiple by

$$\eta + \sum_j \frac{\gamma_j}{s + \sigma + \alpha_j + \sum_l ik_l\beta_{jl} + \sum_l \varepsilon_{jl}k_l^2} = \frac{p(s,k)}{r(s,k)}, \tag{21}$$

$$\sum_l ik_l\xi_l + \mu + \sum_j \frac{\sum_l ik_l\delta_{jl} + \nu_j}{s + \sigma + \alpha_j + \sum_l ik_l\beta_{jl} + \sum_l k_l^2\varepsilon_{jl}} = \frac{q(s,k)}{r(s,k)}. \tag{22}$$

We will argue that for $\Re s$ sufficiently large these blocks have eigenvalues with negative and positive real parts, respectively. In particular we note that

$$\lim_{|s|\to\infty} \frac{p}{r} = \eta, \qquad \lim_{|s|\to\infty} \frac{q}{r} = \sum_l ik_l\xi_l + \mu. \tag{23}$$

Thus for large s the eigenvalues are approximately

$$\frac{\lambda_j - \sigma(\sum_l ik_l\xi_l + \mu)}{1 + \sigma\eta}. \tag{24}$$

Now by (11) and (4) we conclude that the signs of their real parts are the same as the signs of $\Re\lambda_j$ if we choose $\Re s$ sufficiently large, which was what we wished to prove.

From this argument we conclude that the transform of the causal solution in $x > 0$ takes the form

$$\hat{u}_2^R = Q^{-1} \begin{pmatrix} e^{(r+\sigma p)^{-1}(rS^- - \sigma q I)x} v^- \\ 0 \end{pmatrix}. \tag{25}$$

We see that this can be "perfectly matched" to the restriction of $\hat{u}_1$ to $x < 0$ by setting

$$v^- = \int_{-H}^{-h} e^{-S^- y} f^-(y) dy. \tag{26}$$

Thus we have proven that u_1 and u_2 restricted to $x < 0$ are identical.

We note that we can interpret the layer as an (s, k)-dependent change of variables:

$$\hat{u}(x) \to e^{-a\tilde{x}} \hat{u}(\tilde{x}), \tag{27}$$

where

$$\tilde{x} = \frac{r}{r + \sigma p} x, \qquad a = \sigma \frac{q}{r}. \tag{28}$$

With this interpretation, the new layer can be viewed as a generalization of the Bérenger layer from the viewpoint of complex coordinate stretching, as introduced by Chew and Weedon [9].

2.2. Well-posedness of the layer equations. For the applications considered in this paper it will be sufficient to include only one set of auxiliary variables, leading to the PML model

$$\begin{aligned} &\frac{\partial u}{\partial t} + A_x \left((1 + \sigma\eta) \frac{\partial u}{\partial x} + \sigma \left(\sum \xi_l \frac{\partial u}{\partial y_l} + \mu u \right) + \phi \right) + \sum A_{y_l} \frac{\partial u}{\partial y_l} + Cu = 0, \\ &\frac{\partial \phi}{\partial t} + \sigma\phi + \alpha\phi + \sum \beta_l \frac{\partial \phi}{\partial y_l} - \sum \varepsilon_l \frac{\partial^2 \phi}{\partial y_l^2} = \sigma \left(\gamma \frac{\partial u}{\partial x} + \sum \delta_l \frac{\partial u}{\partial y_l} + \nu u \right). \end{aligned} \tag{29}$$

Even with just one set of auxiliary variables, there are many free parameters that must be chosen. Our experience is that the parameters $\mu, \xi, \beta_l, \delta_l, \gamma$ can be determined from the coefficients of the matrices A_x and A_y. The parameter η is introduced in the model to increase the damping of evanescent modes. The parameter α, which is usually referred to as the CFS, typically enhances stability properties at late time.

To our knowledge the parabolic terms $\varepsilon_l \phi_{y_l y_l}$ (hereafter called parabolic CFS) have not been included in PML models before. We have chosen to include them to guarantee the well-posedness of the model (29). To see this we freeze the coefficients and perform a Fourier transform in space (k_x is the dual of x). Excluding the zero order terms in the symbol of the equations (29), we obtain

$$P_1(ik) = - \begin{bmatrix} (ik_x(1 + \sigma\eta) + \sum ik_l \xi_l \sigma) A_x + \sum ik_l A_{y_l} & 0 \\ -(ik_x \sigma\gamma + \sum ik_l \delta_l \sigma) I & \sum \varepsilon_l k_l^2 I + \sum ik_l \beta_l I \end{bmatrix}. \tag{30}$$

Denote the upper diagonal block in P_1, (30), by P_{11}. By the hyperbolicity of the original problem, P_{11} is diagonalizable with imaginary eigenvalues.

Without the parabolic complex frequency shift the lower diagonal block also has purely imaginary eigenvalues, but the system may be only weakly hyperbolic. This is the case if, for some set of $k_1, \ldots, k_{d-1}$,

$$\sum_l i k_l \beta_l \tag{31}$$

coincides with one of P_1's eigenvalues while

$$k_x \sigma \gamma + \sum k_l \delta_l \sigma \neq 0. \tag{32}$$

Then it is not possible to diagonalize P_1, and the problem is not well-posed. Otherwise the system is strongly hyperbolic and thus well-posed.

If all $\varepsilon_l \neq 0$, the lower diagonal block always has eigenvalues that are distinct from the eigenvalues of P_{11}, as shown by the following argument. When at least one k_l is nonzero, the eigenvalue of the lower block has negative real part. For $k_1 = \cdots = k_{d-1} = 0$, P_{11} is nonsingular since A_x is nonsingular, while the lower diagonal block is zero. It follows that P_1 is always diagonalizable, and the system is well-posed. This proves the following claim.

LEMMA 1. *If $\varepsilon_l > 0$, $l = 1, \ldots, d-1$, and the original system* (1) *is well-posed, then the PML* (29) *is also well-posed.*

We conclude this section by noting that the PML for many problems is well-posed without the parabolic CFS. Additionally, if the parabolic CFS is used, ε_l should be chosen relative to the grid size such that it does not impose restrictions on the time-stepping.

3. Construction of energy estimates for constant coefficient Cauchy problems via annihilating polynomials. As we have seen in the previous section, the construction of a layer which is well-posed and perfectly matched is rather straightforward. However, it is not so straightforward to choose the free parameters η, ξ, μ, α_j, β_{jl}, δ_{jl}, ε_{jl}, and ν_j, for a given hyperbolic system, such that the solution does not grow with time. Related to this question is the stability of the constant coefficient Cauchy problem

$$\frac{\partial u(x,t)}{\partial t} = P\left(\frac{\partial}{\partial x}\right) u(x,t), \quad u(x,0) = u_0(x), \quad x \in \mathbf{R}^d, \quad 0 \geq t \geq T. \tag{33}$$

If we perform a Fourier transform in space, (33) reduces to a system of ordinary differential equations

$$\frac{\partial \hat{u}(k,t)}{\partial t} = P(ik)\hat{u}(k,t), \qquad k \in \mathbf{R}^s, \quad 0 \geq t \geq T, \tag{34}$$

$$\hat{u}(k,0) = \hat{u}_0(k). \tag{35}$$

We will distinguish between the following two types of stability.

DEFINITION 2 (stability). *We say that the Cauchy problem* (33) *is*

(i) strongly stable *if all solutions satisfy an estimate* $\|u(\cdot,t)\|_{L^2} \leq K\|u_0(\cdot)\|_{L^2}$*;*

(ii) weakly stable *if the solutions satisfy an estimate* $\|u(\cdot,t)\|_{L^2} \leq K(1+t)^p\|u_0(\cdot)\|_{H^s}$, *where* $s > 0$.

Note that if (33) is well-posed, we can replace H^s by L^2 in (ii). In the remainder of this paper we will drop the subscript of the L^2-norm, i.e., $\|\cdot\| \equiv \|\cdot\|_{L^2}$.

A necessary and sufficient condition for weak stability is that all eigenvalues λ_j of the symbol $P(ik)$ satisfy

$$\Re\{\lambda_j(P(ik))\} \leq 0. \tag{36}$$

Condition (36) can be checked by various methods that determine the number of zeros of polynomials in a half-plane. Below, we will first present a method that automatically generates a finite number of algebraic inequalities that can be used to check (36). Then we will show that if (36) holds, the method can also be used to construct a local energy density that decays with time.

We begin by recalling some definitions from matrix theory (see, e.g., [12]).

DEFINITION 3 (annihilating polynomial). *We say that a scalar polynomial $f(\lambda)$ is an* annihilating polynomial *of the square matrix A if*

$$f(A) = 0.$$

Two important annihilating polynomials are the characteristic polynomial and the minimal polynomial.

DEFINITION 4 (characteristic polynomial). *The scalar polynomial $f(\lambda)$ defined as*

$$f(\lambda) \equiv \det(\lambda I - A)$$

is called the characteristic polynomial of the matrix A.

DEFINITION 5 (minimal polynomial). *By $m_A(\lambda)$ we will denote the uniquely defined annihilating polynomial of lowest degree and with lead coefficient* 1. *The polynomial $m_A(\lambda)$ is called the* minimal polynomial *of A.*

Now, let $m_P(\lambda)$ be the minimal polynomial of the symbol $P(ik)$. Suppose its degree is n. To determine the number of roots with positive and negative real part of $m_P(\lambda) = 0$ for fixed k we can use the following lemma, which is a special case of Corollary (38,1b) in [21].

LEMMA 6. *Consider any polynomial $q(\lambda)$ of degree n. Let D be a real number, and define the polynomials Q_0 and Q_1 with real coefficients by*

$$q(iD) \equiv i^n \left[Q_0(D) + iQ_1(D)\right]. \tag{37}$$

Then there is a continued fraction

$$\frac{Q_1(D)}{Q_0(D)} = \cfrac{1}{c_1 D + d_1 - \cfrac{1}{c_2 D + d_2 - \cfrac{1}{c_3 D + d_3 - \cdots - \cfrac{1}{c_{n_r} D + d_{n_r}}}}}, \tag{38}$$

with $c_j \neq 0$ and $n_r \leq n$. The number of roots with positive (negative) real part equals the number of positive (negative) c_j. There are $n - n_r$ roots on the imaginary axis.

When we apply Lemma 6 to $m_p(\lambda)$, the number of nonzero coefficients c_j may depend on k. A change in sign corresponds to a root crossing the imaginary axis. We have the following corollary.

COROLLARY 7. *A necessary and sufficient condition for weak stability is that all c_j defined in* (38) *are negative, i.e.,*

$$c_j(k) < 0, \quad j = 1, 2, \ldots, n_r(k). \tag{39}$$

Remark. Strong stability follows if all eigenvalues (i.e., the roots of $m_P(\lambda) = 0$) have strictly negative real part for all k. However, in many cases there are certain k for which some roots have zero real part. If the corresponding eigenvectors span their respective invariant subspace, then the problem is still strongly stable. This condition must be checked in each case. We note that if (33) is well-posed, then, for sufficiently large $|k| \geq K$, $P(ik)$ can always be diagonalized. Thus, we need to check the eigenvectors only for roots that have zero real part at bounded $|k|$.

PROPOSITION 8. *Let $\hat{u}$ be the solution of the Fourier transformed system* (34). *Then any component $\hat{u}_i$ satisfies the equation*

$$g\left(\frac{\partial}{\partial t}\right)\hat{u}_i = 0,$$

where $g(\lambda)$ is any annihilating polynomial of the symbol $P(ik)$. In particular we have for the minimal polynomial of $P(ik)$

$$m_P\left(\frac{\partial}{\partial t}\right)\hat{u}_i(k,t) = 0. \tag{40}$$

Proof. By definition we have $g(P(ik)) = 0$. Multiplying by the solution vector from the right yields $g(P(ik))\hat{u} = 0$. By an easy induction argument we have, for any integer q, $(P(ik))^q\hat{u} = \frac{\partial^q \hat{u}}{\partial t^q}$. The proposition follows. ∎

By the following theorem we can construct decaying energies for the problem (33).

THEOREM 9. *Let $\hat{u}_i$ satisfy*

$$q\left(\frac{\partial}{\partial t}\right)\hat{u}_i = 0. \tag{41}$$

If (39) *holds for Q_0 and Q_1 defined as in Lemma* 6*, there exists an energy*

$$\mathcal{E}(t;k) \equiv \frac{1}{2}\sum_{j=1}^{n_r} |c_j||\hat{z}^{(j)}(k,t)|^2 \tag{42}$$

satisfying

$$\frac{\partial}{\partial t}\mathcal{E}(t;k) = -|\hat{z}^{(1)}(k,t)|^2. \tag{43}$$

The functions $\hat{z}^{(j)}$, $j = 1, \ldots, n$, are related to $\hat{u}_i(k,t)$ via the equations

$$\begin{aligned} &\prod_{j=1}^{n-n_r}\left(\frac{\partial}{\partial t} + ib_j(k)\right)\hat{u}_i(k,t) = -i\hat{z}^{(1)}(k,t), \qquad \Im b_j(k) = 0, \\ &\frac{\partial}{\partial t}\begin{bmatrix} |\,c_1|\hat{z}^{(1)} \\ \vdots \\ \vdots \\ |\,c_{n_r}|\hat{z}^{(n_r)} \end{bmatrix} = \begin{bmatrix} id_1 - 1 & -i & \cdots & 0 \\ -i & id_2 & \cdots & 0 \\ 0 & \vdots & \ddots & -i \\ 0 & \vdots & -i & id_{n_r} \end{bmatrix}\begin{bmatrix} \hat{z}^{(1)} \\ \vdots \\ \vdots \\ \hat{z}^{(n_r)} \end{bmatrix}. \end{aligned} \tag{44}$$

For the proof of Theorem 9 we refer to [14]. Note that system (44) can be used to eliminate all $\hat{z}^{(j)}$ so that the energy (42) is expressed in $\hat{u}_i$ alone.

3.1. Localization of $\mathcal{E}(t;k)$. Since the coefficients c_j are rational functions in k, localization can be accomplished by multiplication with a suitable polynomial in k. In particular let $\gamma(k)$ be a polynomial such that $\tilde{c}_j(k) = -\gamma(k)^2 c_j(k)$ is also a polynomial in k. Then, since by assumption $\tilde{c}_j(k) > 0$ for all k, it can be decomposed as

$$\tilde{c}_j(k) = \sum_l q_l^2(k), \tag{45}$$

where $q_l(k)$ are real polynomials in k. By multiplying (43) with $\gamma(k)^2$ we get

$$\begin{aligned} \frac{d}{dt}\frac{1}{2}\sum_{j=1}^{n_r} |\gamma(k)^2 c_j(k)|\,|\hat{z}^{(j)}(t,k)|^2 &= \frac{d}{dt}\frac{1}{2}\sum_{j=1}^{n_r}\sum_l \left|q_l(k)\,\hat{z}^{(j)}(t,k)\right|^2 \\ &= -\left|\gamma(k)\,\hat{z}^{(1)}(t,k)\right|^2. \end{aligned} \tag{46}$$

Integrating over k yields

$$\begin{aligned} &\frac{d}{dt}\frac{1}{2}\sum_{j=1}^{n_r}\sum_l \int_{\mathbb{R}^s} \left|q_l(k)\,\gamma(k)\,\hat{z}^{(j)}(t,k)\right|^2 dk \\ &\qquad = -\int_{\mathbb{R}^s} \left|\gamma(k)\,\hat{z}^{(1)}(t,k)\right|^2 dk, \end{aligned}$$

and by applying Parseval's formula $\int |\hat{f}(k)|^2 dk = \|f(x)\|^2$, we obtain the following result.

COROLLARY 10. *There exists a polynomial $\gamma(k)$ such that the inverse transform of* (46) *is*

$$\frac{d}{dt}E(t) = -\left\|\mathcal{F}^{-1}\left\{\gamma(k)\,\hat{z}^{(1)}(t,k)\right\}\right\|^2. \tag{47}$$

Here

$$E(t) = \frac{1}{2}\sum_{j=1}^{n}\sum_l \left\|\mathcal{F}^{-1}\left\{q_l(k)\,\hat{z}^{(j)}(t,k)\right\}\right\|^2 \tag{48}$$

contains only local quantities.

Note that Theorem 9 can be used with any annihilating polynomial of $P(ik)$. If the minimal polynomial is available, it is advantageous to use it since it has lower degree and thus will produce an energy with lower order derivatives. Its lower degree also simplifies the computation of the continued fraction.

4. PML for Maxwell's equations. The first problem we consider is the scaled TM_z problem in a lossless medium. Then Maxwell's equations can be written

$$\begin{gathered} \frac{\partial u}{\partial t} + A_x\frac{\partial u}{\partial x} + A_y\frac{\partial u}{\partial y} = 0, \\ A_x = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix}, \quad A_y = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \end{gathered}$$

where $u = [H_x, H_y, E_z]^T$. We consider a layer in the x direction. Here A_x is singular and there are only two modes that propagate in the x direction. Hence, we add auxiliary variables only to the x-propagating H_y and E_z fields. The layer we will consider is defined by the equations

$$\begin{aligned}
\frac{\partial H_x}{\partial t} + \frac{\partial E_z}{\partial y} &= 0,\\
\frac{\partial H_y}{\partial t} - (1+\eta\sigma)\frac{\partial E_z}{\partial x} &= \sigma\phi_2,\\
\frac{\partial E_z}{\partial t} - (1+\eta\sigma)\frac{\partial H_y}{\partial x} + \frac{\partial H_x}{\partial y} &= \sigma\phi_1,\\
\frac{\partial \phi_1}{\partial t} + \frac{\partial E_z}{\partial x} &= -(\sigma+\alpha)\phi_1 + \varepsilon\frac{\partial^2\phi_1}{\partial y^2},\\
\frac{\partial \phi_2}{\partial t} + \frac{\partial H_y}{\partial x} &= -(\sigma+\alpha)\phi_2 + \varepsilon\frac{\partial^2\phi_2}{\partial y^2}.
\end{aligned} \tag{49}$$

Here we have included the parameter η, which will improve the damping of evanescent modes. Note that if $\varepsilon = 0$, the above equations are only weakly hyperbolic and thus only weakly well-posed. To ensure strong well-posedness we take $\varepsilon > 0$.

4.1. Stability for constant σ. When σ is constant we can take the Fourier transform in x and y. For simplicity let $\eta = 0$. The symbol of (49) then becomes

$$P(ik) = -\begin{bmatrix} ik_xA_x + ik_yA_y & \sigma D \\ ik_xE & (\alpha+\sigma+\varepsilon k_y^2)I \end{bmatrix}, \tag{50}$$

where

$$D = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}^T, \qquad E = -\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The minimal polynomial of (50) coincides with the characteristic polynomial and can be written as a product of the two polynomials $m_P(\lambda) = m_1(\lambda)m_4(\lambda)$:

$$\begin{aligned}
m_1(\lambda) &= \lambda,\\
m_4(\lambda) &= \big(\lambda^4 + 2(\tau+\sigma)\lambda^3 + (k_x^2 + k_y^2 + (\tau+\sigma)^2)\lambda^2\\
&\quad + 2(\tau k_x^2 + (\tau+\sigma)k_y^2)\lambda + k_x^2\tau^2 + (\tau+\sigma)^2k_y^2\big),
\end{aligned}$$

where we have introduced $\tau = \alpha + \varepsilon k_y^2$.

To determine the sign of the eigenvalues we apply Lemma 6 to $m_4(\lambda)$. The coefficients in the continued fraction (38) are

$$c_1 = -\frac{1}{2(\tau+\sigma)}, \tag{51}$$

$$c_2 = -\frac{2(\tau+\sigma)^2}{(\tau+\sigma)^3 + \sigma k_x^2}, \tag{52}$$

$$c_3 = -\frac{\left((\tau+\sigma)^3 + \sigma k_x^2\right)^2}{2\sigma k_x^2\left((\tau+\sigma)(\tau^2+\sigma\tau+k_y^2) + \tau k_x^2\right)(\tau+\sigma)}, \tag{53}$$

$$c_4 = -\frac{2\sigma k_x^2\left((\tau+\sigma)(\tau^2+\tau\sigma+k_y^2) + \tau k_x^2\right)}{\left((\tau+\sigma)^2k_y^2 + k_x^2\tau^2\right)\left((\tau+\sigma)^3 + \sigma k_x^2\right)}, \tag{54}$$

$$d_1 = d_2 = d_3 = d_4 = 0. \tag{55}$$

Clearly all c_j are negative and defined except for the cases $k_x = k_y = 0$ and $k_x = 0, k_y \neq 0$. When $k_x = 0$ the minimal equation reduces to $\lambda(\lambda^2 + k_y^2)(\lambda + \sigma + \tau)^2 = 0$ with solutions $\lambda = 0, \pm ik_y, -(\sigma+\tau), -(\sigma+\tau)$. The eigenvalues with zero real part are distinct as long as $k_y \neq 0$. For the case $k_y = 0$ there could potentially be algebraic growth. However, it is easily checked that there are three independent eigenvectors when $k_x = k_y = 0$. Thus (49) is strongly stable when $\eta = 0$. When $\eta \neq 0$ the coefficients are somewhat more complicated, but strong stability follows similarly. This concludes the proof of the following lemma.

LEMMA 11. *For constant $\sigma > 0$, $\alpha > 0$, $\varepsilon > 0, \eta\sigma + 1 > 0$, the system (49) is strongly stable.*

4.2. Energy estimates. We now consider decaying energies of the system (49). We start by noticing that $m_4(P(ik))$ annihilates $\hat{E}_z$ and $\hat{\phi}_2$, while $m_P(\partial/\partial t)$ annihilates $\hat{H}_x$, $\hat{H}_y$, and $\hat{\phi}_1$. Thus we have that

$$m_4\left(\frac{\partial}{\partial t}\right)\hat{v} = 0 \quad \text{for} \quad \hat{v} = \hat{E}_z, \hat{\phi}_2, \frac{\partial \hat{H}_x}{\partial t}, \frac{\partial \hat{H}_y}{\partial t}, \frac{\partial \hat{\phi}_1}{\partial t}.$$

It follows from Theorem 9 and (51)–(55) that the energy

$$\mathcal{E}(t;k) \equiv \frac{1}{2}\sum_{j=1}^{4} |c_j||\hat{z}^{(j)}(k,t)|^2 \tag{56}$$

decays with time. To express (56) in $\hat{v}$ we use (44). This yields

$$\begin{aligned}
|z^{(1)}| &= |\hat{v}|, \qquad |z^{(2)}| = \left|\left(|c_1|\frac{\partial}{\partial t} + 1\right)\hat{v}\right|, \\
|z^{(3)}| &= \left|\left(|c_2|\frac{\partial}{\partial t}\left(|c_1|\frac{\partial}{\partial t} + 1\right) + 1\right)\hat{v}\right|, \\
|z^{(4)}| &= \left|\left(|c_3|\frac{\partial}{\partial t}\left(|c_2|\frac{\partial}{\partial t}\left(|c_1|\frac{\partial}{\partial t} + 1\right) + 1\right) + \left(|c_1|\frac{\partial}{\partial t} + 1\right)\right)\hat{v}\right|.
\end{aligned} \tag{57}$$

Since $\mathcal{E}$ is a function of c_3 and c_4, whose denominators vanish for certain k_x and k_y, it is not bounded. To formulate energies in physical space, we first remove the singularities of $\mathcal{E}$ by multiplying (56) by a suitable polynomial in k_x and k_y. Here we will consider two different polynomials, the first producing a semilocal energy and the second a fully local energy.

4.2.1. A semilocal energy. We would like the order of the spatial derivatives of v appearing in the energy in physical space to be as low as possible. At the same time, the energy must be bounded for all k_x and k_y so that we can use Parseval. The energy

$$\mathcal{E}_{SL}(t;k) = 2(\tau+\sigma)k_x^2\left((\tau+\sigma)^2k_y^2 + k_x^2\tau^2\right)\mathcal{E}(t;k) \tag{58}$$

satisfies these requirements. We can split $\mathcal{E}_{SL}$ into a local and a nonlocal part

$$\mathcal{E}_{SL}(t;k) = \mathcal{E}_L + \mathcal{E}_{NL}. \tag{59}$$

The local and nonlocal energies are

$$\mathcal{E}_L = k_x^2\left((\tau+\sigma)^2k_y^2 + k_x^2\tau^2\right)|\hat{v}|^2, \tag{60}$$

$$\mathcal{E}_{NL} = 2(\tau+\sigma)k_x^2\left((\tau+\sigma)^2k_y^2 + k_x^2\tau^2\right)\sum_{j=2}^{4}|c_j||\hat{z}^{(j)}(k,t)|^2. \tag{61}$$

Now by using Parseval we get

$$\frac{d}{dt}\left(E_L(t;v)+E_{NL}(t)\right) = -2(\sigma+\alpha)E_L(t;v) - 2\varepsilon E_L(t;\partial_y v), \tag{62}$$

where

$$\begin{aligned} E_L(t;v) = (\alpha+\sigma)^2\|\partial_x\partial_y v(\cdot,t)\|^2 + \alpha^2\|\partial_x^2 v(\cdot,t)\|^2 + 2\varepsilon(\alpha+\sigma)\|\partial_x\partial_y^2 v(\cdot,t)\|^2 \\ + \varepsilon^2\|\partial_x\partial_y^3 v(\cdot,t)\|^2 + 2\varepsilon\alpha\|\partial_x^2\partial_y v(\cdot,t)\|^2 + \varepsilon^2\|\partial_x^3\partial_y v(\cdot,t)\|^2. \end{aligned} \tag{63}$$

We do not state $E_{NL}(t)$ explicitly, since for our purpose it is sufficient to know that it is bounded and nonnegative. However, we note that $E_{NL}(t)$ is nonlocal in space.

We see that

$$E_L(t;v) \le E_L(0\,;v) + E_{NL}(0\,;v), \tag{64}$$

which proves the following claim.

LEMMA 12. *Let v be any of the fields*

$$E_z, \phi_2, \frac{\partial H_x}{\partial t}, \frac{\partial H_y}{\partial t}, \frac{\partial \phi_1}{\partial t}.$$

If $\sigma > 0$, $\alpha > 0$, $\varepsilon > 0$ and constant, then v satisfies the estimate

$$(\alpha+\sigma)^2\|\partial_x\partial_y v(\cdot,t)\|^2 + \alpha^2\|\partial_x^2 v(\cdot,t)\|^2 \le C = E_L(0\,;v) + E_{NL}(0\,;v). \tag{65}$$

4.2.2. A local energy. To obtain a fully local energy we need to clear the denominators of (56) and (57). Again, this is done by multiplying $\mathcal{E}$ by a suitable factor. For this case we define the fully local energy by

$$\begin{aligned} \mathcal{E}_{FL} \equiv\ & (\tau+\sigma)^2((\tau+\sigma)^3+\sigma k_x^2) \\ & \times \left((\tau+\sigma)^2 k_y^2 + k_x^2\tau^2\right) k_x^2 \left((\tau+\sigma)(\tau^2+\sigma\tau+k_y^2)+\tau k_x^2\right)\mathcal{E}. \end{aligned}$$

$\mathcal{E}_{FL}(t;k)$ can be split into

$$\mathcal{E}_{FL}(t;k) = \mathcal{E}^I + \mathcal{E}^{II} + \mathcal{E}^{III} + \mathcal{E}^{IV},$$

where

$$\begin{aligned} \mathcal{E}^I &= \frac{1}{2}(\tau+\sigma)((\tau+\sigma)^3+\sigma k_x^2)\left((\tau+\sigma)^2 k_y^2 + k_x^2\tau^2\right) \\ &\quad \times k_x^2\left((\tau+\sigma)(\tau^2+\sigma\tau+k_y^2)+\tau k_x^2\right)|\hat{v}|^2, \\ \mathcal{E}^{II} &= 2(\tau+\sigma)^2 k_x^2\left((\tau+\sigma)^2 k_y^2 + k_x^2\tau^2\right) \\ &\quad \times \left((\tau+\sigma)(\tau^2+\sigma\tau+k_y^2)+\tau k_x^2\right)\Big|\underbrace{(\frac{1}{2}\partial_t+\tau+\sigma)\hat{v}}_{\hat{\chi}_1}\Big|^2, \\ \mathcal{E}^{III} &= \frac{1}{2\sigma}(\tau+\sigma)((\tau+\sigma)^3+\sigma k_x^2)\left((\tau+\sigma)^2 k_y^2 + k_x^2\tau^2\right) \\ &\quad \times \Big|\underbrace{\left((\sigma+\tau)\partial_t^2 + 2(\sigma+\tau)^2\partial_t + (\tau+\sigma)^3+\sigma k_x^2\right)\hat{v}}_{\hat{\chi}_2}\Big|^2, \\ \mathcal{E}^{IV} &= \left|\left(((\tau+\sigma)^3+\sigma k_x^2)\partial_t\left(\frac{(\tau+\sigma)}{2\sigma}\partial_t^2 + \frac{(\tau+\sigma)^2}{\sigma}\partial_t + \frac{1}{2\sigma}\right)\right.\right. \\ &\quad \left.\left. + k_x^2((\tau+\sigma)(\tau^2+\sigma\tau+k_y^2)+\tau k_x^2)\left(\frac{1}{2}\partial_t+\sigma+\tau\right)\right)\hat{v}\right|^2. \end{aligned}$$

To localize the energies we need to write them in the form (45). This is a straightforward operation, but the resulting expressions become lengthy (they contain many combinations of higher derivatives) and are therefore presented in Appendix A.

Let E^n be the physical space version of the energy $\mathcal{E}^n$. Then we have that

$$\frac{d}{dt}\left(E^I(t) + E^{II}(t) + E^{III}(t) + E^{IV}(t)\right) = -E^I(t), \tag{66}$$

which means that

$$\left(E^I(t) + E^{II}(t) + E^{III}(t) + E^{IV}(t)\right) \leq \left(E^I(0) + E^{II}(0) + E^{III}(0) + E^{IV}(0)\right).$$

Thus E^I, E^{II}, E^{III}, and E^{IV} all remain bounded.

It may be possible to derive sharper results from this fully local energy by using the system (49). We note that the energy estimates obtained by Bécache, Petropoulos, and Gedney in [7] are stated in terms of the fields rather than the derivatives of the fields, so that strong stability is a straightforward consequence of the energy inequality. We emphasize that our results also imply strong stability, even though the energy is stated in terms of derivatives of the fields.

5. The linearized Euler equations. The next problem we consider is the Euler equations in two dimensions linearized around a subsonic skew flow

$$\frac{\partial u}{\partial t} + A_x \frac{\partial u}{\partial x} + A_y \frac{\partial u}{\partial y} = 0,$$

where

$$u = \begin{bmatrix} \rho \\ v_x \\ v_y \\ p \end{bmatrix}, \quad A_x = \begin{bmatrix} M_x & 1 & 0 & 0 \\ 0 & M_x & 0 & 1 \\ 0 & 0 & M_x & 0 \\ 0 & 1 & 0 & M_x \end{bmatrix}, \quad A_y = \begin{bmatrix} M_y & 0 & 1 & 0 \\ 0 & M_y & 0 & 0 \\ 0 & 0 & M_y & 1 \\ 0 & 0 & 1 & M_y \end{bmatrix}.$$

Here ρ is the density; v_x and v_y are the velocities in the x and y directions, respectively; p is the pressure; and M_x and M_y are the Mach numbers in the x and y directions. We have that $0 < M_x < 1, 0 < M_y < 1$ since the flow is assumed to be subsonic.

From [17] we conclude that a suitable layer in the x-direction should be of the form

$$\begin{aligned} &\frac{\partial u}{\partial t} + A_x\left(\frac{\partial u}{\partial x} + \mu\sigma u + \sigma\phi\right) + A_y \frac{\partial u}{\partial y} = 0, \\ &\frac{\partial \phi}{\partial t} + \frac{\partial u}{\partial x} + M_y \frac{\partial \phi}{\partial y} + (\sigma + \alpha)(\mu u + \phi) = 0. \end{aligned} \tag{67}$$

The symbol $P(ik)$ of (67) is

$$P(ik) = -\begin{bmatrix} (ik_x + \mu\sigma)A_x + ik_y A_y & \sigma A_x \\ ik_x I + (\sigma + \alpha)\mu I & (ik_y M_y + \sigma + \alpha)I \end{bmatrix}. \tag{68}$$

Note that here we do not need to include the parabolic CFS. To establish well-posedness, we simply freeze the coefficients and consider the principal part of $P(ik)$,

$$P_1(ik) = -\begin{bmatrix} ik_x A_x + ik_y A_y & 0 \\ ik_x I & ik_y M_y I \end{bmatrix}. \tag{69}$$

The eigenvalues of the upper diagonal block are easy to compute. They coincide with $ik_y M_y I$ only when $k_x = 0$. Thus P_1 is diagonalizable, and well-posedness follows.

5.1. Stability for constant σ. In [17] it was shown that the choice

$$\mu = \frac{M_x}{1 - M_x^2} \tag{70}$$

is necessary for the solution in a layer closely related to (67) to decay in space. Similar conclusions, from another point of view, were reached by Hu in [19]. The results for decay in space from [17] apply directly to (67). Here we will show that (70) is also necessary and sufficient for stability (in time) for (67).

First we show that (70) is sufficient. We note that the real part of the eigenvalues of $P(ik)$ (with μ given by (70)) coincides with the real part of the eigenvalues of the matrix $\tilde{P}(ik) \equiv P(ik) - ik_y M_y I$. Since $\tilde{P}(ik)$ has a sparser structure, it is easier to check that its eigenvalues have nonpositive real part.

The minimal polynomial of $\tilde{P}(ik)$ can be factored $m_{\tilde{P}}(\lambda) = m_1(\lambda)m_2(\lambda)$, where

$$m_1(\lambda) = \lambda^2 + \left(ik_x M_x + \frac{\sigma}{\zeta} + \alpha \right) \lambda + ik_x \alpha M_x \tag{71}$$

and

$$\begin{aligned} m_2(\lambda) =& \lambda^4 + 2\left(ik_x M_x + \frac{\sigma}{\zeta} + \alpha \right) \lambda^3 \\ &+ \left(4\alpha i k_x M_x + \zeta k_x^2 + k_y^2 + \frac{(\sigma+\alpha)^2 - M_x^2 \alpha^2}{\zeta} \right) \lambda^2 \\ &+ 2\left(ik_x \alpha M_x + \zeta \alpha k_x^2 + (\alpha + \sigma) k_y^2 \right) \lambda + \zeta \alpha^2 k_x^2 + (\alpha+\sigma)^2 k_y^2. \end{aligned} \tag{72}$$

Here we have introduced $\zeta = 1 - M_x^2$. The continued fraction coefficients for (71) are

$$c_1 = -\frac{\zeta}{2(\sigma + \alpha\zeta)}, \tag{73}$$

$$c_2 = -\frac{2(\sigma + \alpha\zeta)^3}{\alpha\sigma M_x^2 \zeta^2 k_x^2}. \tag{74}$$

For (72) the coefficients are

$$c_1 = -\frac{\zeta}{2(\sigma + \alpha\zeta)}, \tag{75}$$

$$c_2 = -\frac{2(\sigma + \alpha\zeta)^3}{c_{2a}}, \tag{76}$$

$$c_3 = -\frac{c_{2a}^3}{2\sigma\zeta(\sigma + \alpha\zeta)^4 c_{3a}}, \tag{77}$$

$$c_4 = -\frac{2(\alpha\zeta + \sigma)^4 c_{3a}^3 \sigma}{c_{2a}^4 (k_y^2 M_x^2 - \zeta k_x^2)^2 (\alpha^2 \zeta k_x^2 + (\alpha+\sigma)^2 k_y^2) c_{4a}}, \tag{78}$$

where c_{2a}, c_{3a}, c_{4a} are positive for all k_x and k_y and can be found in Appendix B. We see that all the coefficients are negative and defined for all k_x and k_y except the cases (a) $k_x = k_y = 0$, (b) $k_x = 0$, $k_y \neq 0$, and (c) $(1 - M_x^2)k_x^2 = M_x^2 k_y^2$. We will consider these cases separately.

First we consider the case (a), for which we easily can compute the eigenvalues of $\tilde{P}(k_x = 0, k_y = 0) = P(k_x = 0, k_y = 0)$. They are

$$0, \quad -\frac{\sigma}{1 - M_x} - \alpha, \quad -\frac{\sigma}{1 + M_x} - \alpha, \quad -\frac{\sigma}{1 - M_x^2} - \alpha.$$

The zero eigenvalue has multiplicity four, and there could potentially be algebraic growth. However, straightforward calculations show that there are also four independent eigenvectors, and this mode will be strongly stable.

For the case (b), the minimal polynomial of $\tilde{P}(ik)$ can be factored into

$$m_{\tilde{P}}(\lambda) = n_1(\lambda)n_2(\lambda)n_3(\lambda),$$
$$n_1(\lambda) = \lambda, \quad n_2(\lambda) = \lambda + \frac{\sigma + \alpha\zeta}{\zeta}, \quad n_3(\lambda) = m_2(\lambda; k_x = 0).$$

Directly, we see that the eigenvalues $\lambda = 0$ and $\lambda = -(\sigma/\zeta + \alpha)$, being solutions to $n_1(\lambda) = 0$ and $n_2(\lambda) = 0$, have nonpositive real parts. The double zero eigenvalue of $\tilde{P}(ik)$ corresponds to the double eigenvalue $\lambda = -ik_y M_y$ of $P(ik)$. Associated with $\lambda = -ik_y M_y$ there are two linearly independent eigenvectors, and thus stability will not be lost.

For $n_3(\lambda)$, we compute the coefficients in the continued fraction. They are

$$c_1 = -\frac{\zeta}{2(\sigma + \alpha\zeta)},$$
$$c_2 = -\frac{2(\sigma + \zeta\alpha)^2}{\sigma M_x^2 k_y^2 \zeta + (\sigma + \alpha(1 + M_x))(\sigma + \alpha(1 - M_x))(\sigma + \zeta\alpha)},$$
$$c_3 = -\frac{(\sigma M_x^2 k_y^2 \zeta + (\sigma + \alpha(1 + M_x))(\sigma + \alpha(1 - M_x))(\sigma + \zeta\alpha))^2}{2k_y^2 \zeta \sigma M_x^2 (\sigma + \alpha)(\sigma\alpha + \zeta(\alpha^2 + k_y^2))},$$
$$c_4 = -\frac{2\sigma M_x^2 (\sigma\alpha + \zeta(\alpha^2 + k_y^2))}{(\sigma + \alpha)(\sigma M_x^2 k_y^2 \zeta + (\sigma + \alpha(1 + M_x))(\sigma + \alpha(1 - M_x))(\sigma + \zeta\alpha))}.$$

Due to the assumptions $0 < M_x < 1$, $\zeta > 0$, $\sigma > 0$, and $\alpha > 0$ they are all negative.

Finally we consider case (c). For this case $m_{\tilde{P}}(\lambda)$ again factors into three polynomials

$$m_{\tilde{P}}(\lambda) = o_1(\lambda)o_2(\lambda)o_3(\lambda),$$
$$o_1(\lambda) = \lambda - ik_x \frac{\zeta}{M_x}, \quad o_2(\lambda) = m_1(\lambda),$$

$$o_3(\lambda) = \lambda^3 + \left(2\frac{\sigma + \zeta\alpha}{1 - M_x^2} + ik_x \frac{1 + M_x^2}{M_x}\right)\lambda^2 + \left(\frac{(\sigma + \alpha)^2 - \alpha^2 M_x^2}{\zeta} + ik_x \frac{2(\sigma + \alpha(1 + M_x^2))}{M_x}\right)\alpha + ik_x \frac{(\sigma + \alpha)^2 + \alpha^2 M_x^2}{M_x}.$$

The eigenvalue belonging to $o_1(\lambda)$ is distinct and does not affect strong stability, and the polynomial $o_2(\lambda) = m_1(\lambda)$ has already been checked. It only remains to check the coefficients of the continued fraction arising from $o_3(\lambda)$. They are

$$c_1 = -\frac{\zeta}{2(\sigma + \alpha\zeta)}, \tag{79}$$
$$c_2 = -\frac{2(\sigma + \alpha\zeta)^3}{c_{2a}}, \tag{80}$$
$$c_3 = -\frac{c_{2a}^3 M_x^2}{4(\sigma + \alpha\zeta)^4 k_x^2 \sigma \zeta^2 (\alpha^2 M_x^2 + (\sigma + \alpha)^2)} \times \frac{1}{\alpha M_x^2 (\alpha + \sigma)(\sigma + \alpha\zeta)^2 + \zeta^2 k_x^2 (\sigma + \alpha(1 + M_x^2))^2}, \tag{81}$$

where

$$c_{2a} = (\sigma + \alpha\zeta)^2((\sigma + \alpha)^2 - \alpha^2 M_x^2) + 2\sigma k_x^2 \zeta^2(\sigma + \alpha(1 + M_x^2)). \tag{82}$$

Clearly, (79)–(81) are negative and defined unless $k_x = k_y = 0$. However, that particular case has already been checked.

To see that (70) is a necessary condition we use the parameterization $k_y = \kappa$, $k_x = \gamma\kappa$ and compute the minimal polynomial of $\tilde{P}(\kappa, \gamma)$ with μ as a free parameter. Again the minimal polynomial can be factored into a quadratic and a quartic. If we compute the coefficients in the continued fraction for the quartic, we see that for κ large the sign of the coefficients c_3 and c_4 will be determined by the sign of the expression

$$\kappa^4(M_x^2 - \gamma^2 + \gamma^2 M_x^2)(M_x^2\mu^2\gamma^2 + 2\gamma^2\mu M_x + \gamma^2 - \mu^2\gamma^2 - \mu^2). \tag{83}$$

Since the expression $(M_x^2 - \gamma^2 + \gamma^2 M_x^2)$ will change sign when $\gamma^2 = M_x^2/(1 - M_x^2)$ we must choose μ such that the sign of the last expression in (83) changes simultaneously. Hence μ must satisfy

$$(M_x^4 - 1)\mu^2 + 2M_x^3\mu + M_x^2 = 0;$$

i.e., we must choose

$$\mu = -\frac{M_x}{1 + M_x^2} \quad \text{or} \quad \mu = \frac{M_x}{1 - M_x^2}.$$

The first choice will violate the conditions for c_2 when k_x is large and cannot be used, while the second choice, as we have seen above, yields a strongly stable PML.

We summarize the results in the following lemma.

LEMMA 13. *For constant $\sigma > 0$, $\alpha > 0$, and $0 < M_x < 1$, a necessary and sufficient condition for strong stability of the system* (67) *is that*

$$\mu = \frac{M_x}{1 - M_x^2}. \tag{84}$$

6. A stable PML for general 2×2 symmetric hyperbolic systems. Our final example is the symmetric hyperbolic system

$$\frac{\partial u}{\partial t} + \underbrace{\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}}_{A} \frac{\partial u}{\partial x} + \underbrace{\begin{bmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{bmatrix}}_{B} \frac{\partial u}{\partial y} = 0. \tag{85}$$

Here A and B are real matrices, and we can choose $a_{12} = 0$ without loss of generality. Note that the convective wave equation

$$\left(\frac{\partial}{\partial t} + M\frac{\partial}{\partial x}\right)^2 u = C^2\nabla^2 u$$

is a special case of (85) if we choose

$$a_{11} = M + C, \quad a_{22} = M - C, \quad b_{12} = C, \quad a_{12} = b_{11} = b_{22} = 0.$$

Equation (85) also contains the anisotropic wave equation as a special case:

$$\frac{\partial^2 u}{\partial t^2} = \nabla \cdot (T\,\nabla u),$$
$$T = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \quad a > 0, \quad c > 0, \quad ac - b^2 > 0,$$

describing electromagnetic waves propagating in an anisotropic dielectric media. Here

$$a_{11} = -a_{22}, \quad a_{12} = 0, \quad b_{11} = -b_{22},$$
$$a_{11} = \sqrt{a}, \quad b_{11} = \frac{b}{\sqrt{a}}, \quad b_{12} = \sqrt{c - \frac{b^2}{a}}.$$

The direction in which the waves supported by the system (85) propagate depends on the coefficients of A and B. If A is nonsingular, there are three distinct cases:

(i) $a_{11}a_{22} < 0, b_{12} \neq 0$: coupled waves moving in opposite x-directions.
(ii) $a_{11}a_{22} > 0, b_{12} \neq 0$: coupled waves moving in the same x-direction.
(iii) $b_{12} = 0$, $a_{11}a_{22} \neq 0$: decoupled waves.

For the cases (ii) and (iii), there is no need to use a PML since waves can be damped without reflection by simply adding a damping term

$$\begin{bmatrix} \sigma_1(x) & 0 \\ 0 & \sigma_2(x) \end{bmatrix} u$$

to (85). The appropriate signs of σ_1 and σ_2 can be determined by the sign of a_{jj}; see also [4]. The case (i) is more interesting. In [4], the following PML model is suggested:

$$\begin{aligned} \frac{\partial u}{\partial t} + A\left(\frac{\partial u}{\partial x} + \sigma\mu u + \phi\right) + B\frac{\partial u}{\partial y} &= 0, \\ \frac{\partial \phi}{\partial t} + (\sigma+\alpha)\phi + \beta\frac{\partial \phi}{\partial y} &= \sigma\left(\gamma\frac{\partial u}{\partial x} + \mu(\sigma+\alpha)u + \delta\frac{\partial u}{\partial y}\right), \end{aligned} \tag{86}$$

where

$$\delta = \frac{b_{22} - b_{11}}{a_{11} - a_{22}}, \qquad \mu = -\frac{a_{11} + a_{22}}{2|a_{11}a_{22}|}, \tag{87}$$

$$\beta = -\frac{b_{11}a_{22} - b_{22}a_{11}}{a_{11} - a_{22}}, \qquad \gamma = -1, \quad \alpha \geq 0. \tag{88}$$

We note that this is a special case of the general layer studied above. With this choice of parameters we have the following claim.

LEMMA 14. *For $\sigma > 0$ and constant, $a_{11}a_{22} < 0$, and $\alpha \geq 0$, the system* (86) *is at least weakly stable.*

Proof. For simplicity we give the proof only for the case $\alpha = 0$. For the general choice of $\alpha \neq 0$ the coefficients are more complicated, but stability follows similarly. We consider three different cases, (a) $k_x \neq 0, k_y \neq 0$, (b) $k_x \neq 0, k_y = 0$, and (c) $k_x = k_y = 0$. First we consider the case (a) and compute the coefficients in the continued fraction, (38) in Lemma 6. They are

$$c_1 = \frac{2a_{22}a_{11}}{\sigma\left(a_{11} - a_{22}\right)^2},$$
$$c_2 = -\frac{2\sigma\left(a_{11} - a_{22}\right)^4}{c_{2a}},$$
$$\begin{aligned} c_{2a} = \sigma^2(a_{11} - a_{22})^4 + 4a_{11}^2a_{22}^2\left(k_x(a_{11} - a_{22}) + k_y(b_{11} - b_{22})\right)^2 \\ - 4a_{11}a_{22}b_{12}^2k_y^2(a_{11} + a_{22})^2\big), \end{aligned}$$

$$c_3 = \frac{1}{32}\frac{c_{2a}^3}{(a_{11}-a_{22})^2 c_{3a} a_{22}^2 a_{11}^2 b_{12}^2 k_y^2 \sigma},$$

$$c_{3a} = \sigma^2(a_{11}-a_{22})^4 c_{3b} + 4a_{22}a_{11}c_{3b}^2,$$

$$c_{3b} = -k_y^2 b_{12}^2 (a_{11}+a_{22})^2 + a_{22}a_{11}(k_y(b_{11}-b_{22}) + k_x(a_{11}-a_{22}))^2,$$

$$c_4 = -8\frac{c_{3a}^3 a_{22} a_{11}}{c_{4a}^4 \sigma c_{4b}^2},$$

where c_{4a} and c_{4b} can be found in Appendix C. The coefficients are negative except for the cases (b) and (c), but then the eigenvalues can be computed directly. For case (b) they are

$$0, \quad ika_{11}a_{22} - \frac{\sigma}{2}\left(1 - \frac{a_{11}}{a_{22}}\right), \quad ika_{11}a_{22} - \frac{\sigma}{2}\left(1 - \frac{a_{22}}{a_{11}}\right),$$

and for the case (c) they are

$$0, \quad -\frac{\sigma}{2}\left(1 - \frac{a_{11}}{a_{22}}\right), \quad -\frac{\sigma}{2}\left(1 - \frac{a_{22}}{a_{11}}\right).$$

Since $a_{11}a_{22} < 0$ they all have nonnegative real part, and the lemma is proved. ∎

As a final remark, in many cases the words "at least weakly" in Lemma 14 can be replaced by "strongly." However, to prove this we need to consider all cases when c_{4a} or c_{4b} vanish. Considering the complexity of the expressions c_{4a} and c_{4b}, we expect the necessary calculations to be quite tedious.

7. Summary. We have presented a very general PML model for first order hyperbolic systems. We believe that the generality should make the model suitable for many future applications. We have also proven that the equations in the layer are perfectly matched to the equations in the computational domain. For the model formulated with one set of auxiliary variables, we have also shown that the layer equations can always be made strongly well-posed.

The critical step in the construction of a PML is to choose the free parameters so that the solution in the layer is stable. To simplify the analysis of this step, we have presented a method with which the stability of the layer can be determined by checking a fixed number of algebraic inequalities, which in turn can be generated automatically. Additionally, if these inequalities hold, we showed that there is an energy density in Fourier space that decays with time. By simple algebraic manipulations and application of Parseval's relation, this energy density can be converted to a decaying energy in physical space. The energy contains only the solution and its spatial and temporal derivatives; i.e., the energy is local.

We have used the introduced techniques to show strong stability for a PML for Maxwell's equations and a PML for the linearized Euler equations. We also showed weak stability for a PML for a general 2×2 hyperbolic system in $(2+1)$ dimensions. For the PML for Maxwell's equations, we also derived a semilocal and a local energy. These energies guarantee the time-decay of higher order derivatives in space and time of the solution.

Unlike techniques that only involve checking the roots of the characteristic polynomial, our method is applicable to variable coefficient problems. This is important since in "real life" the damping parameter σ is not constant. The stability of the variable coefficient problem can be analyzed as a perturbation of the constant coefficient problem. If the constant coefficient problem is stable, our method generates an energy. Since the energy decays for constant σ we expect it to decay at least for slowly varying σ.

Appendix A. Space-time energies for Maxwell PML. The space-time energy version of $\mathcal{E}^{IV}$ is obtained by integration over all wavenumbers and application of Parseval. It is

$$\begin{aligned} E^{IV} = \Bigg\| \Bigg(& ((\alpha - \varepsilon\partial_y^2 + \sigma)^3 - \sigma\partial_x^2)\partial_t \left(\frac{1}{2\sigma}(\alpha - \varepsilon\partial_y^2 + \sigma)\partial_t^2 + \frac{1}{\sigma}(\alpha - \varepsilon\partial_y^2 + \sigma)^2\partial_t + \frac{1}{2\sigma} \right) \\ & - \partial_x^2 \left((\alpha - \varepsilon\partial_y^2 + \sigma)((\alpha - \varepsilon\partial_y^2)^2 + \sigma(\alpha - \varepsilon\partial_y^2) - \partial_y^2) + \alpha - \varepsilon\partial_y^2 k_x^2 \right) \\ & \times \left(\frac{1}{2}\partial_t + \sigma + \alpha - \varepsilon\partial_y^2 \right) v(\cdot, t) \Bigg) \Bigg\|^2 . \end{aligned}$$

For $\mathcal{E}^{III}$ we first rewrite

$$\begin{aligned} & \frac{1}{2\sigma}(\tau+\sigma)((\tau+\sigma)^3 + \sigma k_x^2)\left((\tau+\sigma)^2 k_y^2 + k_x^2\tau^2\right) \\ & \qquad = \frac{1}{2\sigma}(\tau+\sigma)^6 k_y^2 + \frac{1}{2\sigma}(\tau+\sigma)^4 k_x^2\tau^2 \\ & + \frac{\alpha+\sigma}{2}(\tau+\sigma)^2 k_x^2 k_y^2 + \frac{\alpha+\sigma}{2}\tau^2 k_x^4 + \frac{\varepsilon}{2}(\tau+\sigma)^2 k_x^2 k_y^4 + \frac{\varepsilon}{2}\tau^2 k_x^4 k_y^2. \end{aligned}$$

Integrating over k and applying Parseval to each term in $\mathcal{E}^{III}$, we get

$$\begin{aligned} E^{III} = & \frac{1}{2\sigma}\|(\alpha + \sigma - \varepsilon\partial_y^2)^3\partial_y \mathcal{F}^{-1}\{\hat{\chi}_2\}\|^2 \\ & + \frac{1}{2\sigma}\|(\alpha + \sigma - \varepsilon\partial_y^2)^2(\alpha - \varepsilon\partial_y^2)\partial_x \mathcal{F}^{-1}\{\hat{\chi}_2\}\|^2 \\ & + \frac{\sigma+\alpha}{2}\|(\alpha + \sigma - \varepsilon\partial_y^2)\partial_x\partial_y \mathcal{F}^{-1}\{\hat{\chi}_2\}\|^2 \\ & + \frac{\sigma+\alpha}{2}\|(\alpha - \varepsilon\partial_y^2)\partial_x^2 \mathcal{F}^{-1}\{\hat{\chi}_2\}\|^2 \\ & + \frac{\varepsilon}{2}\|(\alpha + \sigma - \varepsilon\partial_y^2)\partial_x\partial_y^2 \mathcal{F}^{-1}\{\hat{\chi}_2\}\|^2 \\ & + \frac{\varepsilon}{2}\|(\alpha - \varepsilon\partial_y^2)\partial_x^2\partial_y \mathcal{F}^{-1}\{\hat{\chi}_2\}\|^2, \end{aligned}$$

where

$$\mathcal{F}^{-1}\{\hat{\chi}_2\} = \left((\sigma + \alpha - \varepsilon\partial_y^2)\partial_t^2 + 2(\sigma + \alpha - \varepsilon\partial_y^2)^2\partial_t + (\sigma + \alpha - \varepsilon\partial_y^2)^3 + \sigma\partial_x^2\right) v(x,t).$$

In the same way we get for $\mathcal{E}^{II}$

$$\begin{aligned}
E^{II} = {} & 2\alpha\|(\alpha+\sigma-\varepsilon\partial_y^2)^3\partial_x\partial_y\mathcal{F}^{-1}\{\hat{\chi}_1\}\|^2 \\
& + 2\varepsilon\|(\alpha+\sigma-\varepsilon\partial_y^2)^3\partial_x\partial_y^2\mathcal{F}^{-1}\{\hat{\chi}_1\}\|^2 \\
& + 2(\alpha+\sigma)\|(\alpha+\sigma-\varepsilon\partial_y^2)^2\partial_x\partial_y^2\mathcal{F}^{-1}\{\hat{\chi}_1\}\|^2 \\
& + 2\varepsilon\|(\alpha+\sigma-\varepsilon\partial_y^2)^2\partial_x\partial_y^3\mathcal{F}^{-1}\{\hat{\chi}_1\}\|^2 \\
& + 2\alpha\|(\alpha+\sigma-\varepsilon\partial_y^2)^2\partial_x^2\partial_y\mathcal{F}^{-1}\{\hat{\chi}_1\}\|^2 \\
& + 2\varepsilon\|(\alpha+\sigma-\varepsilon\partial_y^2)^2\partial_x^2\partial_y^2\mathcal{F}^{-1}\{\hat{\chi}_1\}\|^2 \\
& + 2\alpha\|(\alpha+\sigma-\varepsilon\partial_y^2)^2(\alpha-\varepsilon\partial_y^2)\partial_x^2\mathcal{F}^{-1}\{\hat{\chi}_1\}\|^2 \\
& + 2\varepsilon\|(\alpha+\sigma-\varepsilon\partial_y^2)^2(\alpha-\varepsilon\partial_y^2)\partial_x^2\partial_y\mathcal{F}^{-1}\{\hat{\chi}_1\}\|^2 \\
& + 2(\alpha+\sigma)\|(\alpha+\sigma-\varepsilon\partial_y^2)(\alpha-\varepsilon\partial_y^2)\partial_x\partial_y\mathcal{F}^{-1}\{\hat{\chi}_1\}\|^2 \\
& + 2\varepsilon\|(\alpha+\sigma-\varepsilon\partial_y^2)(\alpha-\varepsilon\partial_y^2)\partial_x^2\partial_y^2\mathcal{F}^{-1}\{\hat{\chi}_1\}\|^2 \\
& + 2\alpha\|(\alpha+\sigma-\varepsilon\partial_y^2)(\alpha-\varepsilon\partial_y^2)\partial_x^3\mathcal{F}^{-1}\{\hat{\chi}_1\}\|^2 \\
& + 2\varepsilon\|(\alpha+\sigma-\varepsilon\partial_y^2)(\alpha-\varepsilon\partial_y^2)\partial_x^3\partial_y\mathcal{F}^{-1}\{\hat{\chi}_1\}\|^2,
\end{aligned}$$

where

$$\mathcal{F}^{-1}\{\hat{\chi}_1\} = \left(\frac{1}{2}\partial_t + \sigma + \alpha - \varepsilon\partial_y^2\right)v. \tag{89}$$

By similar operations, we can obtain an expression for E^I. However, since we have to split the factor in front of $|\hat{v}|^2$ in $\mathcal{E}^I$ into many terms, the expression for E^I becomes very lengthy and we have chosen not to include it here.

Appendix B.

$$\begin{aligned}
c_{2a} &= \sigma\zeta^2(\sigma+3M_x^2\alpha+\alpha)k_x^2 + M_x^2\sigma\zeta(\sigma+\zeta\alpha)k_y^2 \\
&\quad + (\sigma+\zeta\alpha)^2(\sigma+\alpha(1+M_x))(\sigma+\alpha(1-M_x)), \\
c_{3a} &= (c_{3b}+c_{3c}k_x^4+c_{3d}k_y^4+c_{3e}k_x^2k_y^2+c_{3f}k_x^2+c_{3g}k_y^2), \\
c_{3b} &= (-\zeta k_x^2+M_x^2k_y^2)^2(\sigma\alpha\zeta^3k_x^2+\sigma(\sigma+\alpha)\zeta^2k_y^2), \\
c_{3c} &= \alpha\zeta^3(5{M_x}^4\alpha^3+12\sigma{M_x}^2\alpha^2+10{M_x}^2\alpha^3+2\alpha\sigma^2{M_x}^2+5\alpha\sigma^2 \\
&\quad + 2\sigma^3+4\sigma\alpha^2+\alpha^3), \\
c_{3d} &= M_x^2(\sigma+\alpha)\zeta(\alpha\zeta+\sigma)(\zeta\alpha^2+2\sigma\alpha+\sigma^2+\sigma M_x^2\alpha), \\
c_{3e} &= -\zeta^2(\sigma^4+2\sigma^3\alpha(2+M_x^2)+\sigma^2\alpha^2(6+11M_x^2+M_x^4) \\
&\quad + 4\sigma\alpha^3(1+4M_x^2-M_x^6)+\alpha\eta(1+8M_x^2+3M_x^4)), \\
c_{3f} &= \alpha\zeta(\sigma+\alpha)(3\alpha^2M_x^2+(\alpha+\sigma)^2)(\alpha\zeta+\sigma)^2, \\
c_{3g} &= \alpha M_x^2(\sigma+\alpha)(\sigma+\alpha(1+M_x))(\sigma+\alpha(1-M_x))(\alpha\zeta+\sigma)^2, \\
c_{4a} &= \alpha^2\left(\sigma+\alpha\right)^2\left(\zeta\alpha+\sigma\right)^2 + c_{4b}k_x^4 + c_{4c}k_y^4 + c_{4d}k_x^2k_y^2 + c_{4e}k_x^2 + c_{4f}k_y^2, \\
c_{4b} &= \zeta^4\alpha^2, \quad c_{4c} = \zeta^2\left(\sigma+\alpha\right)^2, \quad c_{4d} = 2\zeta^3\alpha\left(\sigma+\alpha\right), \\
c_{4e} &= 2\alpha^2\left(\sigma+\alpha\right)\zeta^2\left(\alpha(1+{M_x}^2)+\sigma\right), \\
c_{4f} &= 2\alpha\zeta\left(\sigma+\alpha\right)^2\left(\alpha\zeta+\sigma\right).
\end{aligned}$$

Appendix C.

$$
\begin{aligned}
c_{4a} = \big(&4{k_x}^2{a_{11}}^4{a_{22}}^2 + \sigma^2{a_{11}}^4 - 4{k_y}^2{b_{12}}^2{a_{11}}^3a_{22} - 8k_yb_{22}{a_{11}}^3k_xa_{22}{}^2 \\
&+ 8k_xa_{11}{}^3a_{22}{}^2k_yb_{11} - 4\sigma^2{a_{11}}^3a_{22} - 8{k_x}^2{a_{11}}^3a_{22}{}^3 + 6\sigma^2a_{22}{}^2{a_{11}}^2 \\
&+ 4{k_x}^2{a_{11}}^2a_{22}{}^4 - 8{k_y}^2b_{22}{a_{11}}^2b_{11}a_{22}{}^2 - 8{k_y}^2{b_{12}}^2{a_{11}}^2a_{22}{}^2 + 8k_xa_{11}{}^2a_{22}{}^3k_yb_{22} \\
&+ 4{k_y}^2{b_{11}}^2a_{22}{}^2{a_{11}}^2 + 4{k_y}^2b_{22}{}^2{a_{11}}^2a_{22}{}^2 - 8k_xa_{11}{}^2a_{22}{}^3k_yb_{11} - 4{k_y}^2{b_{12}}^2a_{11}a_{22}{}^3 \\
&- 4\sigma^2a_{22}{}^3a_{11} + \sigma^2a_{22}{}^4\big), \\
c_{4b} = \big(&a_{22}{}^2{b_{12}}^2{k_y}^2 + {b_{11}}^2a_{22}a_{11}{k_y}^2 + 2a_{22}a_{11}{k_y}^2{b_{12}}^2 - 2b_{11}a_{22}a_{11}{k_y}^2b_{22} \\
&+ a_{22}a_{11}{k_y}^2b_{22}{}^2 + {b_{12}}^2{k_y}^2{a_{11}}^2 - 2b_{11}k_ya_{11}a_{22}{}^2k_x + 2k_ya_{11}a_{22}{}^2k_xb_{22} \\
&+ 2b_{11}k_ya_{11}{}^2a_{22}k_x - 2k_ya_{11}{}^2a_{22}k_xb_{22} + a_{22}{}^3a_{11}k_x{}^2 - 2a_{11}{}^2a_{22}{}^2k_x{}^2 + a_{22}a_{11}{}^3k_x{}^2\big).
\end{aligned}
$$

REFERENCES

[1] S. Abarbanel, D. Gottlieb, and J. Hesthaven, *Well-posed perfectly matched layers for advective acoustics*, J. Comput. Phys., 154 (1999), pp. 266–283.

[2] S. Abarbanel and D. Gottlieb, *A mathematical analysis of the PML method*, J. Comput. Phys., 134 (1997), p. 357–363.

[3] S. Abarbanel and D. Gottlieb, *On the construction and analysis of absorbing layers in CEM*, Appl. Numer. Math., 27 (1998), pp. 331–340.

[4] D. Appelö and T. Hagstrom, *Construction of stable PMLs for general 2×2 symmetric hyperbolic systems*, in Proceedings of the Tenth International Conference on Hyperbolic Problems: Theory, Numerics, Applications, F. Asakura, H. Aiso, S. Kawashima, A. Matsumura, S. Nishibata, and K. Nishihara, eds., Yokahama Publishers, Japan, 2004, pp. 262–270.

[5] E. Bécache, S. Fauqueux, and P. Joly, *Stability of perfectly matched layers, group velocities and anisotropic waves*, J. Comput. Phys., 188 (2003), pp. 399–433.

[6] E. Bécache and P. Joly, *On the analysis of Berenger's perfectly matched layers for Maxwell's equations*, Math. Model Numer. Anal., 36 (2002), pp. 87–119.

[7] E. Bécache, P. Petropoulos, and S. Gedney, *On the long-time behavior of unsplit Perfectly Matched Layers*, IEEE Trans. Antennas Prop., 54 (2004), pp. 1335–1342.

[8] J. Bérenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), p. 185.

[9] W. Chew and W. Weedon, *A 3-D perfectly matched medium from modified Maxwell's equations with stretched coordinates*, Microwave Optical Technol. Lett., 7 (1994), pp. 599–604.

[10] F. Collino and C. Tsogka, *Application of the PML absorbing layer model to the linear elastodynamic problem in aniotropic heterogeneous media*, Geophys., 66 (2001), pp. 294–307.

[11] J. Diaz and P. Joly, *Stabilized perfectly matched layer for advective acoustics*, in Mathematical and Numerical Aspects of Wave Propagation, Proceedings of Waves2003, G. Cohen, P. Joly, E. Heikkola, and P. Neittaanmäki, eds., Springer-Verlag, New York, 2003, pp. 115–119.

[12] F. R. Gantmacher, *The Theory of Matrices*, Vol. 1, Chelsea, New York, 1959.

[13] S. Gedney, *An anisotropic perfectly matched layer-absorbing medium for the truncation of fdtd lattices.*, IEEE Trans. Antennas Propagation, 44 (1996), pp. 1630–1639.

[14] T. Hagstrom and D. Appelö, *Automatic Symmetrization and Energy Estimates Using Local Operators for Partial Differential Equations.* Communications in Partial Differential Equations, in press.

[15] T. Hagstrom, *Radiation boundary conditions for the numerical simulation of waves*, Acta. Numer., 8 (1999), pp. 47–106.

[16] T. Hagstrom, *New Results on Absorbing Layers and Radiation Boundary Conditions*, in Topics in Computational Wave Propagation, Lecture Notes in Comput. Sci. Engrg. 31, M. Ainsworth, P. Davies, D. Duncan, P. Martin, and B, Rynne, eds., Springer-Verlag, New York, 2003, pp. 1–42.

[17] T. Hagstrom, *Perfectly matched layers for hyperbolic systems with applications to the linearized Euler equations*, in Mathematical and Numerical Aspects of Wave Propagation, Proceedings of Waves2003, G. Cohen, P. Joly, E. Heikkola, and P. Neittaanmäki, eds., Springer-Verlag, New York, 2003, pp. 125–129.

[18] F. Q. Hu, *On absorbing boundary conditions for linearized Euler equations by a perfectly matched layer*, J. Comput. Phys., 129 (1996), pp. 201–209.
[19] F. Q. Hu, *A stable perfectly matched layer for linearized Euler equations in unsplit physical variables*, J. Comput. Phys., 173 (2001), pp. 455–480.
[20] H.-O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier-Stokes Equations*, Academic Press, New York, 1989.
[21] M. Marden, *The Geometry of the Zeros of a Polynomial in a Complex Variable*, American Mathematical Society, Providence, RI, 1949.
[22] I. Navon, B. Neta, and M. Hussaini, *A perfectly matched layer approach to the linearized shallow water equations models*, Monthly Weather Rev., 132 (2004), pp 1369–1378.
[23] S. Tsynkov, *Numerical solution of problems on unbounded domains, A review*, Appl. Numer. Math., 27 (1998), pp. 465–532.
[24] L. Zhao and A. Cangellaris, *GT-PML: Generalized theory of perfectly matched layers and its application to reflectionless truncation of finite-difference time-domain grids*, IEEE Trans. Microwave Theory Tech., 44 (1996), pp. 2555–2563.

SIAM J. APPL. MATH.
Vol. 67, No. 1, pp. 24–45

BIFURCATION ANALYSIS OF A MATHEMATICAL MODEL FOR MALARIA TRANSMISSION*

NAKUL CHITNIS†, J. M. CUSHING‡, AND J. M. HYMAN§

Abstract. We present an ordinary differential equation mathematical model for the spread of malaria in human and mosquito populations. Susceptible humans can be infected when they are bitten by an infectious mosquito. They then progress through the exposed, infectious, and recovered classes, before reentering the susceptible class. Susceptible mosquitoes can become infected when they bite infectious or recovered humans, and once infected they move through the exposed and infectious classes. Both species follow a logistic population model, with humans having immigration and disease-induced death. We define a reproductive number, R_0, for the number of secondary cases that one infected individual will cause through the duration of the infectious period. We find that the disease-free equilibrium is locally asymptotically stable when $R_0 < 1$ and unstable when $R_0 > 1$. We prove the existence of at least one endemic equilibrium point for all $R_0 > 1$. In the absence of disease-induced death, we prove that the transcritical bifurcation at $R_0 = 1$ is supercritical (forward). Numerical simulations show that for larger values of the disease-induced death rate, a subcritical (backward) bifurcation is possible at $R_0 = 1$.

Key words. malaria, epidemic model, reproductive number, bifurcation theory, endemic equilibria, disease-free equilibria

AMS subject classifications. Primary, 92D30; Secondary, 37N25

DOI. 10.1137/050638941

1. Introduction.

Malaria is an infectious disease caused by the *Plasmodium* parasite and transmitted between humans through the bite of the female *Anopheles* mosquito. An estimated 40% of the world's population live in malaria endemic areas. The disease kills about 1 to 3 million people a year, 75% of whom are African children. The incidence of malaria has been growing recently due to increasing parasite drug-resistance and mosquito insecticide-resistance. Therefore, it is important to understand the important parameters in the transmission of the disease and develop effective solution strategies for its prevention and control.

Mathematical modeling of malaria began in 1911 with Ross's model [25], and major extensions are described in Macdonald's 1957 book [20]. The first models were two-dimensional with one variable representing humans and the other representing mosquitoes. An important addition to the malaria models was the inclusion of acquired immunity proposed by Dietz, Molineaux, and Thomas [11]. Further work on acquired immunity in malaria has been conducted by Aron [2] and Bailey [5]. Anderson and May [1], Aron and May [3], Koella [15] and Nedelman [21] have written some good reviews on the mathematical modeling of malaria. Some recent papers have also included environmental effects [19], [27], and [28]; the spread of resistance to drugs

*Received by the editors August 25, 2005; accepted for publication (in revised form) June 30, 2006; published electronically November 3, 2006. The authors thank the United States National Science Foundation for the following grants: NSF DMS-0414212 and NSF DMS-0210474. This research has also been supported under Department of Energy contract W-7405-ENG-36. Analysis of a similar model was published in the Ph.D. dissertation of the first author; see [7].

http://www.siam.org/journals/siap/67-1/63894.html

†Corresponding author. Department of Public Health and Epidemiology, Swiss Tropical Institute, Socinstrasse 57, P. O. Box, CH-4002 Basel, Switzerland (Nakul.Chitnis@unibas.ch).

‡Department of Mathematics, University of Arizona, Tucson, AZ 85721 (cushing@math.arizona.edu).

§Mathematical Modeling and Analysis (T-7), Los Alamos National Laboratory, Los Alamos, NM 87545 (hyman@lanl.gov).

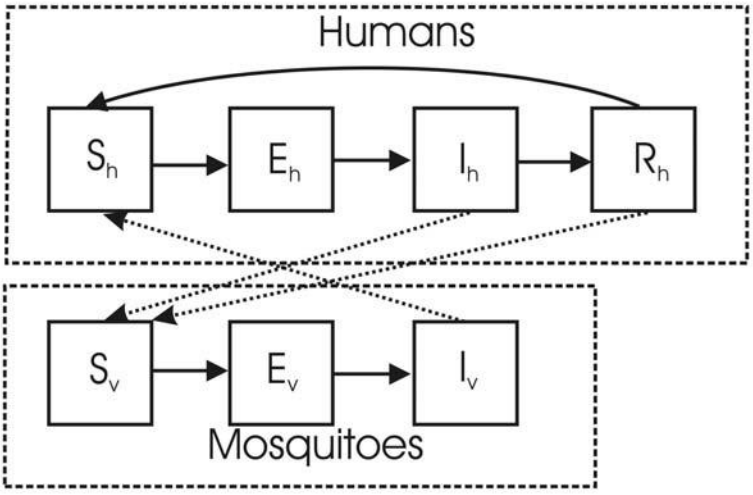


FIG. 1.1. *Susceptible humans, S_h, can be infected when they are bitten by infectious mosquitoes. They then progress through the exposed, E_h, infectious, I_h, and recovered, R_h, classes, before re-entering the susceptible class. Susceptible mosquitoes, S_v, can become infected when they bite infectious or recovered humans. The infected mosquitoes then move through the exposed, E_v, and infectious, I_v, classes. Both species follow a logistic population model, with humans having additional immigration and disease-induced death. Birth, death, and migration into and out of the population are not shown in the figure.*

[4] and [16]; and the evolution of immunity [17].

Recently, Ngwa and Shu [23] and Ngwa [22] proposed an ordinary differential equation (ODE) compartmental model for the spread of malaria with a susceptible-exposed-infectious-recovered-susceptible (SEIRS) pattern for humans and a susceptible-exposed-infectious (SEI) pattern for mosquitoes. In a Ph.D. dissertation, Chitnis [7] analyzed a similar model for malaria transmission. In this paper we extend the Chitnis model.

The new model (Figure 1.1) divides the human population into four classes: susceptible, S_h; exposed, E_h; infectious, I_h; and recovered (immune), R_h. People enter the susceptible class either through birth (at a constant per capita rate) or through immigration (at a constant rate). When an infectious mosquito bites a susceptible human, there is some finite probability that the parasite (in the form of sporozoites) will be passed on to the human and that the person will move to the exposed class. The parasite then travels to the liver where it develops into its next life stage. After a certain period of time, the parasite (in the form of merozoites) enters the blood stream, usually signaling the clinical onset of malaria. In our model, people from the exposed class enter the infectious class at a rate that is the reciprocal of the duration of the latent period. After some time, the infectious humans recover and move to the recovered class. The recovered humans have some immunity to the disease and do not get clinically ill, but they still harbor low levels of parasite in their blood streams and can pass the infection to mosquitoes. After some period of time, they lose their immunity and return to the susceptible class. Humans leave the population through a density-dependent per capita emigration and natural death rate, and through a per capita disease-induced death rate.

We divide the mosquito population into three classes: susceptible, S_v; exposed,

E_v; and infectious, I_v. Female mosquitoes (we do not include male mosquitoes in our model because only female mosquitoes bite animals for blood meals) enter the susceptible class through birth. The parasite (in the form of gametocytes) enters the mosquito with some probability when the mosquito bites an infectious human or a recovered human (the probability of transmission of infection from a recovered human is much lower than that from an infectious human), and the mosquito moves from the susceptible to the exposed class. After some period of time, dependent on the ambient temperature and humidity, the parasite develops into sporozoites and enters the mosquito's salivary glands, and the mosquito moves from the exposed class to the infectious class. The mosquito remains infectious for life. Mosquitoes leave the population through a per capita density-dependent natural death rate.

The extension of the Ngwa and Shu model [23] includes human immigration, excludes direct human recovery from the infectious to the susceptible class, and generalizes the mosquito biting rate so that it applies to wider ranges of populations. In [23], the total number of mosquito bites on humans depends only on the number of mosquitoes, while in our model, the total number of bites depends on both the human and mosquito population sizes. Human migration is present throughout the world and plays a large role in the epidemiology of diseases, including malaria. In many parts of the developing world, there is rapid urbanization as many people leave rural areas and migrate to cities in search of employment. We include this movement as a constant immigration rate into the susceptible class. We do not include immigration of infectious humans, as we assume that most people who are sick will not travel. We also exclude the movement of exposed humans because, given the short time of the exposed stage, the number of exposed people is small. We make the simplifying assumption that there is no immigration of recovered humans. We also exclude the direct infectious-to-susceptible recovery that the model of Ngwa and Shu [23] contains. This is a realistic simplifying assumption because most people show some period of immunity before becoming susceptible again. As our model includes an exponential distribution of movement from the recovered to the susceptible class, it will include the quick return to susceptibility of some individuals. The model in Chitnis [7] is the same as the model in this paper except for the mosquito biting rate, which is the same as in [23].

We first describe the mathematical model including the definition of a domain where the model is mathematically and epidemiologically well-posed. Next, we prove the existence and stability of a disease-free equilibrium point, define the reproductive number, and describe the existence and stability of the endemic equilibrium point(s).

2. Malaria model. The state variables (Table 2.1) and parameters (Table 2.2) for the malaria model (Figure 1.1) satisfy the equations in (2.1). All parameters

TABLE 2.1
The state variables for the malaria model (2.1).

S_h:	Number of susceptible humans
E_h:	Number of exposed humans
I_h:	Number of infectious humans
R_h:	Number of recovered (immune and asymptomatic, but slightly infectious) humans
S_v:	Number of susceptible mosquitoes
E_v:	Number of exposed mosquitoes
I_v:	Number of infectious mosquitoes
N_h:	Total human population
N_v:	Total mosquito population

TABLE 2.2
The parameters for the malaria model (2.1) *and their dimensions.*

Λ_h:	Immigration rate of humans. Humans $\times$ Time^{-1}.
ψ_h:	Per capita birth rate of humans. Time^{-1}.
ψ_v:	Per capita birth rate of mosquitoes. Time^{-1}.
σ_v:	Number of times one mosquito would want to bite humans per unit time, if humans were freely available. This is a function of the mosquito's gonotrophic cycle (the amount of time a mosquito requires to produce eggs) and its anthropophilic rate (its preference for human blood). Time^{-1}.
σ_h:	The maximum number of mosquito bites a human can have per unit time. This is a function of the human's exposed surface area. Time^{-1}.
β_{hv}:	Probability of transmission of infection from an infectious mosquito to a susceptible human, given that a contact between the two occurs. Dimensionless.
β_{vh}:	Probability of transmission of infection from an infectious human to a susceptible mosquito, given that a contact between the two occurs. Dimensionless.
$\tilde{\beta}_{vh}$:	Probability of transmission of infection from a recovered (asymptomatic carrier) human to a susceptible mosquito, given that a contact between the two occurs. Dimensionless.
ν_h:	Per capita rate of progression of humans from the exposed state to the infectious state. $1/\nu_h$ is the average duration of the latent period. Time^{-1}.
ν_v:	Per capita rate of progression of mosquitoes from the exposed state to the infectious state. $1/\nu_v$ is the average duration of the latent period. Time^{-1}.
γ_h:	Per capita recovery rate for humans from the infectious state to the recovered state. $1/\gamma_h$ is the average duration of the infectious period. Time^{-1}.
δ_h:	Per capita disease-induced death rate for humans. Time^{-1}.
ρ_h:	Per capita rate of loss of immunity for humans. $1/\rho_h$ is the average duration of the immune period. Time^{-1}.
μ_{1h}:	Density-independent part of the death (and emigration) rate for humans. Time^{-1}.
μ_{2h}:	Density-dependent part of the death (and emigration) rate for humans. Humans^{-1} $\times$ Time^{-1}.
μ_{1v}:	Density-independent part of the death rate for mosquitoes. Time^{-1}.
μ_{2v}:	Density-dependent part of the death rate for mosquitoes. Mosquitoes^{-1} $\times$ Time^{-1}.

are strictly positive with the exception of the disease-induced death rate, δ_h, which is nonnegative. The mosquito birth rate is greater than the density-independent mosquito death rate, $\psi_v > \mu_{1v}$, ensuring that we have a stable positive mosquito population.

$$\frac{dS_h}{dt} = \Lambda_h + \psi_h N_h + \rho_h R_h - \lambda_h(t) S_h - f_h(N_h) S_h, \tag{2.1a}$$

$$\frac{dE_h}{dt} = \lambda_h(t) S_h - \nu_h E_h - f_h(N_h) E_h, \tag{2.1b}$$

$$\frac{dI_h}{dt} = \nu_h E_h - \gamma_h I_h - f_h(N_h) I_h - \delta_h I_h, \tag{2.1c}$$

$$\frac{dR_h}{dt} = \gamma_h I_h - \rho_h R_h - f_h(N_h) R_h, \tag{2.1d}$$

$$\frac{dS_v}{dt} = \psi_v N_v - \lambda_v(t) S_v - f_v(N_v) S_v, \tag{2.1e}$$

$$\frac{dE_v}{dt} = \lambda_v(t) S_v - \nu_v E_v - f_v(N_v) E_v, \tag{2.1f}$$

$$\frac{dI_v}{dt} = \nu_v E_v - f_v(N_v) I_v, \tag{2.1g}$$

where $f_h(N_h) = \mu_{1h} + \mu_{2h} N_h$ is the per capita density-dependent death and emigration rate for humans and $f_v(N_v) = \mu_{1v} + \mu_{2v} N_v$ is the per capita density-dependent death rate for mosquitoes. The total population sizes are $N_h = S_h + E_h + I_h + R_h$ and

$N_v = S_v + E_v + I_v$, with

$$\frac{dN_h}{dt} = \Lambda_h + \psi_h N_h - f_h(N_h)N_h - \delta_h I_h, \tag{2.2a}$$

$$\frac{dN_v}{dt} = \psi_v N_v - f_v(N_v)N_v, \tag{2.2b}$$

and the infection rates are

$$\lambda_h = b_h(N_h, N_v) \cdot \beta_{hv} \cdot \frac{I_v}{N_v} \quad \text{and} \quad \lambda_v = b_v(N_h, N_v) \cdot \left(\beta_{vh} \cdot \frac{I_h}{N_h} + \tilde{\beta}_{vh} \cdot \frac{R_h}{N_h}\right). \tag{2.3}$$

We define the force of infection from mosquitoes to humans, λ_h, as the product of the number of mosquito bites that one human has per unit time, b_h, the probability of disease transmission from the mosquito to the human, β_{hv}, and the probability that the mosquito is infectious, I_v/N_v. We define the force of infection from humans to mosquitoes, λ_v, as the sum of the force of infection from infectious humans and from recovered humans. These are defined as the number of human bites one mosquito has per unit time, b_v; the probability of disease transmission from the human to the mosquito, β_{vh} and $\tilde{\beta}_{vh}$; and the probability that the human is infectious or recovered, I_h/N_h and R_h/N_h. Here, we model the total number of mosquito bites on humans as

$$b = b(N_h, N_v) = \frac{\sigma_v N_v \sigma_h N_h}{\sigma_v N_v + \sigma_h N_h} = \frac{\sigma_v \sigma_h}{\sigma_v (N_v/N_h) + \sigma_h} N_v, \tag{2.4}$$

so that the total number of mosquito-human contacts depends on the populations of both species. We define $b_h = b_h(N_h, N_v) = b(N_h, N_v)/N_h$ as the number of bites per human per unit time, and $b_v = b_v(N_h, N_v) = b(N_h, N_v)/N_v$ as the number of bites per mosquito per unit time. In the limit that the mosquito population goes to zero or the human population goes to infinity, the model reduces to that in Chitnis [7] and has the same mosquito-human interaction as in Ngwa and Shu [23] and the Ross–Macdonald model (as described by Anderson and May [1]), where the total number of bites is limited by the mosquito population. The number of bites per mosquito is then σ_v (denoted by σ_{vh} in [7]), and the number of bites per human is $\sigma_v N_v/N_h$. We show a summary of the model of mosquito-human interactions and its limits in Table 2.3.

TABLE 2.3
Number of mosquito bites on humans in the malaria transmission model (2.1) *and its limiting cases with population changes.*

	Number of bites per human, b_h	Number of bites per mosquito, b_v	Total number of bites, b
General model	$\frac{\sigma_v N_v \sigma_h}{\sigma_v N_v + \sigma_h N_h}$	$\frac{\sigma_v \sigma_h N_h}{\sigma_v N_v + \sigma_h N_h}$	$\frac{\sigma_v N_v \sigma_h N_h}{\sigma_v N_v + \sigma_h N_h}$
As $N_h \to \infty$ or $N_v \to 0$	$\frac{\sigma_v N_v}{N_h}$	σ_v	$\sigma_v N_v$
As $N_h \to 0$ or $N_v \to \infty$	σ_h	$\frac{\sigma_h N_h}{N_v}$	$\sigma_h N_h$

To simplify the analysis of the malaria model (2.1), we work with fractional quantities instead of actual populations by scaling the population of each class by the

total species population. We let

$$e_h = \frac{E_h}{N_h}, \quad i_h = \frac{I_h}{N_h}, \quad r_h = \frac{R_h}{N_h}, \quad e_v = \frac{E_v}{N_v}, \quad \text{and} \quad i_v = \frac{I_v}{N_v}, \tag{2.5}$$

with

$$S_h = s_h N_h = (1 - e_h - i_h - r_h)N_h \quad \text{and} \quad S_v = s_v N_v = (1 - e_v - i_v)N_v. \tag{2.6}$$

Differentiating the scaling equations (2.5) and solving for the derivatives of the scaled variables, we obtain

$$\frac{de_h}{dt} = \frac{1}{N_h}\left[\frac{dE_h}{dt} - e_h\frac{dN_h}{dt}\right] \quad \text{and} \quad \frac{de_v}{dt} = \frac{1}{N_v}\left[\frac{dE_v}{dt} - e_v\frac{dN_v}{dt}\right] \tag{2.7}$$

and so on for the other variables.

This creates a new seven-dimensional system of equations with two dimensions for the two total population variables, N_h and N_v, and five dimensions for the fractional population variables with disease, e_h, i_h, r_h, e_v, and i_v:

$$\frac{de_h}{dt} = \left(\frac{\sigma_v\sigma_h N_v\beta_{hv}i_v}{\sigma_v N_v + \sigma_h N_h}\right)(1 - e_h - i_h - r_h) - \left(\nu_h + \psi_h + \frac{\Lambda_h}{N_h}\right)e_h + \delta_h i_h e_h, \tag{2.8a}$$

$$\frac{di_h}{dt} = \nu_h e_h - \left(\gamma_h + \delta_h + \psi_h + \frac{\Lambda_h}{N_h}\right)i_h + \delta_h i_h^2, \tag{2.8b}$$

$$\frac{dr_h}{dt} = \gamma_h i_h - \left(\rho_h + \psi_h + \frac{\Lambda_h}{N_h}\right)r_h + \delta_h i_h r_h, \tag{2.8c}$$

$$\frac{dN_h}{dt} = \Lambda_h + \psi_h N_h - (\mu_{1h} + \mu_{2h}N_h)N_h - \delta_h i_h N_h, \tag{2.8d}$$

$$\frac{de_v}{dt} = \left(\frac{\sigma_v\sigma_h N_h}{\sigma_v N_v + \sigma_h N_h}\right)\left(\beta_{vh}i_h + \tilde{\beta}_{vh}r_h\right)(1 - e_v - i_v) - (\nu_v + \psi_v)e_v, \tag{2.8e}$$

$$\frac{di_v}{dt} = \nu_v e_v - \psi_v i_v, \tag{2.8f}$$

$$\frac{dN_v}{dt} = \psi_v N_v - (\mu_{1v} + \mu_{2v}N_v)N_v. \tag{2.8g}$$

The model (2.8) is epidemiologically and mathematically well-posed in the domain

$$\mathcal{D} = \left\{ \begin{pmatrix} e_h \\ i_h \\ r_h \\ N_h \\ e_v \\ i_v \\ N_v \end{pmatrix} \in \mathbb{R}^7 \;\middle|\; \begin{array}{c} e_h \geq 0, \\ i_h \geq 0, \\ r_h \geq 0, \\ e_h + i_h + r_h \leq 1, \\ N_h > 0, \\ e_v \geq 0, \\ i_v \geq 0, \\ e_v + i_v \leq 1, \\ N_v > 0 \end{array} \right\}. \tag{2.9}$$

This domain, $\mathcal{D}$, is valid epidemiologically as the fractional populations e_h, i_h, r_h, e_v, and i_v are all nonnegative and have sums over their species type that are less than or equal to 1. The human and mosquito populations, N_h and N_v, are positive. We use the notation f' to denote df/dt. We denote points in $\mathcal{D}$ by $x = (e_h, i_h, r_h, N_h, e_v, i_v, N_v)$.

THEOREM 2.1. *Assuming that the initial conditions lie in $\mathcal{D}$, the system of equations for the malaria model* (2.8) *has a unique solution that exists and remains in $\mathcal{D}$ for all time $t \geq 0$.*

Proof. The right-hand side of (2.8) is continuous with continuous partial derivatives in $\mathcal{D}$, so (2.8) has a unique solution. We now show that $\mathcal{D}$ is forward-invariant. We can see from (2.8) that if $e_h = 0$, then $e_h' \geq 0$; if $i_h = 0$, then $i_h' \geq 0$; if $r_h = 0$, then $r_h' \geq 0$; if $e_v = 0$, then $e_v' \geq 0$; and if $i_v = 0$, then $i_v' \geq 0$. It is also true that if $e_h + i_h + r_h = 1$, then $e_h' + i_h' + r_h' < 0$; and if $e_v + i_v = 1$, then $e_v' + i_v' < 0$. Finally, we note that if $N_h = 0$, then $N_h' > 0$ and if $N_v = 0$, then $N_v' = 0$. If $N_h > 0$ at time $t = 0$, then $N_h > 0$ for all $t > 0$. Similarly, if $N_v > 0$ at time $t = 0$, then $N_v > 0$ for all $t > 0$. Therefore, none of the orbits can leave $\mathcal{D}$, and a unique solution exists for all time. ∎

3. Disease-free equilibrium point and reproductive number.

3.1. Existence of the disease-free equilibrium point. Disease-free equilibrium points are steady-state solutions where there is no disease. We define the "*diseased*" classes as the human or mosquito populations that are either exposed, infectious, or recovered, that is, e_h, i_h, r_h, e_v, and i_v. We denote the positive orthant in $\mathbb{R}^7$ by $\mathbb{R}^7_+$, and the boundary of $\mathbb{R}^7_+$ by $\partial\mathbb{R}^7_+$. The positive equilibrium human and mosquito population values, in the absence of disease, for (2.8) are

$$N_h^* = \frac{(\psi_h - \mu_{1h}) + \sqrt{(\psi_h - \mu_{1h})^2 + 4\mu_{2h}\Lambda_h}}{2\mu_{2h}} \quad \text{and} \quad N_v^* = \frac{\psi_v - \mu_{1v}}{\mu_{2v}}. \tag{3.1}$$

THEOREM 3.1. *The malaria model* (2.8) *has exactly one equilibrium point, $x_{dfe} = (0, 0, 0, N_h^*, 0, 0, N_v^*)$, with no disease in the population (on $\mathcal{D} \cap \partial\mathbb{R}^7_+$).*

Proof. We need to show that x_{dfe} is an equilibrium point of (2.8) and that there are no other equilibrium points on $\mathcal{D} \cap \partial\mathbb{R}^7_+$. Substituting x_{dfe} into (2.8) shows all derivatives equal to zero, so x_{dfe} is an equilibrium point. We know from Lemma A.1 that on $\mathcal{D} \cap \partial\mathbb{R}^7_+$, $e_h = i_h = r_h = e_v = i_v = 0$. For $i_h = 0$, the only equilibrium point for N_h from (2.8d) is N_h^*, and the only equilibrium point for N_v in $\mathcal{D}$ from (2.8g) is N_v^*. Thus, the only equilibrium point on $\mathcal{D} \cap \partial\mathbb{R}^7_+$ is x_{dfe}. ∎

3.2. Reproductive number. We use the next generation operator approach as described by Diekmann, Heesterbeek, and Metz in [10] to define the reproductive number, R_0, as the number of secondary infections that one infectious individual would create over the duration of the infectious period, provided that everyone else is susceptible. We define the next generation operator, K, which provides the number of secondary infections in humans and mosquitoes caused by one generation of infectious humans and mosquitoes, as

$$K = \begin{pmatrix} 0 & K_{hv} \\ K_{vh} & 0 \end{pmatrix}, \tag{3.2}$$

where we use the following definitions:

K_{hv}: The number of humans that one mosquito infects through its infectious lifetime, assuming all humans are susceptible.

K_{vh}: The number of mosquitoes that one human infects through the duration of the infectious period, assuming all mosquitoes are susceptible.

Using the ideas of Hyman and Li [14], we define K_{hv} and K_{vh} as products of the probability of surviving till the infectious state, the number of contacts per unit

time, the probability of transmission per contact, and the duration of the infectious period:

$$K_{hv} = \left(\frac{\nu_v}{\nu_v + \mu_{1v} + \mu_{2v}N_v^*}\right) \cdot b_v^* \cdot \beta_{hv} \cdot \left(\frac{1}{\mu_{1v} + \mu_{2v}N_v^*}\right), \tag{3.3a}$$

$$\begin{aligned} K_{vh} = &\left(\frac{\nu_h}{\nu_h + \mu_{1h} + \mu_{2h}N_h^*}\right) \cdot b_h^* \cdot \beta_{vh} \cdot \left(\frac{1}{\gamma_h + \delta_h + \mu_{1h} + \mu_{2h}N_h^*}\right) \\ &+ \left(\frac{\nu_h}{\nu_h + \mu_{1h} + \mu_{2h}N_h^*} \cdot \frac{\gamma_h}{\gamma_h + \delta_h + \mu_{1h} + \mu_{2h}N_h^*}\right) \\ &\cdot b_h^* \cdot \tilde{\beta}_{vh} \cdot \left(\frac{1}{\rho_h + \mu_{1h} + \mu_{2h}N_h^*}\right). \end{aligned} \tag{3.3b}$$

In (3.3a), $\nu_v/(\nu_v + \mu_{1v} + \mu_{2v}N_v^*)$ is the probability that a mosquito will survive the exposed state to become infectious;[1] $b_v^* = b_v(N_h^*, N_v^*)$ is the number of contacts that one mosquito has with humans per unit time; and $1/(\mu_{1v} + \mu_{2v}N_v^*)$ is the average duration of the infectious lifetime of the mosquito. In (3.3b), the total number of mosquitoes infected by one human is the sum of the new infections from the infectious and from the recovered states of the human; $\nu_h/(\nu_h + \mu_{1h} + \mu_{2h}N_h^*)$ is the probability that a human will survive the exposed state to become infectious; $\gamma_h/(\gamma_h + \delta_h + \mu_{1h} + \mu_{2h}N_h^*)$ is the probability that the human will then survive the infectious state to move to the recovered state; $b_h^* = b_h(N_h^*, N_v^*)$ is the number of contacts that one human has with mosquitoes per unit time; $1/(\gamma_h + \delta_h + \mu_{1h} + \mu_{2h}N_h^*)$ is the average duration of the infectious period of a human; and $1/(\rho_h + \mu_{1h} + \mu_{2h}N_h^*)$ is the average duration of the recovered period of a human.

We define R_0 as the spectral radius of the next generation operator, K, i.e., $R_0^2 = K_{vh}K_{hv}$. Then, R_0^2 is the number of humans that one infectious human will infect, through a generation of infections in mosquitoes, assuming that previously all other humans and mosquitoes were susceptible.

DEFINITION 3.2. *We define the reproductive number, R_0, as*

$$R_0 = \sqrt{K_{vh}K_{hv}}, \tag{3.4}$$

where K_{vh} and K_{hv} are defined in (3.3).

The original definition of the reproductive number of the Ross–Macdonald model [1] and [3], and the Ngwa and Shu model [23], is equivalent to the square of this R_0. They ([1], [3], and [23]) use the traditional definition of the reproductive number, which approximates the number of secondary infections in humans caused by one infected human, while the R_0 in Definition 3.2 is consistent with the definition given by the next generation operator approach [10], which approximates the number of secondary infections due to one infected individual (be it human or mosquito). Our definition of R_0 includes the generation of infections in mosquitoes, so is the square root of the original definition. The threshold condition for both definitions is the same.

3.3. Stability of the disease-free equilibrium point.

THEOREM 3.3. *The disease-free equilibrium point, x_{dfe}, is locally asymptotically stable if $R_0 < 1$ and unstable if $R_0 > 1$.*

The proof of this theorem is in the appendix section A.1.

[1] In defining periods of time and probabilities for R_0, we use the original system of equations (2.1) and not the scaled equations (2.8). As the two models are equivalent, the reproductive number is the same with either definition: $\mu_{1h} + \mu_{2h}N_h^* = \psi_h + \Lambda_h/N_h^*$ and $\mu_{1v} + \mu_{2v}N_v^* = \psi_v$.

4. Endemic equilibrium points. Endemic equilibrium points are steady-state solutions where the disease persists in the population (all state variables are positive). We use general bifurcation theory to prove the existence of at least one endemic equilibrium point for all $R_0 > 1$. We prove that the transcritical bifurcation at $R_0 = 1$ is supercritical (forward) when $\delta_h = 0$ (there is no disease-induced death). However, numerical results show that the bifurcation can be subcritical (backward) for some positive values of δ_h, giving rise to endemic equilibria for $R_0 < 1$.

We first rewrite the equilibrium equations for $u = (e_h, e_v)$ in (2.8) as a nonlinear eigenvalue problem in a Banach space:

$$u = G(\zeta, u) = \zeta L u + h(\zeta, u), \tag{4.1}$$

where $u \in Y \subset \mathbb{R}^2$, with Euclidean norm $\|\cdot\|$; $\zeta \in Z \subset \mathbb{R}$ is the bifurcation parameter; L is a compact linear map on Y; and $h(\zeta, u)$ is $\mathcal{O}(\|u\|^2)$ uniformly on bounded ζ intervals. We require that both Y and Z be open and bounded sets, and that Y contain the point 0. We define Z as the open and bounded set $Z = \{\zeta \in \mathbb{R} | -M_Z < \zeta < M_Z\}$. This set is defined to include the characteristic values (reciprocals of eigenvalues) of L, so there is minimum value that M_Z can have, but M_Z may be arbitrarily large. We use

$$\zeta = \frac{\sigma_v \sigma_h}{\sigma_v N_v^* + \sigma_h N_h^*} \tag{4.2}$$

for the bifurcation parameter. We also define $\Omega = Z \times Y$ so that the pair $(\zeta, u) \in \Omega$. We denote the boundary of Ω by $\partial\Omega$.

A corollary by Rabinowitz [24, Corollary 1.12] states that if $\zeta_0 \in Z$ is a characteristic value of L of odd multiplicity, then there exists a continuum of nontrivial solution-pairs (ζ, u) of (4.1) that intersects the trivial solution (that is, $(\zeta, 0)$ for any ζ) at $(\zeta_0, 0)$ and either meets $\partial\Omega$ or meets $(\hat{\zeta}_0, 0)$, where $\hat{\zeta}_0$ is also a characteristic value of L of odd multiplicity. We use this corollary to show that there exists a continuum of solution-pairs $(\zeta, u) \in \Omega$ for the eigenvalue equation (4.1). To each of these solution-pairs there corresponds an equilibrium-pair (ζ, x^*). We define the equilibrium-pair, $(\zeta, x^*) \in Z \times \mathbb{R}^7$, as the collection of a parameter value, ζ, and the corresponding equilibrium point, x^*, for that parameter value, of the malaria model (2.8).

THEOREM 4.1. *The model* (2.8) *has a continuum of equilibrium-pairs,* $(\zeta, x^*) \in Z \times \mathbb{R}^7$*, that connects the point* (ξ_1, x_{dfe}) *to the hyperplane* $\zeta = M_Z$ *in* $\mathbb{R} \times \mathbb{R}^7$ *on the boundary of* $Z \times \mathbb{R}^7$ *for any* $M_Z > \xi_1$*, where* x^* *is in the positive orthant of* $\mathbb{R}^7$*. Here* $\xi_1 = 1/\sqrt{AB}$*, where* A *and* B *are defined in* (A.19).

We show the proof of this theorem and related lemmas in appendix section A.2.

THEOREM 4.2. *The transcritical bifurcation point at* $\zeta = \xi_1$ *corresponds to* $R_0 = 1$*. For the set of* ζ *for which there exists an equilibrium-pair* (ζ, x^*)*, the corresponding set of values for* R_0 *includes, but is not necessarily identical to, the interval* $1 < R_0 < \infty$*. Thus, there exists at least one endemic equilibrium point of the malaria model* (2.8) *for all* $R_0 > 1$*.*

Proof. Using the definition of ζ, (4.2), some algebraic manipulations of R_0 (see (3.4)) produce

$$R_0 = \zeta \sqrt{AB}. \tag{4.3}$$

Thus, R_0 is linearly related to ζ, and when $\zeta = \xi_1$, $R_0 = 1$. For any $R_0 > 1$, (4.3) defines a corresponding ζ. We pick an M_Z larger than this ζ. Then, Theorem 4.1

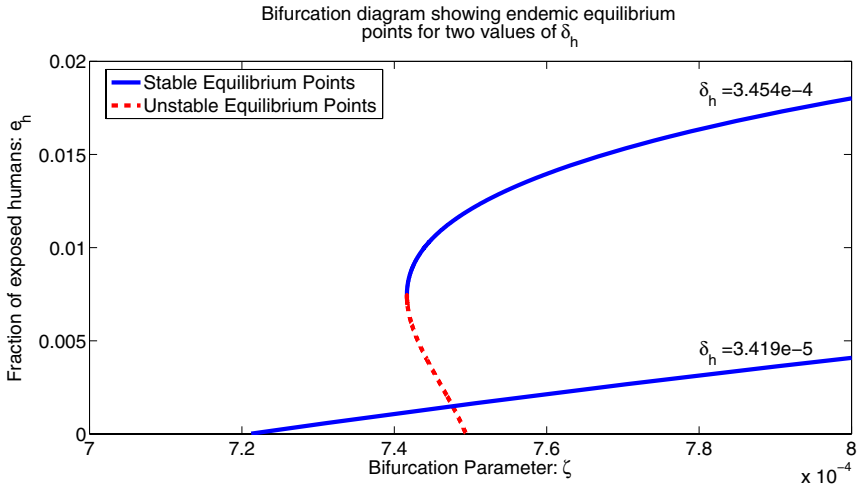


FIG. 4.1. *Bifurcation diagrams for* (2.8) *showing the endemic equilibrium values for the fraction of exposed humans,* e_h*, plotted for the parameters in Table* 4.1 *(except for* σ_v *and* σ_h*, which vary with* ζ*) and two values of the disease-induced death rate* ($\delta_h = 3.454 \times 10^{-4}$ *and* $\delta_h = 3.419 \times 10^{-5}$)*. For the parameter values in Table* 4.1*, there are three equilibrium points in* $\mathcal{D}$*: a locally asymptotically stable disease-free equilibrium point,* x_{dfe}*, on the boundary of the positive orthant of* $\mathbb{R}^7$*, and two endemic equilibrium points inside the positive orthant. Linear stability analysis shows that the "larger" endemic equilibrium point is locally asymptotically stable, while the "smaller" point is unstable. Further linear analysis with an increased value of* $\sigma_v = 0.7000$*,* $\sigma_h = 21.00$*, and all other parameters as in Table* 4.1 *(with* $R_0 = 1.155$*) shows that* x_{dfe} *is unstable, and there is one locally asymptotically stable endemic equilibrium point.*

guarantees the existence of an endemic equilibrium point for ζ, and thereby for the corresponding value of R_0. It is possible, though not necessary, for the continuum of equilibrium-pairs to include values of $\zeta < \xi_1$ ($R_0 < 1$). □

Typically in epidemiological models, bifurcations at $R_0 = 1$ tend to be supercritical (i.e., positive endemic equilibria exist for $R_0 > 1$ near the bifurcation point). In this model (2.8), in the absence of disease-induced death ($\delta_h = 0$), we prove, using the Lyapunov–Schmidt expansion as described by Cushing [9], that the bifurcation is supercritical (forward).

THEOREM 4.3. *In the absence of disease-induced death* ($\delta_h = 0$)*, the transcritical bifurcation at* $R_0 = 1$ *is supercritical (forward).*

Details of this proof are in appendix section A.2.

In the general case, a subcritical (backward) bifurcation can occur for some parameter values, where near the bifurcation point, positive endemic equilibria exist for $R_0 < 1$. Other examples of epidemiological models with subcritical bifurcations at $R_0 = 1$ include those described by Castillo-Chavez and Song [6], Gómez-Acevedo and Yi [13], and van den Driessche and Watmough [26]. The model of Ngwa and Shu [23] exhibits only a supercritical bifurcation at $R_0 = 1$. Although we cannot prove the existence of a subcritical bifurcation, we show through numerical examples that it is possible for some positive values of δ_h. This is important because it implies that there can be a stable endemic equilibrium even if $R_0 < 1$.

We use the bifurcation software program AUTO [12] to create two bifurcation diagrams around $R_0 = 1$ (Figure 4.1) with parameter values in Table 4.1, except for σ_h, σ_v, and δ_h. σ_h and σ_v change as ζ is varied, as shown in the figure; however, their ratio, $\theta = \sigma_h/\sigma_v = 30$, remains constant. One curve has δ_h as in Table 4.1,

TABLE 4.1

The parameter values for which there exist positive endemic equilibrium points when $R_0 < 1$: $R_0 = 0.9898$. The unit of time is days.

$\Lambda_h = 3.285 \times 10^{-2}$	
$\psi_h = 7.666 \times 10^{-5}$	$\psi_v = 0.4000$
$\beta_{vh} = 0.8333$	$\beta_{hv} = 2.000 \times 10^{-2}$
$\tilde{\beta}_{vh} = 8.333 \times 10^{-2}$	
$\sigma_v = 0.6000$	$\sigma_h = 18.00$
$\nu_h = 8.333 \times 10^{-2}$	$\nu_v = 0.1000$
$\gamma_h = 3.704 \times 10^{-3}$	
$\delta_h = 3.454 \times 10^{-4}$	
$\rho_h = 1.460 \times 10^{-2}$	
$\mu_{1h} = 4.212 \times 10^{-5}$	$\mu_{1v} = 0.1429$
$\mu_{2h} = 1.000 \times 10^{-7}$	$\mu_{2v} = 2.279 \times 10^{-4}$

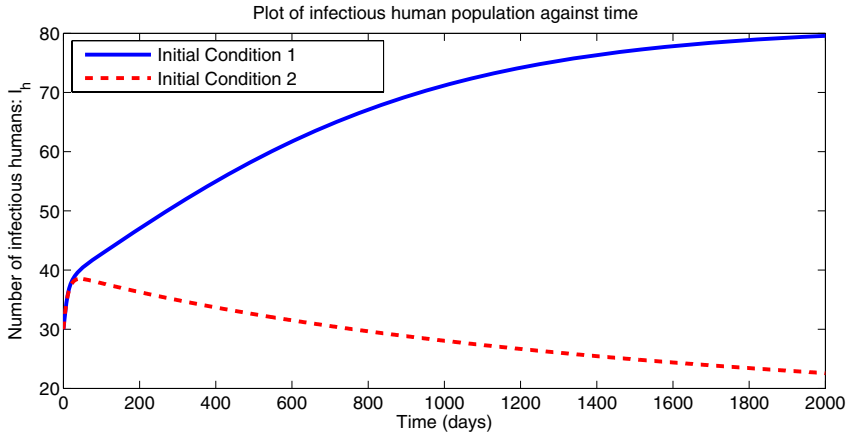


FIG. 4.2. *Solutions of the malaria model* (2.1) *with parameter values defined in Table* 4.1 *showing only the number of infectious humans, I_h, for two different initial conditions. The parameters correspond to $R_0 = 0.9898$. Initial condition* 1 *is $S_h = 400$, $E_h = 10$, $I_h = 30$, $R_h = 0$, $S_v = 1000$, $E_v = 100$, and $I_v = 50$. Initial condition* 2 *is $S_h = 700$, $E_h = 10$, $I_h = 30$, $R_h = 0$, $S_v = 1000$, $E_v = 100$, and $I_v = 50$. The solution for initial condition* 1 *approaches the locally asymptotically stable endemic equilibrium point, while the solution for initial condition* 2 *approaches the locally asymptotically stable disease-free equilibrium point.*

while the other has $\delta_h = 3.419 \times 10^{-5}$. The curve with $\delta_h = 3.454 \times 10^{-4}$ has both unstable and stable endemic equilibrium points. There is a subcritical bifurcation at $\zeta = 7.494 \times 10^{-4}$ ($R_0 = 1$), and a saddle-node bifurcation at $\zeta = 7.417 \times 10^{-4}$ ($R_0 = 0.9897$). Thus a locally asymptotically stable endemic equilibrium is possible for values of R_0 below 1. Further bifurcation analysis (not presented here) indicates that as ζ is increased to $\zeta = 50$ ($R_0 = 66719$), the size of the projection of the endemic equilibrium on the fractional infected groups increases monotonically, and the equilibrium point remains stable. For comparison we show the bifurcation diagram with $\delta_h = 3.419 \times 10^{-5}$. Here, we see only a stable branch of endemic equilibrium points. There is a supercritical bifurcation at $\zeta = 7.209 \times 10^{-4}$ ($R_0 = 1$). There are no endemic equilibrium points for R_0 less than 1. As ζ is increased to $\zeta = 50$ ($R_0 = 69358$), the size of the projection of the endemic equilibrium on the fractional infected groups increases monotonically, and the equilibrium point remains stable.

Figure 4.2 shows the infectious human population, for two different initial condi-

tions, of the solutions to the unscaled equations (2.1) for parameter values in Table 4.1 with $R_0 < 1$. One solution approaches the locally asymptotically stable endemic equilibrium point, while the other approaches the locally asymptotically stable disease-free equilibrium point.

The parameter values in Table 4.1 are within the bounds of a realistically feasible range, except for the mosquito birth and death rates, ψ_v and μ_{1v}, which have been increased to lower R_0 below 1. More realistic values are $\psi_v = 0.13$ and $\mu_{1v} = 0.033$, which result in (with all other parameters as in Table 4.1) $R_0 = 1.6$. More lists of realistic parameter values, and their references, can be found in [7] and [8]. $\delta_h = 3.454 \times 10^{-4}$ corresponds to a death rate of 12.62% of infectious humans per year.

5. Summary and conclusions. We analyzed an ordinary differential equation model for the transmission of malaria, with four variables for humans and three variables for mosquitoes. We showed that there exists a domain where the model is epidemiologically and mathematically well-posed. We proved the existence of an equilibrium point with no disease, x_{dfe}. We defined a reproductive number, R_0, that is epidemiologically accurate in that it provides the expected number of new infections (in mosquitoes or humans) from one infectious individual (human or mosquito) over the duration of the infectious period, given that all other members of the population are susceptible. We showed that if $R_0 < 1$, then the disease-free equilibrium point, x_{dfe}, is locally asymptotically stable, and if $R_0 > 1$, then x_{dfe} is unstable.

We also proved that an endemic equilibrium point exists for all $R_0 > 1$ with a transcritical bifurcation at $R_0 = 1$. The analysis and the numerical simulations showed that for $\delta_h = 0$ (no disease-induced death), and for some small positive values of δ_h, there is a supercritical transcritical bifurcation at $R_0 = 1$ with an exchange of stability between the disease-free equilibrium and the endemic equilibrium. For larger values of δ_h, there is a subcritical transcritical bifurcation at $R_0 = 1$, with an exchange of stability between the endemic equilibrium and the disease-free equilibrium; and there is a saddle-node bifurcation at $R_0 = R_0^*$ for some $R_0^* < 1$. Thus, for some values of $R_0 < 1$, there exist two endemic equilibrium points, the smaller of which is unstable, while the larger is locally asymptotically stable.

Although we cannot prove in general that the endemic equilibrium point is unique and stable for $R_0 > 1$, numerical results for particular parameter sets suggest that there is a unique stable endemic equilibrium point for $R_0 > 1$. Also, from Theorem 2.1 it follows that all orbits of the malaria model (2.8) are bounded. Thus, if there were no stable endemic equilibria in $\mathcal{D}$, then there would exist a nonequilibrium attractor (such as a limit cycle or strange attractor), though for this model we have no evidence for nonequilibrium attractors.

The possible existence of a subcritical bifurcation at $R_0 = 1$ and a saddle-node bifurcation at some $R_0^* < 1$ can have implications for public health, when the epidemiological parameters are close to those in Table 4.1. Simply reducing R_0 to a value below 1 is not always sufficient to eradicate the disease; it is now necessary to reduce R_0 to a value less than R_0^* to ensure that there are no endemic equilibria. The existence of a saddle-node bifurcation also implies that in some areas with endemic malaria, it may be possible to significantly reduce prevalence or eradicate the disease with small increases in control programs (a small reduction in R_0 so that it is less than R_0^*). In some areas where malaria has been eradicated it is possible for a slight disruption, like a change in environmental or control variables or an influx of infectious humans or mosquitoes, to cause the disease to reestablish itself in the population with a significant increase in disease prevalence (increasing R_0 above R_0^*

or moving the system into the basin of attraction of the stable endemic equilibrium).

As we have an explicit expression for R_0, we can analytically evaluate its sensitivity to the different parameter values. We can also numerically evaluate the sensitivity of the endemic equilibrium to the parameter values. This allows us to determine the relative importance of the parameters to disease transmission and prevalence. As each malaria intervention strategy affects different parameters to different degrees, we can thus compare different control strategies for efficiency and effectiveness in reducing malaria mortality and morbidity. This analysis, in the limiting case of the Chitnis model [7] shows that malaria transmission is most sensitive to the mosquito biting rate, and prevalence is most sensitive to the mosquito biting rate and the human recovery rate. The sensitivity analysis for the new model (2.8) is forthcoming [8].

We are extending the model to include the effects of the environment on the spread of malaria. Some parameters, such as the mosquito birth rate and the incubation period in mosquitoes, depend on seasonal environmental factors such as rainfall, temperature, and humidity. We can include these effects by modeling these parameters as periodic functions of time. We would like to explore this periodically forced model for features not seen in the autonomous model, including the modifications to the definition of the reproductive number and the endemic states. This would provide a more accurate picture of malaria transmission and prevalence than that obtained from models using parameter values that are averaged over the seasons. Other planned improvements to the model include the addition of age and spatial structure.

An ultimate goal is to validate this model by applying it to a particular malaria-endemic region of the world to compare the predicted endemic states with the prevalence data.

Appendix. Lemmas and proofs of theorems.

LEMMA A.1. *For all equilibrium points on* $\mathcal{D} \cap \partial\mathbb{R}^7_+$, $e_h = i_h = r_h = e_v = i_v = 0$.

Proof. We need to show that for an equilibrium point in $\mathcal{D}$, if any one of the diseased classes is zero, all the rest are also equal to zero. We show, by setting the right-hand side of (2.8) equal to 0, that if any one of e_h, i_h, r_h, e_v, or i_v is equal to 0, then $e_h = i_h = r_h = e_v = i_v = 0$. For $i'_h = 0$, $e_h = 0$ if and only if $i_h = 0$.[2] Similarly, for $r'_h = 0$, $i_h = 0$ if and only if $r_h = 0$. Thus, if $e_h = 0$, $i_h = 0$, or $r_h = 0$, then $e_h = i_h = r_h = 0$. From $e'_h = 0$, we see that if $e_h = i_h = r_h = 0$, then $i_v = 0$. Also, for $i'_v = 0$, $e_v = 0$ if and only if $i_v = 0$. Thus, if $e_v = 0$ or $i_v = 0$, then $e_v = i_v = 0$. Finally, for $e'_v = 0$, if $e_v = i_v = 0$, then $i_h = r_h = 0$. □

A.1. Proof of Theorem 3.3.

Proof. The Jacobian of the malaria model (2.8) evaluated at x_{dfe} is of the form

$$J = \begin{pmatrix} J_{11} & 0 & 0 & 0 & 0 & J_{16} & 0 \\ J_{21} & J_{22} & 0 & 0 & 0 & 0 & 0 \\ 0 & J_{32} & J_{33} & 0 & 0 & 0 & 0 \\ 0 & J_{42} & 0 & J_{44} & 0 & 0 & 0 \\ 0 & J_{52} & J_{53} & 0 & J_{55} & 0 & 0 \\ 0 & 0 & 0 & 0 & J_{65} & J_{66} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & J_{77} \end{pmatrix}. \tag{A.1}$$

[2] As the right-hand side of (2.8b) is a quadratic function of i_h, there are two possible solutions of i_h when $i'_h = 0$ and $e_h = 0$. However, the nonzero solution of i_h is greater than 1 and is thus outside of $\mathcal{D}$.

As the fourth and seventh columns (corresponding to the total human and mosquito populations) contain only the diagonal terms, these diagonal terms form two eigenvalues of the Jacobian:

$$\eta_6 = \psi_h - \mu_{1h} - 2\mu_{2h}N_h^* = -\sqrt{(\psi_h - \mu_{1h})^2 + 4\mu_{2h}\Lambda_h}, \tag{A.2a}$$

$$\eta_7 = \psi_v - \mu_{1v} - 2\mu_{2v}N_v^* = -(\psi_v - \mu_{1v}). \tag{A.2b}$$

As we have assumed that $\psi_v > \mu_{1v}$, both η_6 and η_7 are always negative. The other five eigenvalues are the roots of the characteristic equation of the matrix formed by excluding the fourth and seventh rows and columns of the Jacobian (A.1):

$$A_5\eta^5 + A_4\eta^4 + A_3\eta^3 + A_2\eta^2 + A_1\eta + A_0 = 0 \tag{A.3}$$

with

$$\begin{aligned}
A_5 &= 1,\\
A_4 &= B_1 + B_2 + B_3 + B_4 + B_5,\\
A_3 &= B_1B_2 + B_1B_3 + B_1B_4 + B_1B_5 + B_2B_3 + B_2B_4 + B_2B_5 + B_3B_4\\
&\quad + B_3B_5 + B_4B_5,\\
A_2 &= B_1B_2B_3 + B_1B_2B_4 + B_1B_2B_5 + B_1B_3B_4 + B_1B_3B_5 + B_1B_4B_5 + B_2B_3B_4\\
&\quad + B_2B_3B_5 + B_2B_4B_5 + B_3B_4B_5,\\
A_1 &= B_1B_2B_3B_4 + B_1B_2B_3B_5 + B_1B_2B_4B_5 + B_1B_3B_4B_5 + B_2B_3B_4B_5\\
&\quad - B_6B_7B_8B_9,\\
A_0 &= B_1B_2B_3B_4B_5 - (B_3B_6B_7B_8B_9 + B_6B_7B_9B_{10}B_{11}),
\end{aligned}$$

and $B_1 = \nu_h + \psi_h + \Lambda_h/N_h^*$, $B_2 = \gamma_h + \delta_h + \psi_h + \Lambda_h/N_h^*$, $B_3 = \rho_h + \psi_h + \Lambda_h/N_h^*$, $B_4 = \nu_v + \psi_v$, $B_5 = \psi_v$, $B_6 = b_h^*\beta_{hv}$, $B_7 = \nu_h$, $B_8 = b_v^*\beta_{vh}$, $B_9 = \nu_v$, $B_{10} = \gamma_h$, and $B_{11} = b_v^*\tilde{\beta}_{vh}$.

To evaluate the signs of the roots of (A.3), we first use the Routh–Hurwitz criterion to prove that when $R_0 < 1$, all roots of (A.3) have negative real part. Then, using Descartes's rule of sign, we prove that when $R_0 > 1$, there is one positive real root.

The Routh–Hurwitz criterion [18, section 1.6-6(b)] for a real algebraic equation

$$a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 = 0 \tag{A.4}$$

states that, given $a_n > 0$, all roots have negative real part if and only if $T_0 = a_n$, $T_1 = a_{n-1}$,

$$T_2 = \begin{vmatrix} a_{n-1} & a_n \\ a_{n-3} & a_{n-2} \end{vmatrix}, T_3 = \begin{vmatrix} a_{n-1} & a_n & 0 \\ a_{n-3} & a_{n-2} & a_{n-1} \\ a_{n-5} & a_{n-4} & a_{n-3} \end{vmatrix}, \ldots, T_n = \begin{vmatrix} a_{n-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_0 \end{vmatrix}$$

are all positive, with $a_i = 0$ for $i < 0$. This is true if and only if all a_i and either all even-numbered T_k or all odd-numbered T_k are positive (the Liénard–Chipart test). Korn and Korn [18] in section 1.6-6(c) state Descartes's rule of sign as the number of positive real roots of a real algebraic equation (A.4) is equal to the number, N_a, of sign changes in the sequence, $a_n, a_{n-1}, \ldots, a_0$, of coefficients, where the vanishing terms are disregarded, or it is less than N_a by a positive even integer.

We show that when $R_0 < 1$, all the coefficients, A_i, of the characteristic equation (A.3), and T_0, T_2, and T_4, are positive, so by the Routh–Hurwitz criterion, all the eigenvalues of the Jacobian (A.1) have negative real part. We then show that when $R_0 > 1$, there is one and only one sign change in the sequence $A_5, A_4, \ldots, A_0$, so by Descartes's rule of sign there is one eigenvalue with positive real part, and the disease-free equilibrium point is unstable.

The expression for R_0^2 in (3.4) can be written, in terms of B_i, as

$$R_0^2 = \frac{B_3B_6B_7B_8B_9 + B_6B_7B_9B_{10}B_{11}}{B_1B_2B_3B_4B_5}. \tag{A.5}$$

For $R_0 < 1$, by (A.5),

$$B_3B_6B_7B_8B_9 + B_6B_7B_9B_{10}B_{11} < B_1B_2B_3B_4B_5, \tag{A.6}$$

$$B_6B_7B_8B_9 < B_1B_2B_4B_5. \tag{A.7}$$

As all the B_i are positive, A_5, A_4, A_3, and A_2 are always positive. From (A.7) we see that $A_1 > 0$, and from (A.6) we see that $A_0 > 0$. Thus, for $R_0 < 1$, all A_i are positive. We now show that the even-numbered T_k are positive for $R_0 < 1$. For the fifth-degree polynomial (A.3), $T_0 = A_5$, which is always positive. $T_2 = A_3A_4 - A_2A_5$, which we can show to be a positive sum of products of B_i's, so $T_2 > 0$. Lastly,

$$T_4 = A_1[A_2A_3A_4 - (A_1A_4^2 + A_2^2A_5)] - A_0[A_3(A_3A_4 - A_2A_5) - (2A_1A_4A_5 - A_0A_5^2)].$$

For ease of notation, we introduce

$$C_1 = A_2A_3A_4 - (A_1A_4^2 + A_2^2A_5),$$
$$C_2 = A_3(A_3A_4 - A_2A_5) - (2A_1A_4A_5 - A_0A_5^2),$$

where we can show that $C_1 > 0$ and $C_2 > 0$, so that $T_4 = A_1C_1 - A_0C_2$. We define

$$C_2^{(1)} = C_2 + B_6B_7B_9B_{10}B_{11}.$$

As $C_2^{(1)} > C_2$ and $A_0 > 0$, for $T_4^{(1)} = A_1C_1 - A_0C_2^{(1)}$, $T_4 > T_4^{(1)}$. Similarly, we define

$$A_0^{(1)} = A_0 + (B_3B_6B_7B_8B_9 + B_6B_7B_9B_{10}B_{11}).$$

As $A_0^{(1)} > A_0$ and $C_2^{(1)} > 0$, for $T_4^{(2)} = A_1C_1 - A_0^{(1)}C_2^{(1)}$, $T_4^{(1)} > T_4^{(2)}$. Finally, we define

$$A_1^{(1)} = A_1 - (B_1B_2B_4B_5 - B_6B_7B_8B_9).$$

As $A_1^{(1)} < A_1$ (for $R_0 < 1$) and $C_1 > 0$, for $T_4^{(3)} = A_1^{(1)}C_1 - A_0^{(1)}C_2^{(1)}$, $T_4^{(2)} > T_4^{(3)}$. We can show that $T_4^{(3)}$ is a sum of positive terms, so $T_4^{(3)} > 0$. As $T_4 > T_4^{(1)} > T_4^{(2)} > T_4^{(3)}$, $T_4 > 0$. Thus, for $R_0 < 1$, all roots of (A.3) have negative real parts.

When $R_0 > 1$

$$B_3B_6B_7B_8B_9 + B_6B_7B_9B_{10}B_{11} > B_1B_2B_3B_4B_5,$$

and so $A_0 < 0$. As A_5, A_4, A_3, and A_2 are positive, the sequence, $A_5, A_4, A_3, A_2, A_1, A_0$ has exactly one sign change. Thus, by Descartes's rule of sign, (A.3) has one positive real root when $R_0 > 1$.

Thus, the disease-free equilibrium point, x_{dfe}, is locally asymptotically stable if $R_0 < 1$ and unstable if $R_0 > 1$. If $R_0 < 1$, on average each infected individual infects fewer than one other individual, and the disease dies out. If $R_0 > 1$, on average each infected individual, infects more than one other individual, so we would expect the disease to spread. The Jacobian of (2.8) at x_{dfe} has one eigenvalue equal to 0 at $R_0 = 1$. □

A.2. Proofs of theorems and lemmas for the existence of endemic equilibrium points. The equilibrium equations for (2.8) are shown below in (A.8). In this analysis, we use the terms e_h, i_h, r_h, N_h, e_v, i_v, and N_v to represent their respective equilibrium values and not their actual values at a given time, t.

$$\left(\frac{\sigma_v\sigma_h N_v\beta_{hv}i_v}{\sigma_v N_v+\sigma_h N_h}\right)(1-e_h-i_h-r_h)-(\nu_h+\psi_h+\Lambda_h/N_h)e_h+\delta_h i_h e_h=0, \tag{A.8a}$$

$$\nu_h e_h-(\gamma_h+\delta_h+\psi_h+\Lambda_h/N_h)i_h+\delta_h i_h^2=0, \tag{A.8b}$$

$$\gamma_h i_h-(\rho_h+\psi_h+\Lambda_h/N_h)r_h+\delta_h i_h r_h=0, \tag{A.8c}$$

$$\Lambda_h+\psi_h N_h-(\mu_{1h}+\mu_{2h}N_h)N_h-\delta_h i_h N_h=0, \tag{A.8d}$$

$$\left(\frac{\sigma_v\sigma_h N_h}{\sigma_v N_v+\sigma_h N_h}\right)\left(\beta_{vh}i_h+\tilde{\beta}_{vh}r_h\right)(1-e_v-i_v)-(\nu_v+\psi_v)e_v=0, \tag{A.8e}$$

$$\nu_v e_v-\psi_v i_v=0, \tag{A.8f}$$

$$\psi_v N_v-(\mu_{1v}+\mu_{2v}N_v)N_v=0. \tag{A.8g}$$

We rewrite (A.8a) and (A.8e) in terms of the bifurcation parameter, ζ (4.2), and a new parameter, $\theta=\sigma_h/\sigma_v$, to obtain

$$\zeta\left(\frac{N_v^*+\theta N_h^*}{N_v+\theta N_h}\right)N_v\beta_{hv}i_v(1-e_h-i_h-r_h)-(\nu_h+\psi_h+\Lambda_h/N_h-\delta_h i_h)e_h=0, \tag{A.9a}$$

$$\zeta\left(\frac{N_v^*+\theta N_h^*}{N_v+\theta N_h}\right)N_h\left(\beta_{vh}i_h+\tilde{\beta}_{vh}r_h\right)(1-e_v-i_v)-(\nu_v+\psi_v)e_v=0. \tag{A.9b}$$

We can vary the bifurcation parameter, ζ, while keeping all other parameters fixed. In terms of the original variables, this corresponds to changing σ_h and σ_v while keeping the ratio between them fixed. We can pick θ, the ratio between them, and sweep out the entire parameter space.

We reduce the equilibrium equations to a two-dimensional system for e_h and e_v by solving for the other variables, either explicitly as functions of the parameters, or in terms of e_h and e_v. We solve (A.8g) for N_v, explicitly expressing the positive equilibrium for the total mosquito population in terms of parameters (exactly as in the disease-free case (3.1)):

$$N_v=\frac{\psi_v-\mu_{1v}}{\mu_{2v}}. \tag{A.10}$$

Solving for i_v in (A.8f) in terms of e_v, we find

$$i_v=\frac{\nu_v}{\psi_v}e_v. \tag{A.11}$$

We write the positive equilibrium for N_h in terms of i_h from (A.8d) as

$$N_h=\frac{(\psi_h-\mu_{1h}-\delta_h i_h)+\sqrt{(\psi_h-\mu_{1h}-\delta_h i_h)^2+4\mu_{2h}\Lambda_h}}{2\mu_{2h}}. \tag{A.12}$$

Using (A.12) in (A.8c), we solve for r_h in terms of i_h:

$$r_h = \frac{2\gamma_h i_h}{2\rho_h + (\psi_h + \mu_{1h} - \delta_h i_h) + \sqrt{(\psi_h - \mu_{1h} - \delta_h i_h)^2 + 4\mu_{2h}\Lambda_h}}. \tag{A.13}$$

Given the nonlinear nature of (A.8b), it is not feasible (or useful) to solve for i_h in terms of e_h explicitly. We therefore use (A.12) to rewrite (A.8b), and define the function $e_h = g(i_h)$ as

$$g(i_h) = \frac{\gamma_h + \delta_h + \frac{1}{2}\left((\psi_h + \mu_{1h} - \delta_h i_h) + \sqrt{(\psi_h - \mu_{1h} - \delta_h i_h)^2 + 4\mu_{2h}\Lambda_h}\right)}{\nu_h} i_h.$$

We note that $g(0) = 0$, and label the positive constant $g(1) = e_h^{max}$. As $g(i_h)$ is a smooth function of i_h with $g'(i_h) > 0$ for $i_h \in [0,1]$ and $e_h \in [0, e_h^{max}]$, there exists a smooth function $i_h = y(e_h)$ with domain $[0, e_h^{max}]$ and range $[0,1]$. As $g'(0) > 0$, the smooth function $y(e_h)$ would extend to some small $e_h < 0$. Substituting $i_h = y(e_h)$ into (A.12) and (A.13), we can also express N_h and r_h as functions of e_h.

We now introduce the bounded open subset of $\mathbb{R}^2$,

$$Y = \left\{ \begin{pmatrix} e_h \\ e_v \end{pmatrix} \in \mathbb{R}^2 \;\middle|\; \begin{array}{l} -\epsilon_h < e_h < e_h^{max} \\ -\epsilon_v < e_v < 1 \end{array} \right\}, \tag{A.14}$$

for some $\epsilon_v > 0$ and some $\epsilon_h > 0$. By substituting (A.10), (A.11), (A.12), (A.13), and $i_h = y(e_h)$ into (A.8a) and (A.8e), we reformulate the seven equilibrium equations (A.8) equivalently as two equations for the components $(e_h, e_v) \in Y$. To place these two equations into the Rabinowitz form (4.1), we need to determine lower order terms. We rewrite (A.8b) as $f(e_h, i_h) = 0$, where $f(e_h, i_h) =$

$$\nu_h e_h - \left[\gamma_h + \delta_h + \frac{1}{2}\left((\psi_h + \mu_{1h} - \delta_h i_h) + \sqrt{(\psi_h - \mu_{1h} - \delta_h i_h)^2 + 4\mu_{2h}\Lambda_h}\right)\right] i_h,$$

and use implicit differentiation to write $i_h = y(e_h)$ as a Taylor polynomial of the form

$$i_h = y_1 e_h + \mathcal{O}(e_h^2), \tag{A.15}$$

where

$$y_1 = -\left.\frac{\frac{\partial f}{\partial e_h}}{\frac{\partial f}{\partial i_h}}\right|_{i_h = e_h = 0} = \frac{\nu_h}{\gamma_h + \delta_h + \frac{1}{2}\left((\psi_h + \mu_{1h}) + \sqrt{(\psi_h - \mu_{1h})^2 + 4\mu_{2h}\Lambda_h}\right)}.$$

Finally, we substitute the Taylor approximation for i_h (A.15) into r_h (A.13) and N_h (A.12), and then all three, along with i_v (A.11) and N_v (A.10) into the equilibrium equations for e_h (A.9a) and e_v (A.9b), to provide first order approximations to the equilibrium equations:

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} f_{1\_10} & f_{1\_01} \\ f_{2\_10} & f_{2\_01} \end{pmatrix} \begin{pmatrix} e_h \\ e_v \end{pmatrix} + \mathcal{O}\left(\begin{pmatrix} e_h \\ e_v \end{pmatrix}^2\right), \tag{A.16}$$

where

$$f_{1\_10} = -\left[\nu_h + \frac{1}{2}\left((\psi_h + \mu_{1h}) + \sqrt{(\psi_h - \mu_{1h})^2 + 4\mu_{2h}\Lambda_h}\right)\right], \tag{A.17a}$$

$$f_{1\_01} = \zeta \cdot \frac{\nu_v \beta_{hv}(\psi_v - \mu_{1v})}{\psi_v \mu_{2v}}, \tag{A.17b}$$

$$f_{2\_10} = \zeta \cdot \frac{\nu_h\left((\psi_h - \mu_{1h}) + \sqrt{(\psi_h - \mu_{1h})^2 + 4\Lambda_h\mu_{2h}}\right)}{2\mu_{2h}\left(\gamma_h + \delta_h + \frac{1}{2}\left((\psi_h + \mu_{1h}) + \sqrt{(\psi_h - \mu_{1h})^2 + 4\mu_{2h}\Lambda_h}\right)\right)} \times \left[\beta_{vh} + \frac{\gamma_h\tilde{\beta}_{vh}}{\rho_h + \frac{1}{2}\left((\psi_h + \mu_{1h}) + \sqrt{(\psi_h - \mu_{1h})^2 + 4\mu_{2h}\Lambda_h}\right)}\right], \tag{A.17c}$$

$$f_{2\_01} = -\left(\psi_v + \nu_v\right). \tag{A.17d}$$

To apply Corollary 1.12 of Rabinowitz [24], we algebraically manipulate (A.16) to produce

$$u = \zeta L u + h(\zeta, u), \tag{A.18}$$

where

$$u = \begin{pmatrix} e_h \\ e_v \end{pmatrix} \quad \text{and} \quad L = \begin{pmatrix} 0 & A \\ B & 0 \end{pmatrix} \quad \text{with}$$

$$A = \frac{\nu_v \beta_{hv}(\psi_v - \mu_{1v})}{\psi_v \mu_{2v}\left(\nu_h + \frac{1}{2}\left((\psi_h + \mu_{1h}) + \sqrt{(\psi_h - \mu_{1h})^2 + 4\mu_{2h}\Lambda_h}\right)\right)}, \tag{A.19a}$$

$$B = \left(\beta_{vh} + \frac{\gamma_h\tilde{\beta}_{vh}}{\rho_h + \frac{1}{2}\left((\psi_h + \mu_{1h}) + \sqrt{(\psi_h - \mu_{1h})^2 + 4\mu_{2h}\Lambda_h}\right)}\right) \times \frac{\nu_h\left((\psi_h - \mu_{1h}) + \sqrt{(\psi_h - \mu_{1h})^2 + 4\mu_{2h}\Lambda_h}\right)}{2\mu_{2h}(\psi_v + \nu_v)\left(\gamma_h + \delta_h + \frac{1}{2}\left((\psi_h + \mu_{1h}) + \sqrt{(\psi_h - \mu_{1h})^2 + 4\mu_{2h}\Lambda_h}\right)\right)}, \tag{A.19b}$$

and $h(\zeta, u)$ is $\mathcal{O}(u^2)$. The matrix, L, has two distinct eigenvalues: $\pm\sqrt{AB}$. Characteristic values of a matrix are the reciprocals of its eigenvalues. We denote the two characteristic values of L by $\xi_1 = 1/\sqrt{AB}$ and $\xi_2 = -1/\sqrt{AB}$. As both A and B are always positive (because we have assumed that $\psi_v > \mu_{1v}$), ξ_1 is real and corresponds to the dominant eigenvalue of L. The right and left eigenvectors corresponding to ξ_1 are, respectively,

$$v = \begin{pmatrix} \sqrt{A} \\ \sqrt{B} \end{pmatrix} \quad \text{and} \quad w = \begin{pmatrix} \sqrt{B} & \sqrt{A} \end{pmatrix}. \tag{A.20}$$

For $M_Z > \xi_1$, as $0 \in Y$, $(\xi_1, 0) \in \Omega$. By Corollary 1.12 of Rabinowitz [24], we know that there is a continuum of solution-pairs $(\zeta, u) \in \Omega$, whose closure contains the point $(\xi_1, 0)$, that either meets the boundary of Ω, $\partial\Omega$, or the point $(\xi_2, 0)$. We

denote the continuum of solution-pairs emanating from $(\xi_1, 0)$ by $\mathscr{C}_1$, where $\mathscr{C}_1 \subset \Omega$, and from $(\xi_2, 0)$ by $\mathscr{C}_2$, where $\mathscr{C}_2 \subset \Omega$. We introduce the sets

$$Z_1 = \{\zeta \in Z |\ \exists u \text{ such that } (\zeta, u) \in \mathscr{C}_1\}, \tag{A.21a}$$

$$U_1 = \{u \in Y |\ \exists \zeta \text{ such that } (\zeta, u) \in \mathscr{C}_1\}, \tag{A.21b}$$

$$Z_2 = \{\zeta \in Z |\ \exists u \text{ such that } (\zeta, u) \in \mathscr{C}_2\}, \tag{A.21c}$$

$$U_2 = \{u \in Y |\ \exists \zeta \text{ such that } (\zeta, u) \in \mathscr{C}_2\}. \tag{A.21d}$$

We denote the part of Y in the positive quadrant of $\mathbb{R}^2$ by $Y^+ = \{(e_h, e_v) \in Y \,|\, e_h > 0 \text{ and } e_v > 0\}$, and the internal boundary of Y^+ by

$$\partial Y^+ = \left\{ \begin{pmatrix} e_h \\ e_v \end{pmatrix} \in Y \,\middle|\, \begin{pmatrix} e_h > 0 \\ \text{and} \\ e_v = 0 \end{pmatrix} \text{ or } \begin{pmatrix} e_h = 0 \\ \text{and} \\ e_v > 0 \end{pmatrix} \text{ or } \begin{pmatrix} e_h = 0 \\ \text{and} \\ e_v = 0 \end{pmatrix} \right\}.$$

We can determine the initial direction of the continua of solution-pairs, $\mathscr{C}_1$ and $\mathscr{C}_2$, using the Lyapunov–Schmidt expansion, as described by Cushing [9]. Although we show the proofs only for the expansion of $\mathscr{C}_1$ around the bifurcation point at $\zeta = \xi_1$ in Lemmas A.2 and A.3, the results for $\mathscr{C}_2$ around $\zeta = \xi_2$ are similar. We begin by expanding the terms of the nonlinear eigenvalue equation (A.18) about the bifurcation point, $(\xi_1, 0)$. The expanded variables are

$$u = 0 + \varepsilon u^{(1)} + \varepsilon^2 u^{(2)} + \cdots, \tag{A.22a}$$

$$\zeta = \xi_1 + \varepsilon\zeta_1 + \varepsilon^2\zeta_2 + \cdots, \tag{A.22b}$$

$$\begin{aligned} h(\zeta, u) &= h(\xi_1 + \varepsilon\zeta_1 + \varepsilon^2\zeta_2 + \cdots, \varepsilon u^{(1)} + \varepsilon^2 u^{(2)} + \cdots) \\ &= \varepsilon^2 h_2(\xi_1, u^{(1)}) + \cdots. \end{aligned} \tag{A.22c}$$

We substitute the expansions (A.22) into the eigenvalue equation (A.18) and evaluate at different orders of ε. Evaluating the substitution of the expansions (A.22) into the eigenvalue equation (A.18) at $\mathcal{O}(\varepsilon^0)$ produces $0 = 0$, which gives us no information.

LEMMA A.2. *The initial direction of the branch of equilibrium points, $u^{(1)}$, near the bifurcation point, $(\xi_1, 0)$, is equal to the right eigenvector of L corresponding to the characteristic value, ξ_1.*

Proof. Evaluating the substitution of the expansions (A.22) into the eigenvalue equation (A.18) at $\mathcal{O}(\varepsilon^1)$, we obtain $u^{(1)} = \xi_1 L u^{(1)}$. This implies that $u^{(1)}$ is the right eigenvector of L corresponding to the eigenvalue $1/\xi_1$, v (A.20). Thus, close to the bifurcation point, the equilibrium point can be approximated by $e_h = \varepsilon\sqrt{A}$ and $e_v = \varepsilon\sqrt{B}$. ∎

LEMMA A.3. *The bifurcation at $\zeta = \xi_1$ of the nonlinear eigenvalue equation* (A.18) *is supercritical if $\zeta_1 > 0$ and subcritical if $\zeta_1 < 0$, where*

$$\zeta_1 = -\frac{w \cdot h_2}{w \cdot Lv}, \tag{A.23}$$

where v and w are the right and left eigenvectors of L corresponding to the characteristic value ξ_1, respectively.

Proof. Evaluating the substitution of the expansions (A.22) into the eigenvalue equation (A.18) at $\mathcal{O}(\varepsilon^2)$, we obtain $u^{(2)} = \xi_1 L u^{(2)} + \zeta_1 L u^{(1)} + h_2$, which we can

rewrite as

$$(\mathbb{I} - \xi_1 L)u^{(2)} = \zeta_1 Lv + h_2, \tag{A.24}$$

where $\mathbb{I}$ is the 2×2 identity matrix. As ξ_1 is a characteristic value of L, $(\mathbb{I} - \xi_1 L)$ is a singular matrix. Thus, for (A.24) to have a solution, $\zeta_1 Lv + h_2$ must be in the range of $(\mathbb{I} - \xi_1 L)$; i.e., it must be orthogonal to the null space of the adjoint of $(\mathbb{I} - \xi_1 L)$. The null space of the adjoint of $(\mathbb{I} - \xi_1 L)$ is spanned by the left eigenvector of L (corresponding to the eigenvalue $1/\xi_1$), w (A.20). The Fredholm condition for the solvability of (A.24) is $w \cdot (\zeta_1 Lv + h_2) = 0$. Solving for ζ_1 provides (A.23). If ζ_1 is positive, then for small positive ε, $u > 0$ and $\zeta > \xi_1$, and the bifurcation is supercritical. Similarly, if ζ_1 is negative, then for small positive ε, $u > 0$ and $\zeta < \xi_1$, and the bifurcation is subcritical. □

LEMMA A.4. *For all $u \in U_1$, $e_h > 0$ and $e_v > 0$.*

Proof. By Lemma A.1, there are no equilibrium points on ∂Y^+ other than $e_h = e_v = 0$, so $U_1 \cap \partial Y^+ = 0$. We know from Lemma A.2 that close to the bifurcation point $(\xi_1, 0)$, the direction of U_1 is equal to v, the right eigenvector corresponding to the characteristic value, ξ_1. As v contains only positive terms, U_1 is entirely contained in Y^+. Thus, for all $u \in U_1$, $e_h > 0$ and $e_v > 0$. □

LEMMA A.5. *The point $u = 0 \in Y$ corresponds to $x_{dfe} \in \mathbb{R}^7$ (on the boundary of the positive orthant of $\mathbb{R}^7$). For every solution-pair $(\zeta, u) \in \mathscr{C}_1$, there corresponds one equilibrium-pair $(\zeta, x^*) \in Z \times \mathbb{R}^7$, where x^* is in the positive orthant of $\mathbb{R}^7$.*

Proof. We first show that $u = 0$ corresponds to x_{dfe}. As $e_h = e_v = 0$, by Theorem 3.1 we know that the only possible equilibrium point is x_{dfe}. We now show that for every $\zeta \in Z_1$ there exists an x^* in the positive orthant of $\mathbb{R}^7$ for the corresponding $u \in U_1$. By Lemma A.4, we know that $e_h > 0$ and $e_v > 0$. We now need to show that for every positive e_h and e_v there exist corresponding positive i_h, r_h, i_v, N_h, and N_v. By looking at the equilibrium equation for i_v (A.11), we see that for every positive e_v there exists a positive i_v. The equilibrium equation for N_v has a positive and bounded solution, depending only on parameter values (A.10). From $i_h = y(e_h)$, we see that for every positive e_h there exists a positive i_h. The equilibrium equations for r_h (A.13) and N_h (A.12) show that for every positive i_h there exists a positive r_h and N_h. □

LEMMA A.6. *The set U_1 does not meet the boundary of Y.*

Proof. As Lemma A.4 shows us that for all $u \in U_1$, $e_h > 0$ and $e_v > 0$, we need to show that $e_h < e_h^{max}$ and $e_v < 1$. By Lemma A.5, we know that all state variables are positive. Therefore, for (A.8e) to have a solution, $e_v + i_v < 1$ so $e_v < 1$. From the properties of $e_h = g(i_h)$, we know that as i_h increases, e_h increases monotonically, reaching e_h^{max} at $i_h = 1$. However, we have already shown that when $e_h + i_h + r_h = 1$, $e_h' + i_h' + r_h' < 0$, and thus there can be no equilibrium point at $e_h + i_h + r_h = 1$. Therefore, i_h is always less than 1, and e_h is always less than e_h^{max}. □

Proof of Theorem 4.1. As shown in Lemma A.4, $U_1 \cap \partial Y^+ = 0$ and U_1 is entirely contained in Y^+. We can similarly show that U_2 is entirely outside of Y^+ because the right eigenvector corresponding to ξ_2 is $\begin{pmatrix} -\sqrt{A} & \sqrt{B} \end{pmatrix}^{\mathrm{T}}$. Therefore, $\mathscr{C}_1$ and $\mathscr{C}_2$ do not intersect, and by Corollary 1.12 of Rabinowitz [24], $\mathscr{C}_1$ meets $\partial\Omega$. By Lemma A.6, the set U_1 does not meet the boundary of Y, so $\mathscr{C}_1$ meets $\partial\Omega$ only at $\zeta = M_Z$.

By Lemma A.5, for every $u \in U_1$, there corresponds an x^* in the positive orthant of $\mathbb{R}^7$, and $u = 0$ corresponds to x_{dfe} (on the boundary of the positive orthant of $\mathbb{R}^7$). Thus, there exists a continuum of equilibrium-pairs $(\zeta, x^*) \in Z \times \mathbb{R}^7$ that connects the point (ξ_1, x_{dfe}) to the hyperplane $\zeta = M_Z$ in $\mathbb{R} \times \mathbb{R}^7$. □

Proof of Theorem 4.3. When $\delta_h = 0$, we can explicitly evaluate $h(\zeta, u)$ in the nonlinear eigenvalue equation (A.18) from the equilibrium equations (A.8) as

$$h = \zeta \begin{pmatrix} C_{(\delta_h=0)} e_h e_v \\ D_{(\delta_h=0)} e_h e_v \end{pmatrix} \tag{A.25}$$

since the coefficients of all the other higher order terms are zero. Although we do not show the explicit representations for $C_{(\delta_h=0)}$ and $D_{(\delta_h=0)}$, they are both negative. From (A.25) and (A.22) we can evaluate the second order expansion

$$h_2 = \xi_1 \begin{pmatrix} C_{(\delta_h=0)} \sqrt{A}\sqrt{B} \\ D_{(\delta_h=0)} \sqrt{A}\sqrt{B} \end{pmatrix} = \begin{pmatrix} C_{(\delta_h=0)} \\ D_{(\delta_h=0)} \end{pmatrix}. \tag{A.26}$$

As h_2 contains only negative terms and w, v, and L contain only nonnegative terms, (A.23) implies that ζ_1 is positive. Thus, by Lemma A.3, with no disease-induced death, for any positive values of the other parameters there is a supercritical bifurcation at $R_0 = 1$. □

Acknowledgements. The authors thank Karl Hadeler for his discussions and ideas on improving the model, including the mosquitoes' human-biting rates; Alain Goriely, Joceline Lega, Jia Li, Seymour Parter, and Joel Miller for their careful reading of the manuscript and valuable comments; and two anonymous referees for many helpful suggestions.

REFERENCES

[1] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, Oxford, UK, 1991.

[2] J. L. Aron, *Mathematical modeling of immunity to malaria*, Math. Biosci., 90 (1988), pp. 385–396.

[3] J. L. Aron and R. M. May, *The population dynamics of malaria*, in The Population Dynamics of Infectious Disease: Theory and Applications, R. M. Anderson, ed., Chapman and Hall, London, 1982, pp. 139–179.

[4] N. Bacaër and C. Sokhna, *A reaction-diffusion system modeling the spread of resistance to an antimalarial drug*, Math. Biosci. Engrg., 2 (2005), pp. 227–238.

[5] N.J.T. Bailey, *The Mathematical Theory of Infectious Diseases and Its Application*, Griffin, London, 1975.

[6] C. Castillo-Chavez and B. Song, *Dynamical models of tuberculosis and their applications*, Math. Biosci. Engrg., 1 (2004), pp. 361–404.

[7] N. Chitnis, *Using Mathematical Models in Controlling the Spread of Malaria*, Ph.D. thesis, Program in Applied Mathematics, University of Arizona, Tucson, AZ, 2005.

[8] N. Chitnis, J. M. Hyman, and J. M. Cushing, *Determining Important Parameters in the Spread of Malaria Through the Sensitivity Analysis of a Mathematical Model*, in preparation.

[9] J. M. Cushing, *An Introduction to Structured Population Dynamics*, CBMS-NSF Reg. Conf. Ser. Appl. Math. 71, SIAM, Philadelphia, 1998.

[10] O. Diekmann, J.A.P. Heesterbeek, and J.A.J. Metz, *On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations*, J. Math. Biol., 28 (1990), pp. 365–382.

[11] K. Dietz, L. Molineaux, and A. Thomas, *A malaria model tested in the African savannah*, Bull. World Health Organ., 50 (1974), pp. 347–357.

[12] E. J. Doedel, R. C. Paffenroth, A. R. Champneys, T. F. fairgrieve, Y. A. Kuznetsov, B. Sandstede, and X. Wang, *AUTO* 2000: *Continuation and Bifurcation Software for Ordinary Differential Equations (with HomCont)*, v.0.9.7, 2002; online at http://sourceforge.net/projects/auto2000/.

[13] H. Gómez-Acevedo and M. Y. Li, *Backward bifurcation in a model for HTLV-*I *infection of CD4+ T cells*, Bull. Math. Biol., 67 (2005), pp. 101–114.

[14] J. M. Hyman and J. Li, *An intuitive formulation for the reproductive number for the spread of diseases in heterogeneous populations*, Math. Biosci., 167 (2000), pp. 65–86.
[15] J. C. Koella, *On the use of mathematical models of malaria transmission*, Acta Tropica, 49 (1991), pp. 1–25.
[16] J. C. Koella and R. Antia, *Epidemiological models for the spread of anti-malarial resistance*, Malaria J., 2 (2003).
[17] J. C. Koella and C. Boëte, *A model for the coevolution of immunity and immune evasion in vector-borne disease with implications for the epidemiology of malaria*, The American Naturalist, 161 (2003), pp. 698–707.
[18] G. A. Korn and T. M. Korn, *Mathematical Handbook for Scientists and Engineers: Definitions, Theorems, and Formulas for Reference and Review*, Dover Publications, Mineola, NY, 2000.
[19] J. Li, R. M. Welch, U. S. Nair, T. L. Sever, D. E. Irwin, C. Cordon-Rosales, and N. Padilla, *Dynamic malaria models with environmental changes*, in Proceedings of the Thirty-Fourth Southeastern Symposium on System Theory, Huntsville, AL, 2002, pp. 396–400.
[20] G. Macdonald, *The Epidemiology and Control of Malaria*, Oxford University Press, London, 1957.
[21] J. Nedelman, *Introductory review: Some new thoughts about some old malaria models*, Math. Biosci., 73 (1985), pp. 159–182.
[22] G. A. Ngwa, *Modelling the dynamics of endemic malaria in growing populations*, Discrete Contin. Dyn. Syst. Ser. B, 4 (2004), pp. 1173–1202.
[23] G. A. Ngwa and W. S. Shu, *A mathematical model for endemic malaria with variable human and mosquito populations*, Math. Comput. Modelling, 32 (2000), pp. 747–763.
[24] P. H. Rabinowitz, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.
[25] R. Ross, *The Prevention of Malaria*, John Murray, London, 1911.
[26] P. van den Driessche and J. Watmough, *A simple SIS epidemic model with a backward bifurcation*, J. Math. Biol., 40 (2000), pp. 525–540.
[27] H. M. Yang, *Malaria transmission model for different levels of acquired immunity and temperature-dependent parameters (vector)*, Revista de Saúde Pública, 34 (2000), pp. 223–231.
[28] H. M. Yang and M. U. Ferreira, *Assessing the effects of global warming and local social and economic conditions on the malaria transmission*, Revista de Saúde Pública, 34 (2000), pp. 214–222.

SIAM J. APPL. MATH.
Vol. 67, No. 1, pp. 46–68

DERIVATION OF NEW QUANTUM HYDRODYNAMIC EQUATIONS USING ENTROPY MINIMIZATION*

ANSGAR JÜNGEL†, DANIEL MATTHES†, AND JOSIPA PINA MILIŠIĆ†

Abstract. New quantum hydrodynamic equations are derived from a Wigner–Boltzmann model, using the quantum entropy minimization method recently developed by Degond and Ringhofer. The model consists of conservation equations for the carrier, momentum, and energy densities. The derivation is based on a careful expansion of the quantum Maxwellian in powers of the Planck constant. In contrast to the standard quantum hydrodynamic equations derived by Gardner, the new model includes vorticity terms and a dispersive term for the velocity. Numerical current-voltage characteristics of a one-dimensional resonant tunneling diode for both the new quantum hydrodynamic equations and Gardner's model are presented. The numerical results indicate that the dispersive velocity term regularizes the solution of the system.

Key words. quantum moment hydrodynamics, entropy minimization, quantum Maxwellian, moment method, finite-difference discretization, numerical simulations, resonant tunneling diode, current-voltage characteristics

AMS subject classifications. 35Q40, 65M06, 76Y05

DOI. 10.1137/050644823

1. Introduction.

Quantum phenomena in semiconductor devices are increasingly important, as the characteristic lengths of modern devices are of the order of only deca-nanometers. In fact, there are devices, like resonant tunneling diodes, whose behavior is essentially based on quantum effects. Since the numerical solution of the Schrödinger equation (or one of its approximations) or the Wigner equation is very time-consuming, fluid-type quantum models seem to provide a compromise between accurate and efficient numerical simulations. Moreover, quantum fluid models have several advantages. First, they are formulated in macroscopic quantities like the current density, which can be measured. Second, for the macroscopic quantum models, the same types of boundary conditions are commonly employed as for their classical counterparts.

A fluid dynamical formulation of the Schrödinger equation has been known since the early years of quantum mechanics [26]. A simple derivation uses WKB wave functions $\psi = \sqrt{n}\exp(iS/\varepsilon)$ for the electron density $n(x,t)$ and the phase $S(x,t)$, where ε is the scaled Planck constant. Separating the real and the imaginary parts of the single-state Schrödinger equation gives Euler-type equations for n and the "velocity" $u = \nabla S$, which are called the *quantum hydrodynamic* (*QHD*) *model*. These equations include the so-called Bohm potential $\Delta\sqrt{n}/\sqrt{n}$ as a quantum correction [17, 20]. In the semiclassical limit $\varepsilon \to 0$, the classical pressureless Euler equations are recovered.

In order to incorporate many-particle effects, we are aware of two approaches. The first approach starts from the mixed-state Schrödinger–Poisson system [17, 20]. Defining the particle and current densities as the superpositions of all single-state

*Received by the editors November 10, 2005; accepted for publication (in revised form) July 24, 2006; published electronically November 3, 2006. The authors acknowledge partial support from the Deutsche Forschungsgemeinschaft (DFG), grants JU359/3 (Gerhard-Hess Award) and JU359/5 (Priority Program "Multi-Scale Problems"), and from the Forschungsfond of the University of Mainz.
http://www.siam.org/journals/siap/67-1/64482.html

†Institut für Mathematik, Universität Mainz, Staudingerweg 9, 55099 Mainz, Germany (juengel@mathematik.uni-mainz.de, matthes@mathematik.uni-mainz.de, milisic@mathematik.uni-mainz.de).

densities, quantum equations for the macroscopic variables (particle density, current density, and energy density) are derived. The system of equations is closed by expressing the heat flux heuristically in terms of the macroscopic variables.

The second approach starts from the (collisional) Wigner equation in position-momentum space,

$$\partial_t f + p\cdot\nabla_x f + \theta[V]f = Q(f), \qquad (x,p)\in\mathbb{R}^{2d},\ t>0, \tag{1.1}$$

where (x,p) is the position-momentum variable, $t>0$ is the time, and $\theta[V]$ is a pseudodifferential operator defined by

$$(\theta[V]w)(x,p,t)$$
$$= \frac{i}{(2\pi)^{d/2}}\int_{\mathbb{R}^{2d}}\frac{1}{\varepsilon}\left[V\left(x+\frac{\varepsilon}{2}\eta,t\right)-V\left(x-\frac{\varepsilon}{2}\eta,t\right)\right]w(x,p',t)e^{i\eta\cdot(p-p')}d\eta dp'.$$

The electric potential $V=V(x,t)$ is self-consistently coupled to the Wigner function $f(x,p,t)$ via Poisson's equation

$$\lambda^2\Delta V = \int_{\mathbb{R}^d} f dp - C, \tag{1.2}$$

where λ is the scaled Debye length and $C=C(x)$ the doping concentration characterizing the semiconductor device. Notice that the collisionless Wigner equation is formally equivalent to the Heisenberg equation for the density matrix.

The above approach allows for an abstract formulation of the collision operator. In fact, we assume only that its kernel consists of the quantum thermal equilibrium distribution (defined in section 2) and that the operator preserves certain moments.

The macroscopic variables are defined as the moments of the Wigner function over momentum space; more precisely, we consider the particle density $n=\langle 1\rangle$, the fluid-dynamical momentum density $nu=\langle p\rangle$, and the energy density $e=\langle\frac{1}{2}|p|^2\rangle$, where we have used the notation $\langle g(p)\rangle=\int f(\cdot,p)g(p)dp$ for functions $g(p)$. In order to obtain macroscopic equations as well, a moment method is applied to (1.1): we multiply the equation by 1, p, and $\frac{1}{2}|p|^2$ and integrate over the momentum space. This yields evolution equations for n, nu, and e. However, the resulting system of moment equations needs to be closed.

As a closure condition, Gardner [12] employed a quantum-corrected thermal equilibrium distribution function in place of f in the derivation of the moment equations. The use of this closure can be—formally—justified by a hydrodynamic scaling and passage to the limit of vanishing scaling parameter. Gardner bases his choice of the quantum equilibrium distribution on a result by Wigner [31]. Arguing that the electric potential is close to $\log n$ near equilibrium, he replaces V by $\log n$, which is the origin of the Bohm potential.

Another approach, avoiding second derivatives of the potential, consists of deriving an approximate solution to the Bloch equation by an asymptotic expansion of the solution for "small" potentials. This leads to the so-called smooth QHD equations in which the potential V is replaced by a smoothed potential $S[V]$, where S is a pseudodifferential operator [14]. The drawback of this approach is that the numerical solution of the "smooth" QHD model is a nontrivial task. Moreover, there is an ambiguity in the interpretation of the temperature (see the remark in section 6 of [28]).

Our approach to defining a closure is based on Levermore's entropy minimization principle. This method was first employed in the context of classical gas dynamics [25]

and has been recently extended to quantum fluids by Degond and Ringhofer [9]. The idea is to define the equilibrium distribution as the minimizer M_f of the quantum entropy subject to the constraints of given moments. (Here, we adopt the mathematical sign convention of decreasing entropy.) The minimizer is called the *quantum Maxwellian* since there are some similarities to the classical Maxwellian of gas dynamics (see section 2). The quantum Maxwellian M_f, as the solution of a constrained minimization problem, depends on Lagrange multipliers which can be interpreted in the $O(\varepsilon^2)$ approximation as the logarithm of the particle density, the fluid velocity, and the temperature, respectively. Expanding M_f in powers of ε^2 and assuming as in [12] that spatial variations of the temperature $T = T(x,t)$ are of the order $O(\varepsilon^2)$, we derive the following QHD equations up to order $O(\varepsilon^4)$:

$$\partial_t n + \operatorname{div}(nu) = 0, \tag{1.3}$$

$$\partial_t(nu) + \operatorname{div}(nu\otimes u) + \operatorname{div} P - n\nabla V = 0, \tag{1.4}$$

$$\partial_t e + \operatorname{div}\big((P+eI)u\big) + \operatorname{div} S - nu\cdot\nabla V = 0, \tag{1.5}$$

where I is the unit matrix in $\mathbb{R}^d$; the energy density equals

$$e = \frac{d}{2}nT + \frac{1}{2}n|u|^2 - \frac{\varepsilon^2}{24}n\bigg(\Delta\log n - \frac{1}{T}\operatorname{tr}(R^\top R)\bigg),$$

with the trace "tr" of a matrix; the quantities P (stress tensor) and S are given by

$$\begin{aligned}
P &= nTI - \frac{\varepsilon^2}{12}n\bigg((\nabla\otimes\nabla)\log n - \frac{1}{T}R^\top R\bigg),\\
S &= -\frac{\varepsilon^2}{12}n\bigg(\bigg(\frac{d}{2}+1\bigg)R\nabla\log n + \bigg(\frac{d}{2}+2\bigg)\operatorname{div} R + \frac{3}{2}\Delta u\bigg)\\
&\quad + \frac{\varepsilon^2}{12}\bigg(\frac{d}{2}+1\bigg)n(R\nabla\log n + \operatorname{div} R);
\end{aligned}$$

and the vorticity matrix $R = (R_{ij})$ is the antisymmetric part of the velocity derivative,

$$R_{ij} = \partial_{x_j}u_i - \partial_{x_i}u_j. \tag{1.6}$$

A more general model, allowing arbitrarily large spatial deviations of the temperature, is derived in section 3. Employing a Caldeira–Leggett-type collision operator, relaxation-time terms can also be included (see section 3.1).

The quantum correction $(\varepsilon^2/12)n(\nabla\otimes\nabla)\log n$ to the stress tensor in the QHD equations was first stated in the semiconductor context by Ancona and Iafrate [1] and Ancona and Tiersten [2]. Since

$$\frac{\varepsilon^2}{12}\operatorname{div}(n(\nabla\otimes\nabla)\log n) = \frac{\varepsilon^2}{6}n\nabla\bigg(\frac{\Delta\sqrt{n}}{\sqrt{n}}\bigg),$$

the quantum correction can be interpreted as a force including the Bohm potential $\Delta\sqrt{n}/\sqrt{n}$ [11]. The hydrodynamic formulation of quantum mechanics has been employed in solid-state physics for many years; see, for instance, [18] and the references in the review [22].

For $\varepsilon = 0$ in (1.3)–(1.5), we recover the classical hydrodynamic equations. For $\varepsilon > 0$ and constant temperature, we obtain the same equations as derived in [21],

where also the quantum entropy minimization method has been used. Our model differs from Gardner's QHD equations (formulas (1)–(3) in [12]) by the vorticity term R and the dispersive velocity term in the energy equation (1.5),

$$\operatorname{div} q_S = \frac{\varepsilon^2}{8}\operatorname{div}(n\Delta u). \tag{1.7}$$

The origin of this difference lies in the different choices of the quantum Maxwellian. We refer to section 3.5 for a detailed discussion.

The term q_S—but not the vorticity R—also appears in other QHD derivations. It was derived in [13] from a mixed-state Wigner model and interpreted as a dispersive "heat flux" (see formula (36) in [13]). Moreover, it appears in the QHD equations of [16] involving a "smoothed" potential, derived from the Wigner–Boltzmann equation by a Chapman–Enskog expansion.

An interesting feature of the dispersive term (1.7) is that it stabilizes the QHD system numerically. This statement needs some explanation. It is known that the numerical approximation of Gardner's QHD model (see (6.5)–(6.7)) is quite delicate. The usual approach is to employ a hyperbolic solver, for instance an upwind method [12] or a shock-capturing discontinuous Galerkin method [6], originally devised for the classical hydrodynamic equations. It has been argued in [24] that a hyperbolic solver may be inadequate for the QHD equations since the *numerical viscosity* might destroy the dispersive quantum effects. Therefore, a central finite-difference scheme provides an alternative (but still simple) numerical approach. In fact, a central finite-difference approach for Gardner's QHD equations fails, and a stabilization in the form of numerical viscosity seems to be necessary. The dispersive term (1.7) allows us to solve the new QHD equations by using a *central* scheme, thus avoiding numerical viscosity.

Another QHD model with *physical viscosity* has been derived in [19] using a Fokker–Planck collision operator. This operator describes the interaction of the electrons with a heat bath modeling the phonons of the semiconductor lattice. In numerical simulations of a resonant tunneling diode, it turns out that the shape of the current-voltage characteristic is unphysical if the temperature is kept constant [24], and that the diffusion effects are too strong compared to the quantum dispersion [23].

In this paper we present the first numerical simulations of a QHD model involving the term (1.7). More precisely, a simple one-dimensional resonant tunneling diode is simulated. The current-voltage characteristics show multiple regions of negative differential resistance. The dispersive term (1.7) has the effect of "smoothing" the current-voltage curve; i.e., it decreases the peak-to-valley ratio, the quotient of the peak to the valley current.

We also examine the existence of conserved quantities of the new QHD equations. Clearly, the mass is conserved. We prove that also the energy $E = \int (e+\lambda^2|\nabla V|^2/2)dx$ is conserved. This provides gradient estimates for the particle density, velocity, and temperature, which is useful in the mathematical analysis of the equations.

We summarize the advantages of our approach:

- Starting from the Wigner–BGK (Bhatganar–Gross–Krook) equation, no ad hoc assumptions are needed in order to derive the QHD equations.
- An energy for the new model can be defined, leading to useful mathematical estimates.
- The dispersive velocity term seems to stabilize the (numerical) solution of the system.
- The new model provides current-voltage characteristics showing negative differential resistance effects.

The paper is organized as follows. In section 2 we specify our definition of the quantum Maxwellian, which is used as the closure in the moment method developed in section 3. Section 4 is devoted to simplified QHD models, and the system (1.3)–(1.5) is derived. In section 5 we prove that the energy of the system is conserved. Finally, in section 6, the new QHD model (1.3)–(1.5) is numerically discretized and solved in one space dimension, and simulations of a resonant tunneling diode are presented.

2. Definition of the quantum Maxwellian. In order to define the quantum Maxwellian, we first recall the Wigner transform. Let A_ρ be an operator on $L^2(\mathbb{R}^d)$ with integral kernel $\rho(x,x')$, i.e.,

$$(A_\rho\phi)(x) = \int_{R^d} \rho(x,x')\phi(x')dx' \quad \text{for all } \phi \in L^2(\mathbb{R}^d).$$

The Wigner transform of A_ρ is defined by

$$W(A_\rho)(x,p) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \rho\left(x+\frac{\varepsilon}{2}\eta, x-\frac{\varepsilon}{2}\eta\right)e^{i\eta\cdot p}d\eta.$$

Its inverse W^{-1}, also called Weyl quantization, is defined as an operator on $L^2(\mathbb{R}^d)$:

$$(W^{-1}(f)\phi)(x) = \int_{\mathbb{R}^{2d}} f\left(\frac{x+y}{2}\right)\phi(y)e^{ip\cdot(x-y)/\varepsilon}dpdy \quad \text{for all } \phi \in L^2(\mathbb{R}^d).$$

With these definitions we are able to introduce the *quantum exponential* and the *quantum logarithm* formally by

$$\exp f = W(\exp W^{-1}(f)), \qquad \operatorname{Log} f = W(\log W^{-1}(f)),$$

where exp and log are the operator exponential and logarithm, respectively. In [8] it was (formally) shown that the quantum exponential and quantum logarithm are equal to the usual exponential and logarithm, respectively, up to order $O(\varepsilon^2)$,

$$\exp f = \exp f + O(\varepsilon^2), \qquad \operatorname{Log} f = \log f + O(\varepsilon^2). \tag{2.1}$$

The essential ingredient in the definition of the quantum Maxwellian is the relative quantum entropy. Let a quantum mechanical state be described by the Wigner function f solving the Wigner equation (1.1). Then its *relative quantum* (*von Neumann*) *entropy* is given by [9]

$$H(f) = \int_{\mathbb{R}^{2d}} f(x,p)\left((\operatorname{Log} f)(x,p) - 1 + \frac{|p|^2}{2} - V(x)\right)dxdp.$$

Whereas the classical entropy is a function on the configuration space, the above quantum entropy is a real number, underlining the nonlocal nature of quantum mechanics.

We define the quantum thermal equilibrium or *quantum Maxwellian* M_f for some given function $f(x,p)$ as the solution of the constrained minimization problem

$$H(M_f) = \min\left\{H(\hat{f}) : \int_{\mathbb{R}^d} \hat{f}(x,p,t)\begin{pmatrix}1\\ p\\ |p|^2/2\end{pmatrix}dp = \begin{pmatrix}n(x,t)\\ nu(x,t)\\ e(x,t)\end{pmatrix},\ x\in\mathbb{R}^d,\ t>0\right\}, \tag{2.2}$$

where

$$n(x,t) = \langle 1\rangle(x,t) = \int_{\mathbb{R}^d} f(x,p,t)dp,$$
$$nu(x,t) = \langle p\rangle(x,t) = \int_{\mathbb{R}^d} f(x,p,t)pdp,$$
$$e(x,t) = \frac{1}{2}\langle |p|^2\rangle(x,t) = \frac{1}{2}\int_{\mathbb{R}^d} f(x,p,t)|p|^2dp.$$

In [9] it is shown that the solution f^* of the constrained minimization problem (if it exists) is given by

$$M_f(x,p,t) = \exp\left(A(x,t) - \frac{|p-w(x,t)|^2}{2T(x,t)}\right). \tag{2.3}$$

The Lagrange multipliers A, w, and T are uniquely determined by the moments of f. They correspond in the classical setting to the logarithm of the particle density, the velocity, and the temperature, respectively (see Lemma 3.4).

3. Derivation of the general QHD model. The derivation of the new QHD equations is done in several steps. First, we derive the moment equations. Then the quantum exponential is expanded in powers of ε^2 up to order $O(\varepsilon^4)$. The third step is to expand the moments accordingly. Finally, the expansions are substituted into the moment equations.

3.1. Moment equations. We consider the Wigner–Boltzmann equation (1.1) in the hydrodynamic scaling; i.e., we introduce the scaling

$$x' = \delta x, \quad t' = \delta t,$$

for some parameter $\delta > 0$, which is assumed to be small compared to 1. Then (1.1) becomes for $f = f_\delta$ (omitting the primes)

$$\partial_t f_\delta + p\cdot\nabla_x f_\delta + \theta[V]f_\delta = \delta^{-1}Q(f_\delta), \quad (x,p)\in\mathbb{R}^{2d},\ t>0, \tag{3.1}$$

with initial condition $f_\delta(x,p,0) = f_I(x,p)$. We assume that the collision operator has the following properties: its kernel consists exactly of (multiples of) M_f and

$$\int_{\mathbb{R}^d} Q(f)dp = 0, \quad \int_{\mathbb{R}^d} Q(f)pdp = 0, \quad \int_{\mathbb{R}^d} Q(f)\tfrac{1}{2}|p|^2dp = 0 \quad \text{for all } f(x,p). \tag{3.2}$$

An example satisfying these conditions is the relaxation-time or BGK operator $Q(f) = M_f - f$ (with scaled relaxation time $\tau = 1$) [4].

A more general collision operator, allowing for relaxation-time terms in the macroscopic equations, can be defined as follows. Assume that the collision operator can be written as $Q(f) = Q_0(f) + \delta Q_1(f)$, where the operator $Q_0(f)$ models elastic collisions and satisfies the conditions (3.2); $Q_1(f)$ is given by the Caldeira–Leggett operator [5]

$$Q_1(f) = \frac{1}{\tau_p}(\mathrm{div}_p(pf) + \Delta_p f),$$

modeling inelastic collisions; and τ_p is the momentum relaxation time. Then

$$\int_{\mathbb{R}^d} Q_1(f)dp = 0, \quad \int_{\mathbb{R}^d} Q_1(f)pdp = -\frac{nu}{\tau_p}, \quad \int_{\mathbb{R}^d} Q_1(f)\tfrac{1}{2}|p|^2dp = -\frac{2}{\tau_p}\left(e - \frac{d}{2}n\right), \tag{3.3}$$

which is (a special case of) the momentum and energy relaxation-time terms employed in [12].

The formal limit $\delta \to 0$ in (3.1) yields $Q(f) = 0$, where $f = \lim_{\delta\to 0} f_\delta$, which implies that the limit f is equal to M_f. The moment equations are obtained from (3.1) by multiplication with 1, p, and $\frac{1}{2}|p|^2$, respectively, and integration over the momentum space. Since

$$\int_{\mathbb{R}^d} \theta[V] f dp = 0, \quad \int_{\mathbb{R}^d} \theta[V] f p dp = -n\nabla V, \quad \int_{\mathbb{R}^d} \theta[V] f \tfrac{1}{2}|p|^2 dp = -nu\cdot\nabla V$$

(see, e.g., [8]), we obtain

$$\partial_t n + \operatorname{div}(nu) = 0, \tag{3.4}$$

$$\partial_t (nu) + \operatorname{div}\langle p\otimes p\rangle - n\nabla V = 0, \tag{3.5}$$

$$\partial_t e + \operatorname{div}\langle \tfrac{1}{2}|p|^2 p\rangle - nu\cdot\nabla V = 0, \tag{3.6}$$

where $(p\otimes p)_{ij} = p_i p_j$ for $i,j = 1,\ldots,d$. Recall that the brackets denote integration against the Wigner function $f = M_f$; i.e., in multi-index notation,

$$\langle p^\alpha\rangle(x,t) = \int_{\mathbb{R}^d} M_f(x,p,t) p^\alpha dp,$$

for multi-indices $\alpha \in \mathbb{N}^d$. When employing the Caldeira–Leggett operator defined above, the right-hand sides of (3.4)–(3.6) equal the right-hand sides of (3.3). To close the system (3.4)–(3.6), we need to express the integrals $\langle p\otimes p\rangle$ and $\langle \frac{1}{2}|p|^2 p\rangle$ in terms of the moments n, nu, and e. This constitutes the main step of the derivation.

The following computations are simplified by working with the new variable $s = T^{-1/2}(p-w)$, where w is the Lagrange multiplier introduced in (2.3). In terms of s, the quantum Maxwellian reads as

$$M_f(x,p(s)) = \exp\left(A(x) - \frac{1}{2}|s|^2\right) =: g(x,s).$$

From now on, we omit the dependence of the time t since it acts only as a parameter. The substitution $p \mapsto s$ yields

$$\langle s^\alpha\rangle(x) = T^{d/2}\int_{\mathbb{R}^d} g(x,s) s^\alpha ds.$$

In the following lemma we express the moments $\langle p^\alpha\rangle$ in terms of moments in s. This allows for a more canonical form of the QHD equations.

LEMMA 3.1. *The system* (3.4)–(3.6) *is equivalent to*

$$\partial_t n + \operatorname{div}(nu) = 0, \tag{3.7}$$

$$\partial_t (nu) + \operatorname{div}(nu\otimes u) + \operatorname{div} P - n\nabla V = 0, \tag{3.8}$$

$$\partial_t e + \operatorname{div}\big((P+eI)u\big) + \operatorname{div} S - nu\cdot\nabla V = 0, \tag{3.9}$$

where I is the identity matrix, $u = (nu)/n$, $P = \langle (p-u)\otimes(p-u)\rangle$ is the stress tensor, and $S = \langle \frac{1}{2}(p-u)|p-u|^2\rangle$ is the (quantum) heat flux. Moreover, the following expansions hold:

$$P = T\langle s\otimes s\rangle + O(\varepsilon^4), \quad S = \frac{1}{2}T^{3/2}\langle |s|^2 s\rangle - \left(\frac{d}{2}+1\right)T^{3/2}\langle s\rangle + O(\varepsilon^4). \tag{3.10}$$

Proof. The formulation (3.7)–(3.9) follows immediately from (3.4)–(3.6) since

$$\langle p\otimes p\rangle = P + nu\otimes u \quad \text{and} \quad \left\langle \frac{1}{2}|p|^2 p\right\rangle = S + (P+eI)u.$$

Using the expansion (2.1), elementary integrations yield for $i,j=1,\dots,d$,

$$\langle 1\rangle = T^{d/2}e^A \int_{\mathbb{R}^d} e^{-|s|^2/2}ds + O(\varepsilon^2) = (2\pi T)^{d/2}e^A + O(\varepsilon^2), \tag{3.11}$$

$$\langle s_i\rangle = T^{d/2}e^A \int_{\mathbb{R}^d} e^{-|s|^2/2}s_i ds + O(\varepsilon^2) = O(\varepsilon^2), \tag{3.12}$$

$$\langle s_i s_j\rangle = T^{d/2}e^A \int_{\mathbb{R}^d} e^{-|s|^2/2}s_i s_j ds + O(\varepsilon^2) = n\delta_{ij} + O(\varepsilon^2). \tag{3.13}$$

The relations $n=\langle 1\rangle$, $\langle w\rangle = w\langle 1\rangle = nw$, and $nu = \langle p\rangle = \langle T^{1/2}s + w\rangle = T^{1/2}\langle s\rangle + nw$ give for the second moments

$$\begin{aligned}
\langle p\otimes p\rangle &= T\langle s\otimes s\rangle + \Big\langle (T^{1/2}s+w)\otimes(T^{1/2}s+w) - (T^{1/2}s)\otimes(T^{1/2}s)\Big\rangle \\
&= T\langle s\otimes s\rangle + T^{1/2}\langle s\rangle\otimes w + T^{1/2}w\otimes\langle s\rangle + w\otimes w\langle 1\rangle \\
&= T\langle s\otimes s\rangle + \frac{1}{n}\langle T^{1/2}s+w\rangle\otimes\langle T^{1/2}s+w\rangle - \frac{T}{n}\langle s\rangle\otimes\langle s\rangle \\
&= T\langle s\otimes s\rangle + nu\otimes u + O(\varepsilon^4),
\end{aligned}$$

where in the last equality we have employed (3.12). Therefore, $P = \langle p\otimes p\rangle - nu\otimes u = T\langle s\otimes s\rangle + O(\varepsilon^4)$. In a similar way, we compute the third moment:

$$\begin{aligned}
\frac{1}{2}\langle |p|^2 p\rangle &= \frac{1}{2}T^{1/2}\langle |T^{1/2}s+w|^2 s\rangle + \frac{1}{2}w\langle |p|^2\rangle \\
&= \frac{1}{2}T^{3/2}\langle |s|^2 s\rangle + T\langle s\otimes s\rangle w + \frac{1}{2}T^{1/2}|w|^2\langle s\rangle + ew \\
&= \frac{1}{2}T^{3/2}\langle |s|^2 s\rangle + (P+eI)w + \frac{1}{2}T^{1/2}|w|^2\langle s\rangle.
\end{aligned}$$

By (3.11) and (3.12), the energy density can be expanded as

$$e = \frac{1}{2}\langle |p|^2\rangle = \frac{T}{2}\langle |s|^2\rangle + T^{1/2}w\cdot\langle s\rangle + \frac{1}{2}|w|^2\langle 1\rangle = \frac{d}{2}nT + \frac{1}{2}n|w|^2 + O(\varepsilon^2).$$

Thus, since $w = u - T^{1/2}\langle s\rangle/n$ and $P = nTI + O(\varepsilon^2)$, we obtain

$$\begin{aligned}
\frac{1}{2}\langle |p|^2 p\rangle &= \frac{1}{2}T^{3/2}\langle |s|^2 s\rangle + (P+eI)u - \frac{T^{1/2}}{n}\left(P + eI - \frac{1}{2}n|w|^2 I\right)\langle s\rangle \\
&= \frac{1}{2}T^{3/2}\langle |s|^2 s\rangle + (P+eI)u - \frac{T^{3/2}}{n}\left(\left(\frac{d}{2}+1\right)nI + O(\varepsilon^2)\right)\langle s\rangle.
\end{aligned}$$

This shows that $S = \langle \frac{1}{2}|p|^2 p\rangle - (P+eI)u = \frac{1}{2}T^{3/2}\langle |s|^2 s\rangle - (d/2+1)T^{3/2}n\langle s\rangle + O(\varepsilon^4)$. ∎

3.2. Expansion of the quantum exponential. We wish to give asymptotic expansions of P, S, and U up to order $O(\varepsilon^4)$. For this, we first need to expand the quantum Maxwellian. This is done by means of the following lemma, which is adopted from [9].

LEMMA 3.2. *Let $f(x,p)$ be a smooth symbol. Then the quantum exponential* $\exp f$ *can be expanded as follows:*

$$\exp f = e^f - \frac{\varepsilon^2}{8} e^f \mathcal{Q} + O(\varepsilon^4),$$

where, using Einstein's summation convention,

$$\begin{aligned}\mathcal{Q} &= \partial^2_{x_ix_j} f\, \partial^2_{p_ip_j} f - \partial^2_{x_ip_j} f\, \partial^2_{p_ix_j} f + \frac{1}{3}\partial^2_{x_ix_j} f\, \partial_{p_i} f\, \partial_{p_j} f \\ &\quad - \frac{2}{3}\partial^2_{x_ip_j} f\, \partial_{p_i} f\, \partial_{x_j} f + \frac{1}{3}\partial^2_{p_ip_j} f\, \partial_{x_i} f\, \partial_{x_j} f. \end{aligned} \tag{3.14}$$

In the situation at hand, the symbol is $f(x,p) = A(x) - |p - w(x)|^2/2T(x)$. Then we obtain the following result.

LEMMA 3.3. *The quantum correction* (3.14) *can be written for $f(x,p) = A(x) - |p-w(x)|^2/2T(x)$ as follows:*

$$\begin{aligned}\mathcal{Q}(s) &= T^{-1}\big(X^0 + X^1_i s_i + X^2_{ij} s_i s_j + X^3_{ijk} s_i s_j s_k \\ &\quad + Y^0|s|^2 + Y^1_i |s|^2 s_i + Y^2_{ij}|s|^2 s_i s_j + Z^0|s|^4\big), \end{aligned} \tag{3.15}$$

where the coefficients X^i, Y^i, and Z are defined by

$$\begin{aligned} X^0 &= -\Delta A - \frac{1}{3}|\nabla A|^2 + \frac{1}{2T}\mathrm{tr}\,(\widetilde{R}^\top \widetilde{R}), \\ X^1_i &= \frac{2}{T^{1/2}}\partial_{x_m}\left(\frac{1}{3}A - \log T\right)\widetilde{R}_{mi} - \frac{1}{\sqrt{T}}\Delta w_i, \\ X^2_{ij} &= \frac{1}{3}\partial^2_{x_ix_j} A + \frac{2}{3}\partial_{x_i}(\log T)\partial_{x_j} A - \partial_{x_i}(\log T)\partial_{x_j}(\log T) - \frac{1}{3T}(\widetilde{R}^\top \widetilde{R})_{ij} \\ X^3_{ijk} &= \frac{1}{3T^{1/2}}\partial^2_{x_ix_j} w_k, \\ Y^0 &= \nabla\left(\frac{1}{2}\log T - \frac{1}{3}A\right)\cdot\nabla(\log T) - \frac{1}{2}\Delta(\log T), \\ Y^1_i &= \frac{1}{3T^{1/2}}\partial_{x_m}(\log T)\widetilde{R}_{mi}, \\ Y^2_{ij} &= \frac{1}{6}\left(\partial^2_{x_ix_j}(\log T) + \partial_{x_i}(\log T)\partial_{x_j}(\log T)\right), \\ Z^0 &= -\frac{1}{12}|\nabla(\log T)|^2, \end{aligned}$$

and $\widetilde{R}_{ij} = \partial_{x_j} w_i - \partial_{x_i} w_j$. The symbol "tr" denotes the trace of a matrix.

Proof. The proof consists of computing the relevant derivatives of f with respect to x_i and p_j, namely,

$$\begin{aligned}
\partial_{x_i} f &= \partial_{x_i} A + T^{-1}\partial_{x_i} w_k (p-w)_k + \frac{1}{2}T^{-2}\partial_{x_i} T|p-w|^2 \\
&= \partial_{x_i} A + T^{-1/2}\partial_{x_i} w_k s_k + \frac{1}{2}T^{-1}\partial_{x_i} T|s|^2, \\
\partial^2_{x_i x_j} f &= \partial^2_{x_i x_j} A - T^{-1}\partial_{x_i} w_k \partial_{x_j} w_k - T^{-2}\partial_{x_j} T\partial_{x_i} w_k (p-w)_k + T^{-1}\partial^2_{x_i x_j} w_k (p-w)_k \\
&\quad - T^{-2}\partial_{x_i} T\partial_{x_j} w_k (p-w)_k - T^{-3}\partial_{x_i} T\partial_{x_j} T|p-w|^2 + \frac{1}{2}T^{-2}\partial^2_{x_i x_j} T|p-w|^2 \\
&= \partial^2_{x_i x_j} A - T^{-1}\partial_{x_i} w_k \partial_{x_j} w_k - T^{-3/2}\partial_{x_j} T\partial_{x_i} w_k s_k + T^{-1/2}\partial^2_{x_i x_j} w_k s_k \\
&\quad - T^{-3/2}\partial_{x_i} T\partial_{x_j} w_k s_k - T^{-2}\partial_{x_i} T\partial_{x_j} T|s|^2 + \frac{1}{2}T^{-1}\partial^2_{x_i x_j} T|s|^2, \\
\partial_{p_i} f &= -T^{-1}(p-w)_i = -T^{-1/2}s_i, \\
\partial^2_{p_i x_j} f &= T^{-1}\partial_{x_j} w_i + T^{-2}\partial_{x_j} T(p-w)_i = T^{-1}\partial_{x_j} w_i + T^{-3/2}\partial_{x_j} T s_i, \\
\partial^2_{p_i p_j} f &= -T^{-1}\delta_{ij},
\end{aligned}$$

and the products appearing in the sum (3.14), which are

$$\begin{aligned}
\partial^2_{x_i x_j} f\, \partial^2_{p_i p_j} F &= \big(- T^{-1}\Delta A - T^{-3/2}\Delta w_k + T^{-2}\|\nabla w\|^2 + 2T^{-5/2}\nabla T\cdot\nabla w_k\big)s_k \\
&\quad + \left(\frac{1}{2}T^{-2}\Delta T - T^{-3}|\nabla T|^2\right)|s|^2, \\
\partial^2_{x_i p_j} f\, \partial^2_{p_i x_j} f &= T^{-2}\partial_{x_i} w_j \partial_{x_j} w_i + 2T^{-5/2}\partial_{x_j} T\partial_{x_i} w_j s_i + T^{-3}\partial_{x_i} T\partial_{x_j} T s_i s_j, \\
\partial^2_{x_i x_j} f\, \partial_{p_i} f\, \partial_{p_j} f &= \big(T^{-1}\partial^2_{x_i x_j} A - T^{-2}\partial_{x_i} w_\ell \partial_{x_j} w_\ell\big)s_i s_j \\
&\quad + \big(T^{-3/2}\partial^2_{x_i x_j} w_k - 2T^{-5/2}\partial_{x_i} T\partial_{x_j} w_k\big)s_i s_j s_k \\
&\quad + \left(\frac{1}{2}T^{-2}\partial^2_{x_i x_j} T - T^{-3}\partial_{x_i} T\partial_{x_j} T\right)|s|^2 s_i s_j, \\
\partial^2_{x_i p_j} f\, \partial_{p_i} f\, \partial_{x_j} f &= -T^{-3/2}\partial_{x_\ell} A\partial_{x_i} w_\ell s_i - T^{-2}\partial_{x_i} T\partial_{x_j} A s_i s_j - T^{-2}\partial_{x_i} w_j \partial_{x_j} w_k s_i s_k \\
&\quad - T^{-5/2}\partial_{x_i} T\partial_{x_j} w_k s_i s_j s_k - \frac{1}{2}T^{-5/2}\partial_{x_i} w_j \partial_{x_j} T|s|^2 s_i \\
&\quad - \frac{1}{2}T^{-3}\partial_{x_i} T\partial_{x_j} T|s|^2 s_i s_j, \\
\partial^2_{p_i p_j} f\, \partial_{x_i} f\, \partial_{x_j} f &= -T^{-1}|\nabla A|^2 - 2T^{-3/2}\nabla A\cdot\nabla w_k s_k - T^{-2}\nabla A\cdot\nabla T|s|^2 \\
&\quad - T^{-2}\nabla w_k\cdot\nabla w_\ell s_k s_\ell - T^{-5/2}\nabla T\cdot\nabla w_k|s|^2 s_k - \frac{1}{4}T^{-3}|\nabla T|^2|s|^4.
\end{aligned}$$

Inserting these expressions into (3.14) and simplifying, we arrive at (3.15). ∎

3.3. Expansion of the moments. The aim of this subsection is to specify the integrals $\langle s^\alpha\rangle$ in order to expand the moments n, nu, and e. By Lemma 3.2, we obtain

$$\begin{aligned}
\langle s^\alpha\rangle &= T^{d/2}\int_{\mathbb{R}^d} g(x,s)s^\alpha ds \\
&= T^{d/2}\int_{\mathbb{R}^d} e^{A-|s|^2/2}\left(1-\frac{\varepsilon^2}{8}\mathcal{Q}(s)\right)s^\alpha ds + O(\varepsilon^4) \\
&= (2\pi T)^{d/2}e^A\left([s^\alpha] - \frac{\varepsilon^2}{8}[\mathcal{Q}(s)s^\alpha]\right) + O(\varepsilon^4),
\end{aligned}$$

where $[g]$ denotes the integral of a function $g = g(s)$ against the classical Gaussian kernel,

$$[g] = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-|s|^2/2} g(s) ds.$$

Notice that from the expansion

$$n = \langle 1 \rangle = (2\pi T)^{d/2} e^A \left(1 - \frac{\varepsilon^2}{8} [\mathcal{Q}(s)] \right) + O(\varepsilon^4) \tag{3.16}$$

it follows that

$$\langle s^\alpha \rangle = n \left([s^\alpha] + \frac{\varepsilon^2}{8} \big([\mathcal{Q}(s)][s^\alpha] - [\mathcal{Q}(s) s^\alpha] \big) \right) + O(\varepsilon^4). \tag{3.17}$$

Thus it remains to calculate the integrals $[\mathcal{Q}(s) s^\alpha]$.

Integrals of type $[s^\alpha]$ can be computed explicitly. Using

$$\int_{\mathbb{R}} t^m e^{-t^2/2} dt = \sqrt{2\pi} \times \begin{cases} 0 & \text{if } m \text{ is odd,} \\ 1 & \text{if } m = 0 \text{ or } m = 2, \\ 3 & \text{if } m = 4, \\ 15 & \text{if } m = 6, \end{cases}$$

it becomes a matter of combinatorics to conclude for i, j, m, $n = 1, \ldots, d$,

$$\begin{aligned}
[s_i s_j] &= \delta_{ij}, \\
[|s|^2] &= d, \\
[s_i s_j s_m s_n] &= \delta_{ij}\delta_{mn} + \delta_{im}\delta_{jn} + \delta_{in}\delta_{jm}, \\
[s_i s_j |s|^2] &= (d+2)\delta_{ij}, \\
[|s|^4] &= d(d+2), \\
[s_i s_j s_m s_n |s|^2] &= (d+4)(\delta_{ij}\delta_{mn} + \delta_{im}\delta_{jn} + \delta_{in}\delta_{jm}), \\
[s_m s_n |s|^4] &= (d+2)(d+4)\delta_{mn}.
\end{aligned}$$

Then the expansion of $\mathcal{Q}(s)$, given in (3.15), yields the following formulas:

$$[\mathcal{Q}(s)] = X^0 + \sum_\ell X^2_{\ell\ell} + dY^0 + (d+2) \sum_\ell Y^2_{\ell\ell} + d(d+2) Z^0, \tag{3.18}$$

$$[\mathcal{Q}(s) s_m] = X^1_m + \sum_\ell \left(X^3_{m\ell\ell} + X^3_{\ell m\ell} + X^3_{\ell\ell m} \right) + (d+2) Y^1_m, \tag{3.19}$$

$$\begin{aligned}
[\mathcal{Q}(s) s_m^2] = [\mathcal{Q}(s)] &+ 2X^2_{mm} + 2Y^0 + 2\sum_\ell Y^2_{\ell\ell} + 2(d+4) Y^2_{mm} \\
&+ 4(d+2) Z^0,
\end{aligned} \tag{3.20}$$

$$[\mathcal{Q}(s) s_m s_n] = \left(X^2_{mn} + X^2_{nm} \right) + (d+4) \left(Y^2_{mn} + Y^2_{nm} \right), \tag{3.21}$$

$$\begin{aligned}
[\mathcal{Q}(s)|s|^2 s_m] = (d+2) X^1_m &+ (d+4) \sum_\ell \left(X^3_{m\ell\ell} + X^3_{\ell m\ell} + X^3_{\ell\ell m} \right) \\
&+ (d+2)(d+4) Y^1_m.
\end{aligned} \tag{3.22}$$

LEMMA 3.4. *The moments n, nu, and e can be expressed in terms of the Lagrange multipliers A, w, and T asymptotically as follows:*

$$n = (2\pi T)^{d/2}e^A - \frac{\varepsilon^2}{24T}(2\pi T)^{d/2}e^A\bigg\{ -2\Delta A - |\nabla A|^2 + (d-2)\nabla\log T\cdot\nabla A - (d-1)\Delta\log T - \bigg(\frac{d}{2}-1\bigg)\bigg(\frac{d}{2}-2\bigg) \times|\nabla\log T|^2 + \frac{1}{2T}\mathrm{tr}\,(\widetilde{R}^\top\widetilde{R})\bigg\} + O(\varepsilon^4), \tag{3.23}$$

$$nu = nw + T^{-1}U, \tag{3.24}$$

$$e = \frac{d}{2}nT + \frac{1}{2}n|u|^2 - \frac{\varepsilon^2}{24}n\bigg\{\Delta\log n - \frac{1}{T}\mathrm{tr}\,(\widetilde{R}^\top\widetilde{R}) + \frac{d}{2}|\nabla\log T|^2 - \Delta\log T - \nabla\log T\cdot\nabla\log n\bigg\} + O(\varepsilon^4). \tag{3.25}$$

Notice that (3.23) and (3.24) imply the inverse relations

$$A = \log n - \frac{d}{2}\log T - \frac{d}{2}\log(2\pi) + O(\varepsilon^2), \quad w = u + O(\varepsilon^2). \tag{3.26}$$

In particular, the vorticity matrices $\widetilde{R}$ and R, defined in (1.6), coincide up to order $O(\varepsilon^2)$ since $\widetilde{R}_{ij} = \partial_j u_i - \partial_i u_j + O(\varepsilon^2) = R_{ij} + O(\varepsilon^2)$.

Proof. The formula for the particle density (3.23) is obtained by first substituting the expressions for the coefficients X, Y, and Z into (3.18). This yields $[\mathcal{Q}(s)]$ in terms of A, w, and T. Inserting the result into (3.16) then gives (3.23).

In order to derive (3.24), we write, by the definition of U (see (3.10)),

$$nu = \langle T^{1/2}s + w\rangle = T^{1/2}\langle s\rangle + w\langle 1\rangle = T^{-1}U + nw.$$

Hence, $u - w = U/nT = O(\varepsilon^2)$. The above equations also show that $T^{1/2}w\cdot\langle s\rangle = nu\cdot w - n|w|^2$. Hence, using $\langle 1\rangle = n$,

$$\begin{aligned} e &= \frac{1}{2}\langle|T^{1/2}s + w|^2\rangle = \frac{T}{2}\langle|s|^2\rangle + T^{1/2}w\cdot\langle s\rangle + \frac{1}{2}|w|^2\langle 1\rangle \\ &= \frac{T}{2}\langle|s|^2\rangle + nu\cdot w - \frac{1}{2}n|w|^2 = \frac{T}{2}\langle|s|^2\rangle + \frac{1}{2}n|u|^2 - \frac{1}{2}n|u-w|^2. \end{aligned}$$

In view of (3.26), we have $|u-w|^2 = O(\varepsilon^4)$, from which we conclude

$$e = \frac{T}{2}\langle|s|^2\rangle + \frac{1}{2}n|u|^2 + O(\varepsilon^4).$$

The bracket $\langle|s|^2\rangle$ can be computed from (3.17), employing $[|s|^2] = d$,

$$\langle|s|^2\rangle = dn + \frac{\varepsilon^2}{8}n\sum_m\big([\mathcal{Q}(s)] - [\mathcal{Q}(s)s_m^2]\big) + O(\varepsilon^4).$$

Substitution of (3.18) and (3.20) into the above expression and elimination of A and w, using (3.26), gives $\langle|s|^2\rangle$ in terms of n, nu, and T. This finally leads to (3.25). ∎

3.4. Expansion of the terms P, S, and U. The QHD equations (3.7)–(3.9) are determined by the following expansion of the auxiliary terms P, S, and U, defined in (3.10), in terms of the macroscopic variables n, nu, and e.

LEMMA 3.5. *The following expansion holds:*

$$P = nTI + \frac{\varepsilon^2}{12}n\bigg\{\bigg(\frac{d}{2}+1\bigg)\nabla\log T\otimes\nabla\log T - \nabla\log T\otimes\nabla\log n$$

$$- \nabla\log n\otimes\nabla\log T - (\nabla\otimes\nabla)\log(nT^2) + \frac{R^\top R}{T}\bigg\} \tag{3.27}$$

$$+ \frac{\varepsilon^2}{12}T\mathrm{div}\bigg(n\frac{\nabla\log T}{T}\bigg)I + O(\varepsilon^4),$$

$$S = -\frac{\varepsilon^2}{12}n\bigg\{\bigg(\frac{d}{2}+1\bigg)R\nabla\log\bigg(\frac{n}{T}\bigg) + \bigg(\frac{d}{2}+2\bigg)\mathrm{div}\,R + \frac{3}{2}\Delta u\bigg\} \tag{3.28}$$

$$+ \frac{\varepsilon^2}{12}\bigg(\frac{d}{2}+1\bigg)n\bigg\{R\nabla\log\bigg(\frac{n}{T^2}\bigg) + \mathrm{div}\,R\bigg\} + O(\varepsilon^4). \tag{3.29}$$

Proof. We apply formula (3.17) to obtain for all m, $n = 1,\ldots,d$,

$$P_{mn} = nT\bigg(\delta_{mn} + \frac{\varepsilon^2}{8}(\delta_{mn}[\mathcal{Q}(s)] - [\mathcal{Q}(s)s_m s_n])\bigg),$$
$$S_m = -\frac{\varepsilon^2}{16}nT^{3/2}[\mathcal{Q}(s)|s|^2 s_m] + \frac{\varepsilon^2}{8}\bigg(\frac{d}{2}+1\bigg)nT^{3/2}[\mathcal{Q}(s)s_m].$$

Then the components of P are computed by employing (3.18) and (3.21), substituting the definitions of the coefficients X, Y, and Z, and replacing A and w by n and nu according to (3.26). In a similar way, S is evaluated using (3.19) and (3.22). ∎

3.5. Discussion of the QHD equations. The differences between our QHD equations and Gardner's model can be understood as follows. In both approaches, closure is obtained by assuming that the Wigner function f is in thermal equilibrium. However, the notion of "thermal equilibrium" is different.

In order to illustrate the differences, we recall the classical situation. For a system with the Hamiltonian $h(x,p) = |p|^2/2 + V(x)$, the unconstrained thermal equilibrium distribution is given by the Gibbs measure $f_G(x,p) = \exp(-h(x,p)/T_0)$, which minimizes the relative entropy $S = \int f(\log f - 1 - h/T_0)dp$. Here, T_0 denotes a temperature constant. If mass, momentum, and energy densities are given, the constrained thermal equilibrium is realized by a suitable rescaling and a momentum-shift of the Gibbs state,

$$\widetilde{f}_G(x,p) = n(x)\,\exp\bigg(-\frac{h(x,p-u(x))}{T(x)}\bigg). \tag{3.30}$$

The temperature $T(x)$, which is a Lagrange multiplier coming from the minimization procedure, is determined from the given energy density. The choice of $\widetilde{f}_G$ as a thermal equilibrium function has its physical justification in the fact that it is the unique minimizer of the relative entropy S with the prescribed moments.

Analogously, a quantum system, which is characterized by its energy operator $W^{-1}(h)$ (recall that W^{-1} is the Weyl quantization), attains its minimum of the relative (von Neumann) entropy in the mixed state with Wigner function $f_Q = \exp(-h/T_0)$.

This state represents the unconstrained quantum thermal equilibrium. The expansion of f_Q in terms of the scaled Planck constant ε^2 was first given in [31],

$$f_Q(x,p) = \exp\left(-\frac{h(x,p)}{T_0}\right)(1+\varepsilon^2 f_2(x,p)) + O(\varepsilon^4)$$

with an appropriate function f_2. As a definition of the quantum equilibrium with moment constraints, Gardner employed this expansion of f_Q and modified it as follows:

$$\widetilde{f}_Q(x,p) = n(x)\,\exp\left(-\frac{h(x,p-u(x))}{T(x)}\right)\big(1+\varepsilon^2 f_2(x,p-u(x))\big) + O(\varepsilon^4). \tag{3.31}$$

These modifications mimic the passage from the Gibbs state to (3.30) in the classical situation. The use of $\widetilde{f}_Q$ as an equilibrium function results in simple formulas for the moment equations. However, the Wigner function (3.31) is an ad hoc ansatz. Moreover, in contrast to the classical case, $\widetilde{f}_Q$ is *not* the constrained minimizer for the relative von Neumann entropy.

The equilibrium state M_f used here is a genuine minimizer of the relative quantum entropy with respect to the given moments. In the spirit of the classical situation, these equilibria seem to be more natural. The price we have to pay is the appearance of various additional terms in the expansion of M_f.

If the temperature is assumed to be constant and if only the particle density is prescribed, both approaches to defining a thermal equilibrium coincide. In order to see this, we write Gardner's momentum-shifted quantum equilibrium more explicitly than in (3.31):

$$\widetilde{f}_G(x,p,t) = e^{-V/T-|p|^2/2T}\left\{1+\frac{\varepsilon^2}{8T^2}\left(-\Delta V+\frac{1}{3T}|\nabla V|^2+\frac{1}{3T}p_ip_j\partial_{x_ix_j}V\right)\right\}+O(\varepsilon^4).$$

The equilibrium function obtained from the entropy minimization with given particle density equals (see [21, Remark 3.3])

$$\begin{aligned}\widetilde{f}(x,p,t) &= \exp\left(A(x,t)-\frac{|p|^2}{2T}\right)\\ &= e^{A-|p|^2/2T}\left\{1+\frac{\varepsilon^2}{8T}\left(\Delta A+\frac{1}{3}|\nabla A|^2-\frac{1}{3T}p_ip_j\partial_{x_ix_j}A\right)\right\}+O(\varepsilon^4).\end{aligned}$$

Both approximations are essentially derived in the same way. Using $n=\int \widetilde{f}_Q dp = (2\pi T)^{d/2}e^{-V/T}+O(\varepsilon^2)$ and assuming constant (or "slowly varying") temperature, Gardner has substituted $\nabla V = -T\nabla\log n + O(\varepsilon^2)$ into the formula for $\widetilde{f}_Q$ in order to avoid the second-order derivatives of the potential. This substitution in fact yields the approximation $\widetilde{f}$ since, by (3.26), $\nabla A = \nabla\log n + O(\varepsilon^2)$, and thus, both expansions $\widetilde{f}_Q$ and $\widetilde{f}$ coincide.

4. Simplified QHD models. The full QHD model is given by (3.7)–(3.9) with the constitutive relations (3.27)–(3.29). In this section we will discuss some simplified versions. We recall the QHD equations

$$\partial_t n + \operatorname{div}(nu) = 0, \tag{4.1}$$

$$\partial_t(nu) + \operatorname{div}(nu\otimes u) + \operatorname{div} P - n\nabla V = 0, \tag{4.2}$$

$$\partial_t e + \operatorname{div}\big((P+eI)u\big) + \operatorname{div} S - nu\cdot\nabla V = 0, \tag{4.3}$$

where e is the energy density given by (3.25), and P, S, and U are given by (3.27)–(3.29) (without the $O(\varepsilon^4)$ terms).

First, we shall assume that the temperature is slowly varying in the sense of $\nabla\log T = O(\varepsilon^2)$. Then the expressions $\varepsilon^2\nabla\log T$ in (3.27)–(3.29) are of order $O(\varepsilon^4)$ and can therefore be neglected in our approximation:

$$P = nTI - \frac{\varepsilon^2}{12}n\bigg((\nabla\otimes\nabla)\log n - \frac{R^\top R}{T}\bigg), \tag{4.4}$$

$$S = -\frac{\varepsilon^2}{12}n\bigg\{\bigg(\frac{d}{2}+1\bigg)R\nabla\log n + \bigg(\frac{d}{2}+2\bigg)\operatorname{div}R + \frac{3}{2}\Delta u\bigg\}$$

$$+ \frac{\varepsilon^2}{12}\bigg(\frac{d}{2}+1\bigg)n\bigg\{R\nabla\log n + \operatorname{div}R\bigg\}, \tag{4.5}$$

$$e = \frac{d}{2}nT + \frac{1}{2}n|u|^2 - \frac{\varepsilon^2}{24}n\bigg(\Delta\log n - \frac{1}{T}\operatorname{tr}(R^\top R)\bigg). \tag{4.6}$$

The stress tensor P consists of the classical pressure nT on the diagonal, the "quantum pressure" $(\varepsilon^2/12)n(\nabla\otimes\nabla)\log n$, and the vorticity correction $(\varepsilon^2/12)nR^\top R/T$. The term S provides additional quantum corrections not present in [12]. The energy density consists of the thermal energy, kinetic energy, and quantum energy. Again, due to the vorticity R, the energy takes a different form than the expressions in [12, 16].

Further simplifications can be obtained if the vorticity is "small," i.e., $R = O(\varepsilon^2)$. In one space dimension this term always vanishes. If $R = O(\varepsilon^2)$, then $\varepsilon^2 R$ is of order $O(\varepsilon^4)$ and can be neglected. We obtain the QHD equations

$$\partial_t n + \operatorname{div}(nu) = 0, \tag{4.7}$$

$$\partial_t(nu) + \operatorname{div}(nu\otimes u) + \nabla(nT) - \frac{\varepsilon^2}{12}\operatorname{div}\big(n(\nabla\otimes\nabla)\log n\big) - n\nabla V = 0, \tag{4.8}$$

$$\partial_t e + \operatorname{div}\big((P+eI)u\big) - \frac{\varepsilon^2}{8}\operatorname{div}(n\Delta u) - nu\cdot\nabla V = 0, \tag{4.9}$$

with the stress tensor and energy density, respectively,

$$P = nTI - \frac{\varepsilon^2}{12}n(\nabla\otimes\nabla)\log n, \qquad e = \frac{d}{2}nT + \frac{1}{2}n|u|^2 - \frac{\varepsilon^2}{24}n\Delta\log n.$$

This system of equations corresponds to Gardner's QHD model (without relaxation-time terms) except for the dispersive velocity term $(\varepsilon^2/8)\operatorname{div}(n\Delta u)$. We already mentioned in the introduction that this term has been derived also by Gardner and Ringhofer [16] by employing a Chapman–Enskog expansion of the Wigner–Boltzmann equation. They do not obtain vorticity terms, since they assume that the quantum equilibrium distribution is an even function of the momentum p. Roughly speaking, this gives (in our context) the quantum exponential $\exp(A - |p|^2/2T)$ instead of $\exp(A - |p-w|^2/2T)$. The Lagrange multiplier w, however, is responsible for the presence of the vorticity term R.

Interestingly, most quantum terms cancel out in the energy equation. In fact, by substituting the above expression for the energy density in (4.9), a computation yields

$$\partial_t(nT) + \operatorname{div}(nTu) + \frac{2}{d}nT\operatorname{div}u - \frac{\varepsilon^2}{6d}\operatorname{div}(n\Delta u) = 0.$$

5. Conserved quantities. In this subsection we show that the mass and energy are conserved for the system (3.7)–(3.9) and (1.2) with the relations (3.27)–(3.29), neglecting the $O(\varepsilon^4)$ terms. The momentum is not conserved due to the electric force given by $\int n\nabla V dx$.

LEMMA 5.1. *The mass $N(t) = \int n dx$ and the energy*

$$E(t) = \int_{\mathbb{R}^d} \left(e + \frac{\lambda^2}{2}|\nabla V|^2 \right) dx,$$

where e is defined in (3.25) (*without the $O(\varepsilon^4)$ term*), *are conserved; i.e., $dN/dt(t) = 0$ and $dE(t)/dt = 0$ for all $t > 0$. Furthermore, the energy can be written as*

$$E(t) = \int_{\mathbb{R}^d} \left(\frac{d}{2} nT + \frac{1}{2} n|u|^2 + \frac{\lambda^2}{2}|\nabla V|^2 + \frac{\varepsilon^2}{6}|\nabla\sqrt{n}|^2 + \frac{\varepsilon^2 d}{48} n|\nabla \log T|^2 + \frac{\varepsilon^2}{24T} n \operatorname{tr}(R^\top R) \right) dx \geq 0. \tag{5.1}$$

Proof. The conservation of N is clear. In order to prove that E is conserved, we differentiate E and employ (4.3) and (1.2),

$$\frac{dE}{dt} = \int_{\mathbb{R}^d} (\partial_t e + \lambda^2 \nabla V \cdot \nabla \partial_t V) dx = \int_{\mathbb{R}^d} (nu \cdot \nabla V - \lambda^2 V \partial_t \Delta V) dx$$
$$= \int_{\mathbb{R}^d} (-\operatorname{div}(nu) V - V \partial_t n) dx = 0,$$

taking into account (4.1). Next we prove formula (5.1). The integral of the energy density e can be written as

$$E = \int_{\mathbb{R}^d} \left(\frac{d}{2} nT + \frac{1}{2} n|u|^2 + \frac{\varepsilon^2 d}{48} n|\nabla \log T|^2 + \frac{\varepsilon^2}{24T} n \operatorname{tr}(R^\top R) \right) dx$$
$$+ \frac{\varepsilon^2}{24} \int_{\mathbb{R}^d} \big(- n\Delta \log n + n\Delta \log T + n \nabla \log T \cdot \nabla \log n \big) dx.$$

The last integral equals, after an integration by parts,

$$\frac{\varepsilon^2}{24} \int_{\mathbb{R}^d} \big(4|\nabla\sqrt{n}|^2 - \nabla n \cdot \nabla \log T + n \nabla \log T \cdot \nabla \log n \big) dx = \frac{\varepsilon^2}{6} \int_{\mathbb{R}^d} |\nabla\sqrt{n}|^2 dx,$$

which shows (5.1). ∎

The energy (5.1) consists of, in this order, the thermal energy, the kinetic energy, the electrostatic energy, and the energy of the Bohm potential. The remaining two terms represent additional field quantum energies associated with spatial variations of the temperature and the vorticity. These last two energy terms are new; i.e., they do not appear in the QHD equations of [12].

In the case of the QHD equations with slowly varying temperature, i.e., (4.1)–(4.3) and (1.2) with the definitions (4.4)–(4.5), the energy is given by (5.1) except the term involving $|\nabla \log T|^2$. If, additionally, the vorticity is "small," i.e., in the case of the model (4.7)–(4.9) and (1.2), which is used in the numerical simulations of section 6, the energy is equal to (5.1) except for the last two terms.

Unfortunately, we are not able to prove the conservation of the $O(\varepsilon^4)$ approximation of the quantum entropy and the positivity of the particle density (as for the model in [9]) since we obtain $O(\varepsilon^4)$ correction terms which do not vanish.

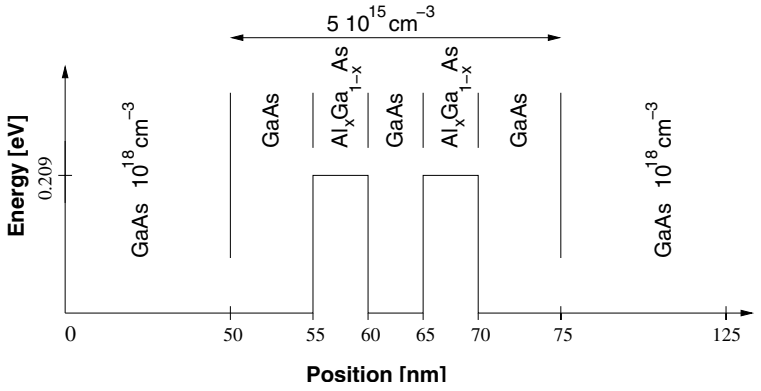


FIG. 6.1. *Geometry of the resonant tunneling diode and external potential modeling the double barriers. The Al mole fraction is* $x = 0.3$.

6. Numerical results. In this section we present the results from our numerical simulations of a simple one-dimensional Gallium arsenide (GaAs) resonant tunneling diode, using the new QHD system. The aim is also to compare the new equations with Gardner's QHD model; in particular, the influence of the dispersive velocity term will be explored.

The geometry of the tunneling diode is chosen essentially as in [12] (see Figure 6.1). The diode consists of highly doped 50 nm GaAs regions near the contacts and a lightly doped middle region (the channel) of 25 nm length. The channel contains a quantum well of 5 nm length, sandwiched between two 5 nm $Al_xGa_{1-x}As$ barriers with Al mole fraction $x = 0.3$. The double barrier heterostructure is placed between two 5 nm GaAs spacer layers. The total length of the device is thus 125 nm. The double barrier height is $B = 0.209\,\mathrm{eV}$. It is incorporated into the QHD equations by replacing V by $V + B$.

For our simulations, we use the one-dimensional stationary QHD equations for small temperature variations $\nabla \log T = O(\varepsilon^2)$ coupled to the Poisson equation for the electric potential. Including the physical parameters, these equations read as follows:

$$(nu)_x = 0, \tag{6.1}$$

$$m(nu^2)_x + k_B(nT)_x - \frac{\hbar^2}{12m}(n(\log n)_{xx})_x - qnV_x = 0, \tag{6.2}$$

$$\frac{5}{2}k_B(nTu)_x + \frac{1}{2}m(nu^3)_x - \frac{\hbar^2}{8m}(nu(\log n)_{xx} + nu_{xx})_x - qnuV_x = k_B\sigma(nT_x)_x, \tag{6.3}$$

$$\varepsilon_s V_{xx} = q(n - C). \tag{6.4}$$

The physical constants in the above equations are the effective mass m, the Boltzmann constant k_B, the reduced Planck constant $\hbar$, the elementary charge q, and the semiconductor permittivity ε_s. The values of these constants are given in Table 6.1. The parameter σ is defined by

$$\sigma = \kappa\tau_0\frac{k_BT_0}{m},$$

with the thermal conductivity κ, the relaxation time τ_0, and the lattice temperature T_0.

TABLE 6.1
Physical parameters for GaAs.

Parameter	Physical meaning	Value
q	Elementary charge	$1.602 \cdot 10^{-19}$ As
m	Eeffective electron mass	$0.067 \cdot m_0$ with $m_0 = 9.11 \cdot 10^{-31}$ kg
k_B	Boltzmann constant	$1.3807 \cdot 10^{-23}$ kg m^2/s^{2}K
$\hbar$	Reduced Planck constant	$1.0546 \cdot 10^{-34}$ kg m^2/s
ε_s	Semiconductor permittivity	$12.9 \cdot 8.8542 \cdot 10^{-12}$ A^{2}s^4/kg m^3
τ_0	Momentum relaxation time	$0.9 \cdot 10^{-12}$ s
T_0	Lattice temperature	77 K

We have allowed the heat flux $k_B\sigma(nT_x)_x$ since this term has also been used by Gardner [12] in his model, with which we wish to compare our numerical results. In fact, we need this term for numerical stability as it is needed in Gardner's QHD equations. We expect that the heat conductivity can be obtained by a Chapman–Enskog expansion of the Wigner–Boltzmann equation, but additional diffusion terms might appear.

Using a standard scaling (see, e.g., [19]), we derive the scaled QHD equations (1.2)–(1.5) of the introduction, where the nondimensional parameters are given by

$$\varepsilon^2 = \frac{\hbar^2}{mk_BT_0L^2}, \qquad \lambda^2 = \frac{\varepsilon_s k_B T_0}{q^2 C^* L^2}.$$

Here, L is the device length and C^* the maximal doping concentration. For the values we used in the numerical simulations below (see Table 6.1), we obtain $\varepsilon^2 \approx 0.011$, which justifies our expansion in ε^2.

We compare the numerical results with Gardner's QHD equations, which do not contain the dispersive expression (1.7) in the velocity but additional relaxation-time terms of Baccarani–Wordeman type [3]:

$$(nu)_x = 0, \tag{6.5}$$

$$m(nu^2)_x + k_B(nT)_x - \frac{\hbar^2}{12m}(n(\log n)_{xx})_x - qnV_x = -\frac{mnu}{\tau_p}, \tag{6.6}$$

$$\begin{aligned}\frac{5}{2}k_B(nTu)_x + \frac{1}{2}m(nu^3)_x - \frac{\hbar^2}{8m}(nu(\log n)_{xx})_x - qnuV_x \\ = k_B\sigma(nT_x)_x - \frac{1}{\tau_w}\left(e - \frac{3}{2}nT_0\right),\end{aligned} \tag{6.7}$$

together with the Poisson equation (6.4). Here, the momentum and energy relaxation times are given by, respectively,

$$\tau_p = \tau_0\frac{T_0}{T}, \qquad \tau_w = \frac{\tau_p}{2}\left(1 + \frac{3T}{mv_s^2}\right),$$

where τ_0 is given in Table 6.1 and $v_s = 2 \cdot 10^7$ cm/s is the saturation velocity. The inclusion of these terms (at least if $\tau_p = \tau_w/2$) can be justified by employing a Caldeira–Leggett scattering operator as exposed in section 3.1. We observed that the relaxation-time terms in Gardner's QHD model are necessary for numerical stability; on the other hand, they lead to severe numerical difficulties when included in the new QHD equations.

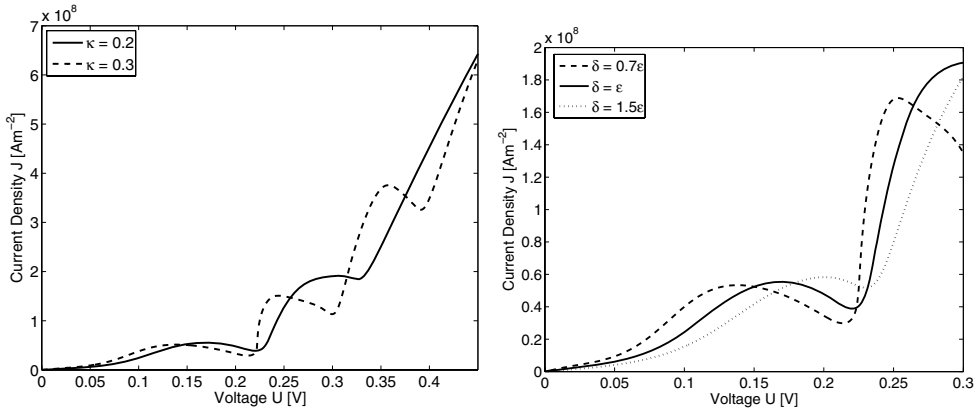


FIG. 6.2. *Left: Current-voltage characteristic for the new QHD system with thermal conductivities* $\kappa = 0.2$ (*solid line*) *and* $\kappa = 0.3$ (*dashed line*). *Right: Influence of the dispersive velocity term* $(\delta^2/8)(nu_{xx})_x$ *on the current-voltage curve for thermal conductivity* $\kappa = 0.2$.

The above QHD equations have to be solved in the interval $(0, 1)$ with the following boundary conditions taken from [12]:

$$n(0) = C(0), \quad n(1) = C(1), \quad n_x(0) = n_x(1) = 0,$$
$$u_x(0) = u_x(1) = 0, \quad T(0) = T(1) = T_0, \quad V(0) = 0, \quad V(1) = U_0,$$

where U_0 is the applied voltage.

First, we discretize the new QHD equations (6.1)–(6.4) using central finite differences on a uniform mesh with $N = 500$ points. This corresponds to a mesh size of $\triangle x = 1/500 = 0.002$. The resulting discrete nonlinear system is solved by a damped Newton method with damping parameter found by a line search method (see Algorithm A6.3.1 in [10]). We employ the following continuation method for the applied voltage: first the system of equations is solved for applied voltage $U_0 = 0\,\mathrm{V}$; then, given the solution corresponding to the voltage U_0, it is taken as an initial guess for the solution of the system with applied voltage $U_0 + \triangle U$. The voltage step is chosen as $\triangle U = 1\,\mathrm{mV}$.

The current-voltage characteristics using the thermal conductivities $\kappa = 0.2$ and $\kappa = 0.3$ are presented in Figure 6.2. There are apparently two regions of negative differential resistance (NDR) if $\kappa = 0.2$, and three NDR regions if $\kappa = 0.3$. It is well known for Gardner's QHD model that the behavior of the solutions is quite sensitive to changes of the value of the thermal conductivity. We observe a similar sensitive dependence: the peak-to-valley ratio, i.e., the ratio of local maximal to local minimal current density, is larger for larger thermal conductivities.

The electron density shows a charge enhancement in the quantum well, which is more pronounced for smaller κ (see Figure 6.3(left)). At the center of the right barrier, the electron density dramatically decreases. After the first valley in the current-voltage characteristics, the density develops a "wiggle." This phenomenon is not a numerical effect, since it has been observed in various numerical simulations [24, 27]. For larger values of the thermal conductivity, the minimum of the particle density increases, which stabilizes the numerical scheme.

Next, we study the influence of the dispersive velocity term $(\varepsilon^2/8)(nu_{xx})_x$. For this, we replace the factor $\varepsilon^2/8$ by $\delta^2/8$ and choose various values for δ. Clearly, only

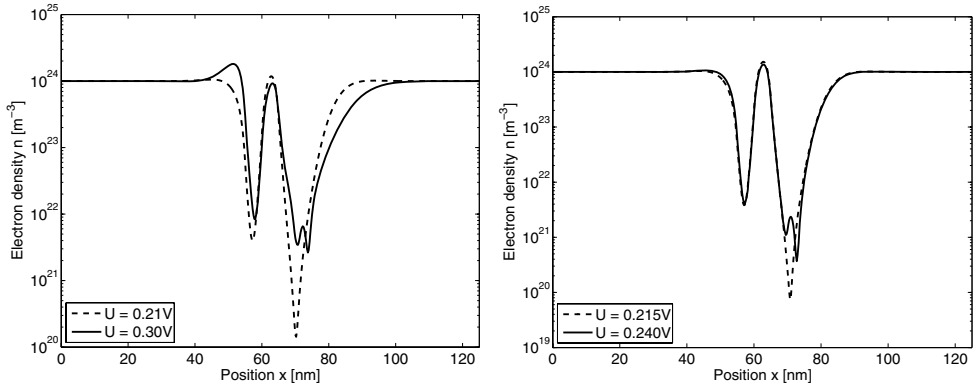


FIG. 6.3. *Electron density before* (*dashed line*) *and after* (*solid line*) *the first valley for thermal conductivities* $\kappa = 0.2$ (*left*) *and* $\kappa = 0.3$ (*right*).

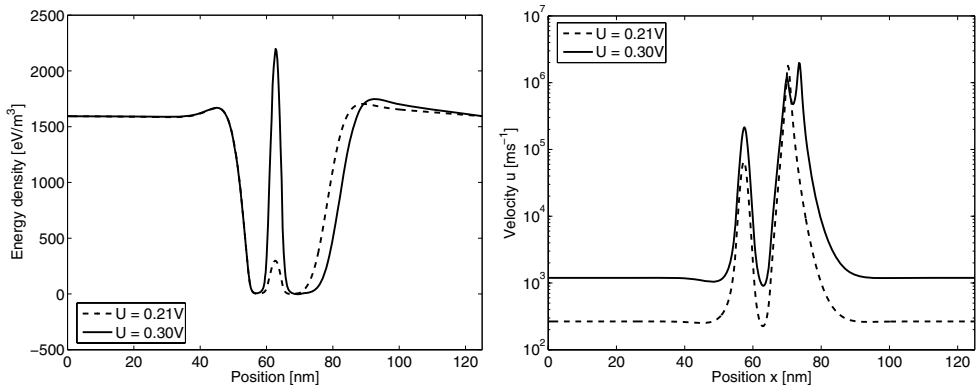


FIG. 6.4. *Thermal energy density* (*left*) *and velocity* (*right*) *before* (*dashed line*) *and after* (*solid line*) *the first valley computed from the new QHD model* (*thermal conductivity* $\kappa = 0.2$).

$\delta = \varepsilon$ corresponds to the physical situation. The dispersive velocity term indeed regularizes the equations in the sense that the current-voltage curves become "smoother" (see Figure 6.3(right)). A similar "smoothing" has been observed in [23, 24] for the viscous QHD equations, but there the smoothing originates from a diffusive and not from a dispersive term. For smaller values of δ, the peak-to-valley ratio of the first NDR region becomes larger. For $\delta = 0$, we arrive at Gardner's QHD equations without relaxation terms. We already mentioned that a central finite-difference discretization fails for this model; therefore, the limit $\delta \to 0$ cannot be done numerically.

In Figure 6.4 the thermal energy density $\frac{3}{2} n k_B T$ and the velocity $u = J/(qn)$ are reported. The velocity profile is very similar to that computed from Gardner's model (see Figure 6.5, $N = 500$). The velocity is high in the barriers and rather small in the well; i.e., the electrons spend more time in the quantum well than in the barriers. On the other hand, the temperature of the new QHD model differs from that obtained by Gardner's QHD model, particularly in the region between the barriers. The heating in the well in our model can probably be explained by the central scheme that we have used. Gardner's upwind scheme involves some numerical diffusion that seems to bring down the thermal energy in the quantum well. We notice that $\nabla \log T$ is *not*

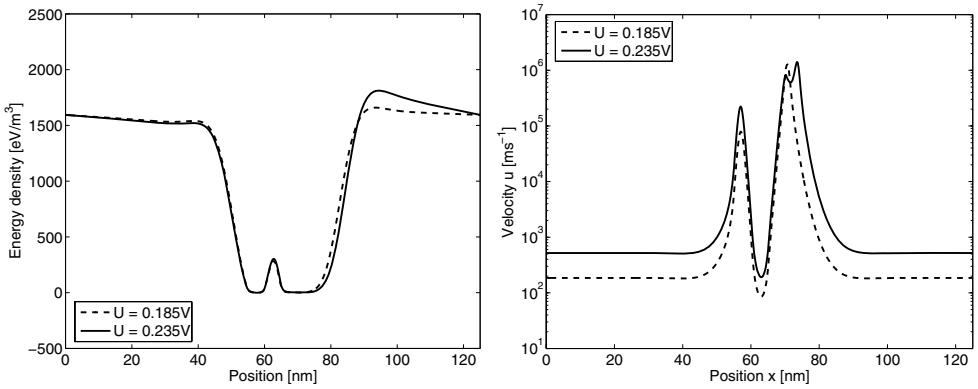


FIG. 6.5. *Thermal energy density (left) and velocity (right) before (dashed line) and after (solid line) the first valley computed from Gardner's QHD model (thermal conductivity $\kappa = 0.2$).*

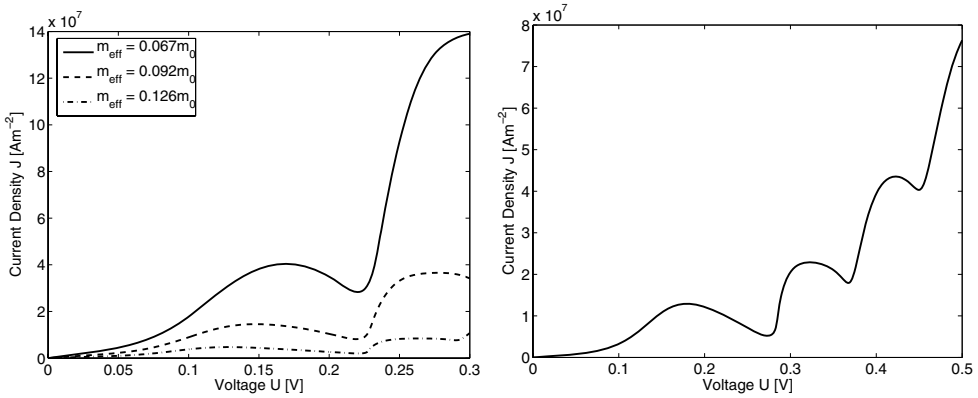


FIG. 6.6. *Left: Influence of the effective mass m_{eff} on the current-voltage characteristic. Right: Current-voltage characteristic for a barrier height of $B = 0.3\,eV$. In both pictures, $\kappa = 0.2$.*

of order $O(\varepsilon^2)$, as assumed in the derivation of the model, except in the high doped contact regions.

The influence of the effective mass on the current-voltage curves is shown in Figure 6.6(left). Corresponding to the effective masses $m = 0.067m_0$, $m = 0.092m_0$, $m = 0.126m_0$, the peak-to-valley ratios are 1.44, 1.79, 2.37, respectively. Here, m_0 denotes the electron mass at rest. Similarly to the quantum drift-diffusion model, the peak-to-valley ratio increases with the effective mass [22]. Strictly speaking, the effective mass is not constant in the whole device, but it is material-dependent. The use of a nonconstant effective mass would be more physical, but the modeling and the numerical approximation is—even in the much simpler quantum drift-diffusion model—a lot more involved [29, 30].

In Figure 6.6(right) the current-voltage curve for the barrier height $B = 0.3\,\mathrm{eV}$ is shown. As expected, the peak-to-valley ratio is larger if the barrier is higher (corresponding to a higher Al mole fraction); the values for the first NDR region are 1.44 for $B = 0.209\,\mathrm{eV}$ and 2.48 for $B = 0.3\,\mathrm{eV}$. The current densities are much smaller than in Figure 6.2, where the lower potential barrier $B = 0.209\,\mathrm{eV}$ has been used. Interestingly, there are at least three NDR regions, whereas there are only two regions

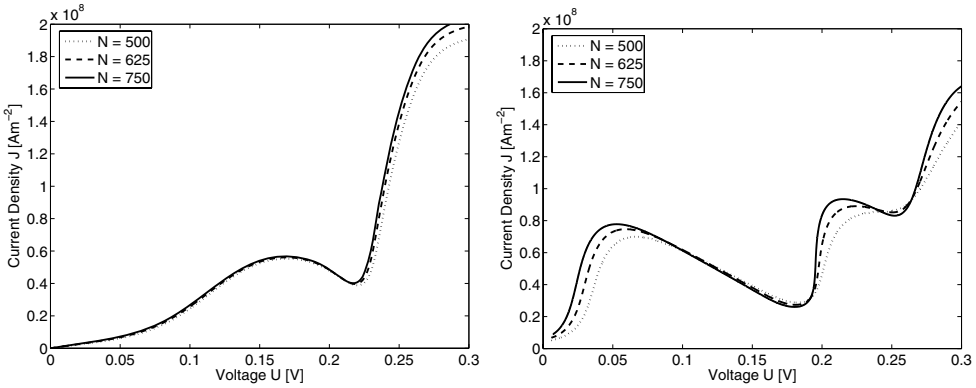


FIG. 6.7. *Influence of the number of discretization points on the current-voltage characteristics for the new QHD equations* (*left, using central finite differences*) *and for Gardner's QHD model* (*right, using an upwind scheme*). *In both pictures,* $\kappa = 0.2$.

for the barrier height $B = 0.209\,\mathrm{eV}$.

In Figure 6.7, the current-voltage curves for the new QHD equations and for Gardner's model are compared. Gardner's model is discretized using a second-order upwind method as in [12]. The right figure with $N = 500$ points corresponds to Figure 2 of the cited paper. Notice that close to thermal equilibrium there are well-known difficulties in computing the solution, which is not the case for our new model. Due to the numerical viscosity introduced by the upwind method, it is clear that the solution of Gardner's model depends on the mesh size. The solution to the new QHD equations is less mesh-dependent. In particular, the numerical results before the first valley are almost the same for $N \geq 500$ grid points. More importantly, the slope of the curve in Gardner's model becomes steeper in the region after the valley when the mesh size $\triangle x$ is decreased. On the other hand, the current-voltage curve of the new QHD model does not seem to develop such singular slopes. Moreover, it is possible to solve the discrete system for grid points $N > 750$ (not shown).

REFERENCES

[1] M. Ancona and G. Iafrate, *Quantum correction to the equation of state of an electron gas in a semiconductor*, Phys. Rev. B, 39 (1989), pp. 9536–9540.

[2] M. Ancona and H. Tiersten, *Macroscopic physics of the silicon inversion layer*, Phys. Rev. B, 35 (1987), pp. 7959–7965.

[3] G. Baccarani and M. Wordeman, *An investigation of steady-state velocity overshoot effects in Si and GaAs devices*, Solid-State Electronics, 28 (1985), pp. 407–416.

[4] P. Bhatnagar, E. Gross, and M. Krook, *A model for collision processes in gases*, I. *Small amplitude processes in charged and neutral one-component systems*, Phys. Rev., 94 (1954), pp. 511–525.

[5] A. Caldeira and A. Leggett, *Path integral approach to quantum Brownian motion*, Phys. A, 121A (1983), pp. 587–616.

[6] Z. Chen, *A finite element method for the quantum hydrodynamic model for semiconductor devices*, Comput. Math. Appl., 31 (1996), pp. 17–26.

[7] P. Degond, F. Méhats, and C. Ringhofer, *Quantum hydrodynamic models derived from the entropy principle*, Contemp. Math., 371 (2005), pp. 107–131.

[8] P. Degond, F. Méhats, and C. Ringhofer, *Quantum energy-transport and drift-diffusion models*, J. Statist. Phys., 118 (2005), pp. 625–665.

[9] P. Degond and C. Ringhofer, *Quantum moment hydrodynamics and the entropy principle*, J. Statist. Phys., 112 (2003), pp. 587–628.

[10] J. E. Dennis, jr., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, 1996.
[11] D. Ferry and J.-R. Zhou, *Form of the quantum potential for use in hydrodynamic equations for semiconductor device modeling*, Phys. Rev. B, 48 (1993), pp. 7944–7950.
[12] C. L. Gardner, *The quantum hydrodynamic model for semiconductor devices*, SIAM J. Appl. Math., 54 (1994), pp. 409–427.
[13] C. Gardner, *Resonant tunneling in the quantum hydrodynamic model*, VLSI Design, 3 (1995), pp. 201–210.
[14] C. Gardner and C. Ringhofer, *Smooth quantum potential for the hydrodynamic model*, Phys. Rev. E, 53 (1996), pp. 157–167.
[15] C. L. Gardner and C. Ringhofer, *Approximation of thermal equilibrium for quantum gases with discontinuous potentials and application to semiconductor devices*, SIAM J. Appl. Math., 58 (1998), pp. 780–805.
[16] C. Gardner and C. Ringhofer, *The Chapman-Enskog expansion and the quantum hydrodynamic model for semiconductor devices*, VLSI Design, 10 (2000), pp. 415–435.
[17] I. Gasser, P. Markowich, D. Schmidt, and A. Unterreiter, *Macroscopic theory of charged quantum fluids*, in Mathematical Problems in Semiconductor Physics, P. Marcati, P. A. Markowich, and R. Natalini, eds., Res. Notes Math. Ser. 340, Pitman, Boston, 1995, pp. 42–75.
[18] H. Grubin and J. Kreskovsky, *Quantum moment balance equations and resonant tunneling structures*, Solid-State Electronics, 32 (1989), pp. 1071–1075.
[19] M. Gualdani and A. Jüngel, *Analysis of the viscous quantum hydrodynamic equations for semiconductors*, European J. Appl. Math., 15 (2004), pp. 577–595.
[20] A. Jüngel, *Quasi-hydrodynamic Semiconductor Equations*, Birkhäuser, Basel, 2001.
[21] A. Jüngel and D. Matthes, *A derivation of the isothermal quantum hydrodynamic equations using entropy minimization*, Z. Angew. Math. Mech., 85 (2005), pp. 806–814.
[22] A. Jüngel and J.-P. Milišić, *Macroscopic quantum models with and without collisions*, in Proceedings of the Sixth International Workshop on Mathematical Aspects of Fluid and Plasma Dynamics, Kyoto, Japan, Transp. Theory Stat. Phys., 2005, to appear.
[23] A. Jüngel and J.-P. Milišić, *Numerical and physical viscosity in the quantum hydrodynamic equations for semiconductors*, preprint, Universität Mainz, Germany, 2006.
[24] A. Jüngel and S. Tang, *Numerical approximation of the viscous quantum hydrodynamic model for semiconductors*, Appl. Numer. Math., 56 (2006), pp. 899–915.
[25] C. Levermore, *Moment closure hierarchies for kinetic theories*, J. Statist. Phys., 83 (1996), pp. 1021–1065.
[26] E. Madelung, *Quantentheorie in hydrodynamischer form*, Z. Physik, 40 (1927), pp. 322–326.
[27] P. Pietra and C. Pohl, *Weak limits of the quantum hydrodynamic model*, VLSI Design, 9 (1999), pp. 427–434.
[28] C. Ringhofer, C. Gardner, and D. Vasileska, *Effective potentials and quantum fluid models: A thermodynamic approach*, Intern. J. High Speed Electronics Sys., 13 (2003), pp. 771–801.
[29] A. Wettstein, *Quantum Effects in MOS Devices*, Series in Microelectronics 94, Hartung-Gorre, Konstanz, Germany, 2000.
[30] A. Wettstein, A. Schenk, and W. Fichtner, *Quantum device-simulation with the density-gradient model on unstructured grids*, IEEE Trans. Electron. Dev., 48 (2001), pp. 279–284.
[31] E. Wigner, *On the quantum correction for thermodynamic equilibrium*, Phys. Rev., 40 (1932), pp. 749–759.

SIAM J. APPL. MATH.
Vol. 67, No. 1, pp. 69–98

ELECTROSEISMIC PROSPECTING IN LAYERED MEDIA*

BENJAMIN S. WHITE† AND MINYAO ZHOU†

Abstract. Electroseismic (ES) prospecting is an experimental method that seeks to use the conversion of electromagnetic (EM) waves to seismic waves in the earth to explore for oil and gas. The wave conversion occurs through the phenomenon of electrokinetics, for which a complete set of partial differential equations was derived by S. Pride. In this paper, we show how Pride's equations in plane layered media can be written in a convenient mathematical form suggested by B. Ursin, who used this form to give a unified treatment of EM waves, acoustic waves, and the waves of isotropic elasticity in plane layered media. We use Ursin's formalism, which we develop and simplify for the case of a stack of homogeneous layers, to derive explicit formulas that can be made the basis of an efficient computer code. Numerical results are presented for spatially extended electrode sources that have been used in field tests of ES prospecting. More generally, the methods developed are applicable to any system that can be put into Ursin's form. In particular, the code that was written for ES can be modified to compute purely seismic waves, purely EM waves, or the waves of Biot theory, since all these phenomena are included in the equations of electrokinetics when parameters are specialized appropriately.

Key words. electrokinetics, layered media, seismic, electromagnetic, poroelastic

AMS subject classifications. 74F15, 74F10, 86A15, 86A25, 74J05

DOI. 10.1137/050633603

1. Introduction.

In prospecting for oil and gas, seismic methods are the main tool for imaging of the earth's subsurface [4] because of the high spatial resolution that is possible. However, skillful interpretation of the seismic images is necessary to distinguish hydrocarbon- from nonhydrocarbon-saturated rocks, because there are often only subtle differences in the rock properties (e.g., densities and compressibilities) that affect the seismic waves. In contrast, the electrical resistivity of hydrocarbon-saturated rocks is one to three orders of magnitude greater than that of the surrounding medium, making resistivity an excellent direct indicator of hydrocarbons. If the earth's electrical properties could be imaged with the same spatial resolution as is routine in seismic imaging, then many hydrocarbon reservoirs could be easily detected and delineated.

Electromagnetic (EM) exploration methods [14, 25] can be used to map the electrical properties of the subsurface. However, the earth is a conducting medium, so high-frequency EM waves are attenuated rapidly as they propagate in the earth. Thus EM surveys, except for shallow depths, typically rely on lower frequency/longer wavelength waves. Then the spatial resolution of the subsurface images are limited by the long wavelengths of the probes.

Electroseismic (ES) prospecting is an experimental method that seeks to use the conversion of EM waves to seismic waves in the earth to search for hydrocarbons. This wave conversion occurs through the phenomenon of electrokinetics [15], i.e., in a porous medium such as the earth, an EM wave will excite a seismic wave of the same frequency, and vice versa, through movement of ions in the pore fluids. Since EM waves have a much faster propagation speed than seismic, they have much

*Received by the editors June 14, 2005; accepted for publication (in revised form) July 18, 2006; published electronically November 14, 2006.
http://www.siam.org/journals/siap/67-1/63360.html
†ExxonMobil Corporate Strategic Research, Route 22 East, Annandale, NJ 08801 (benjamin.s.white@exxonmobil.com, minyaozhou@yahoo.com).

longer wavelengths than seismic waves of the same frequency. Thus using seismic frequencies some of the energy of a long wavelength EM wave probe will be converted by inhomogeneities in the earth to a shorter wavelength seismic wave, which might then be recorded at the earth's surface. The waves recorded will have information related to the earth's electrical properties, as is shown below.

In this paper we derive the mathematical basis for an efficient computer code for solving Pride's equations [15] of electrokinetics in plane layered media, i.e., earth models in which the material parameters are functions of the depth coordinate only. The resulting code has been implemented and, together with the three-dimensional asymptotic theory of [24], has been used for the planning and interpretation of field tests of ES prospecting. Although the tests are described elsewhere [21, 20, 18, 10, 9, 8], we will show here how the calculations were done for the type of experimental source/receiver configurations used in the field.

Our method is based on a formalism introduced by Ursin [23], who showed how Maxwell's equations for electromagnetism, the equations of acoustics, and the equations of isotropic elasticity all have a similar mathematical structure in layered media when each of these systems is written in an appropriate way. In this paper, we add the equations of electrokinetics to Ursin's list. We develop and simplify Ursin's formalism for the case of a stack of homogeneous layers, that is, when the material parameters are piecewise constant functions of depth. In this case many quantities can be computed with explicit algebraic formulas which can then be made the basis of a fast computer code.

Early field tests of electrokinetic phenomena in the earth were reported in the 1936 paper of Thompson [22]. More recent tests are reported in [19, 3, 12, 13, 5]. All these studies are more properly called "seismoelectric," i.e., seismic waves were used as a source and EM waves were recorded. The tests [21, 20, 18, 10, 9, 8] are "electroseismic," with EM sources and seismic geophones as receivers.

General properties of ES waves, e.g., source-receiver reciprocity, eigenvectors, and Green's functions in homogeneous media, were derived by Pride and Haartsen in [17]. These authors also wrote a computer code that computes electroseismics for point sources in layered media, using the "global matrix method," that is, inverting a large but banded matrix to solve simultaneously for all quantities in all layers [7]. In contrast, our method makes use of Ursin's mathematical structure to solve for reflection and/or transmission matrices at each layer boundary. Our method is closer to that adopted by Garambois and Dietrich [6], who also used reflection and transmission matrices, but in the form proposed by Kennett and Kerry [11]. We feel that although Ursin's formulation is less well known, using it one can make better use of the inherent mathematical structure of the wave problems, including, e.g., explicit formulas for the jumps across interfaces.

The asymptotic theory of [24] is applicable when seismic ray theory is, i.e., when the seismic wavelength is much smaller than any of the geometric scales in the earth model. With this theory, ES calculations can be done for media that are not plane-layered, but have nontrivial three-dimensional geometry. The asymptotic theory can then be used to compute the seismic waves that are converted from an EM wave incident on an interface, that is, a boundary between media whose material parameters differ. Numerical comparisons of that theory with the computer code described here are in [24].

The formulas of the asymptotic theory show that seismic signals are generated by an EM wave incident on an interface, and that a large discontinuity in the electric field is one of the factors that can cause large conversions of EM to seismic energy.

Now EM theory dictates that the normal electric field at an interface suffers a jump discontinuity in proportion to the ratio of electrical resistivities on each side of the interface (ignoring displacement currents, as is usual in EM prospecting [25]). Thus the asymptotic theory suggests that ES prospecting should illuminate the boundaries of hydrocarbon reservoirs—because the high contrast between the resistivity of hydrocarbons and that of the surrounding medium will produce a large discontinuity of the electric field there.

This paper is organized as follows: In section 2 we write Pride's equations in Ursin's form, complete with source terms and boundary conditions. In section 3 we give a self-contained derivation of Ursin's diagonalization method, in the form in which it is subsequently used. In section 4 we derive formulas for propagator matrices, jump matrices, and reflection and transmission matrices for any system that can be put in Ursin form, and in section 5 we couple the results of section 4 with general sources and boundary conditions, giving explicit formulas for Pride's equations. In sections 6 and 7 we look in some detail at seismic and EM sources, respectively, and in section 8 we show numerical results for the type of ES modeling used in our field tests. Conclusions are in section 9 and explicit formulas for eigenvalues and eigenvectors are in the appendices.

2. Pride's equations in layered media. Pride's equations for electrokinetics in a porous medium are [15], at each point $\mathbf{x} = (x_1, x_2, x_3)$ of space,

$$\begin{aligned}
\nabla \times \mathbf{E} &= i\omega\mu\mathbf{H}, \\
\nabla \times \mathbf{H} &= (\sigma - i\epsilon\omega)\mathbf{E} + L(-\nabla p + \omega^2\rho_f\mathbf{u} + \mathbf{f}) + \mathbf{j}, \\
-\omega^2(\rho\mathbf{u} + \rho_f\mathbf{w}) &= \nabla \cdot \tau + \mathbf{F}, \\
-i\omega\mathbf{w} &= L\mathbf{E} + (\kappa/\eta)(-\nabla p + \omega^2\rho_f\mathbf{u} + \mathbf{f}), \\
\tau &= (\lambda\nabla \cdot \mathbf{u} + C\nabla \cdot \mathbf{w})\mathbf{I} + G(\nabla\mathbf{u} + \nabla\mathbf{u}^{\mathbf{T}}), \\
-p &= C\nabla \cdot \mathbf{u} + M\nabla \cdot \mathbf{w}.
\end{aligned} \tag{2.1}$$

In (2.1) a time dependence of $\exp(-i\omega t)$, where ω is frequency, is assumed.

The sources in Pride's equations are the following: $\mathbf{F}$, the imposed force on the solid, $\mathbf{f}$, the imposed force on the pore fluid, and $\mathbf{j}$, the externally applied electrical current.

The following are the quantities to be calculated: $\mathbf{E}$, the electric field, $\mathbf{H}$, the magnetic field, $\mathbf{u}$, the solid displacement, $\mathbf{w}$, the relative fluid displacement, τ, the stress tensor, and p, the pressure in the pore fluid. $\mathbf{I}$ is the 3×3 identity matrix.

The material parameters in Pride's equations are as follows: μ, the magnetic permeability, ϵ, the dielectric constant, λ and G, the Lamé parameters, C and M, the Biot moduli, ρ, the bulk density, ρ_f, the density of the pore fluid, κ, the permeability, η, the pore fluid viscosity, and L, the electrokinetic mobility. It is $L \neq 0$ that couples the EM and mechanical systems. For $L = 0$ the equations reduce to uncoupled systems, with Maxwell's equations governing electromagnetism and Biot's equations governing fluid and solid motion in a porous medium [1, 2, 16].

We will write

$$\bar{\sigma} = \sigma - i\epsilon\omega. \tag{2.2}$$

It is customary in EM prospecting to make the quasi-static approximation [25] that $\sigma \gg \epsilon\omega$ in the subsurface $x_3 > 0$, so that $\bar{\sigma}$ can be replaced by σ there. This approximation is equivalent to ignoring displacement currents in the earth. In the

air, $x_3 < 0$, the conductivity is zero, and the dielectric constant is ϵ_0, so that we have $\bar{\sigma} = -i\epsilon_0\omega$ in the air.

We will also define

$$\begin{aligned} \beta_1 &= [C^2 - M(\lambda + 2G)]^{-1}, \\ \beta_2 &= \left[1 - \frac{\eta}{\kappa}\frac{L^2}{\bar{\sigma}}\right]^{-1}. \end{aligned} \tag{2.3}$$

For material parameters which depend only on the depth coordinate $x_3 = z$ we can take Fourier transforms in the two lateral coordinates x_1, x_2. Let $(k_1, k_2)^T$ be the horizontal wavenumber and let

$$k = \sqrt{k_1^2 + k_2^2}, \quad \gamma = k/\omega \tag{2.4}$$

be the magnitude of the horizontal wavenumber and the horizontal slowness, respectively. Define Fourier transforms

$$\hat{\mathbf{F}}(k_1, k_2, z) \equiv \mathcal{F}[\mathbf{F}] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i(k_1x_1+k_2x_2)}\ \mathbf{F}(x_1, x_2, z)dx_1dx_2, \tag{2.5}$$

with similar expressions for $\hat{\mathbf{f}}, \hat{\mathbf{j}}, \hat{\mathbf{E}}, \hat{\mathbf{H}}, \hat{\mathbf{u}}, \hat{\mathbf{w}}, \hat{\tau}, \hat{p}$. The lateral Fourier transforms are inverted by the usual formulas, e.g.,

$$\mathbf{F}(x_1, x_2, z) \equiv \mathcal{F}^{-1}[\hat{\mathbf{F}}] = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i(k_1x_1+k_2x_2)}\ \hat{\mathbf{F}}(k_1, k_2, z)dk_1dk_2, \tag{2.6}$$

with similar expressions for the other variables.

To write the equations in Fourier transform space, we first transform the sources to obtain $\hat{\mathbf{F}}, \hat{\mathbf{f}}, \hat{\mathbf{j}}$. Then for each $(k_1, k_2)^T$, plane wave sources of the form $e^{i(k_1x_1+k_2x_2)}\hat{\mathbf{F}}$, $e^{i(k_1x_1+k_2x_2)}\hat{\mathbf{f}}$, $e^{i(k_1x_1+k_2x_2)}\hat{\mathbf{j}}$ will produce plane wave responses with spatial dependence of the form $e^{i(k_1x_1+k_2x_2)}$. The equations are greatly simplified if we rotate to a coordinate system $(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)^T$ with the first coordinate oriented in the direction of the horizontal wavenumber $(k_1, k_2)^T$, so that all of these plane waves have a spatial dependence of the form $e^{ik\tilde{x}_1}$. Therefore, let

$$\Omega = \begin{bmatrix} \frac{k_1}{k} & \frac{k_2}{k} & 0 \\ -\frac{k_2}{k} & \frac{k_1}{k} & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.7}$$

and define

$$\begin{aligned} &\tilde{\mathbf{x}} = \Omega\mathbf{x}, \ \ \tilde{\mathbf{E}} = \Omega\hat{\mathbf{E}}, \ \ \tilde{\mathbf{H}} = \Omega\hat{\mathbf{H}}, \\ &\tilde{\mathbf{u}} = \Omega\hat{\mathbf{u}}, \ \ \tilde{\mathbf{w}} = \Omega\hat{\mathbf{w}}, \ \ \tilde{\tau} = \Omega\hat{\tau}\Omega^{\mathbf{T}}, \ \ \tilde{p} = \hat{p}, \\ &\tilde{\mathbf{F}} = \Omega\hat{\mathbf{F}}, \ \ \tilde{\mathbf{f}} = \Omega\hat{\mathbf{f}}, \ \ \tilde{\mathbf{j}} = \Omega\hat{\mathbf{j}}. \end{aligned} \tag{2.8}$$

A straightforward calculation yields the following form for the equations, since the governing equations (2.1) are invariant under rotations:

Let

$$\dot{\tilde{\mathbf{u}}} = -i\omega\tilde{\mathbf{u}}, \ \ \dot{\tilde{\mathbf{w}}} = -i\omega\tilde{\mathbf{w}} \tag{2.9}$$

be the solid and relative fluid velocities, and let

$$\Phi^{(1)} = \left[\dot{\tilde{u}}_3, \tilde{\tau}_{13}, -\dot{\tilde{w}}_3, \tilde{H}_2, \tilde{\tau}_{33}, \dot{\tilde{u}}_1, \tilde{p}, \tilde{E}_1\right]^{\mathbf{T}},$$
$$\Phi^{(2)} = \left[\dot{\tilde{u}}_2, \tilde{E}_2, \tilde{\tau}_{23}, -\tilde{H}_1\right]^{\mathbf{T}}. \tag{2.10}$$

Define $n_1 = 4$ and $n_2 = 2$. Then the $2n_m$-dimensional vectors $\Phi^{(m)}$, $m = 1, 2$, satisfy uncoupled systems of linear ordinary differential equations of the form suggested by Ursin [23],

$$\frac{d}{dz}\Phi^{(m)} = -i\omega\mathbf{M}^{(m)}\Phi^{(m)} + \mathbf{S}^{(m)}, \quad m = 1, 2, \tag{2.11}$$

where $\mathbf{S}^{(m)}$ are $2n_m$-dimensional source vectors and the $2n_m \times 2n_m$ matrices $\mathbf{M}^{(m)}$ are of the block form

$$\mathbf{M}^{(m)} = \begin{bmatrix} \mathbf{0} & \mathbf{M_1}^{(m)} \\ \mathbf{M_2}^{(m)} & \mathbf{0} \end{bmatrix} \tag{2.12}$$

with symmetric $n_m \times n_m$ submatrices

$$\mathbf{M_1}^{(m)} = \left(\mathbf{M_1}^{(m)}\right)^{\mathbf{T}}, \ \mathbf{M_2}^{(m)} = \left(\mathbf{M_2}^{(m)}\right)^{\mathbf{T}}. \tag{2.13}$$

The $(2n_1 = 8)$-dimensional System 1 is equivalent to what Haartsen and Pride [7] call the PSVTM system, since, as we will see, it contains compressional (P) waves, vertical shear (SV) waves, and transverse magnetic (TM) waves. For this system the submatrices are

(2.14)

$$\mathbf{M_1}^{(1)} = \begin{bmatrix} -\beta_1 M & \beta_1\gamma(C^2 - \lambda M) & -\beta_1 C & 0 \\ \beta_1\gamma(C^2 - \lambda M) & \rho + i\omega\rho_f^2\frac{\kappa}{\eta} - 4\beta_1\gamma^2 G[C^2 - M(\lambda + G)] & 2\beta_1\gamma GC - i\omega\rho_f\gamma\frac{\kappa}{\eta} & \rho_f L \\ -\beta_1 C & 2\beta_1\gamma GC - i\omega\rho_f\gamma\frac{\kappa}{\eta} & -\beta_1(\lambda + 2G) + i\omega\gamma^2\frac{\kappa}{\eta} & -\gamma L \\ 0 & \rho_f L & -\gamma L & \frac{\bar{\sigma}}{i\omega} \end{bmatrix},$$

$$\mathbf{M_2}^{(1)} = \begin{bmatrix} \rho & \gamma & -\rho_f & 0 \\ \gamma & \frac{1}{G} & 0 & 0 \\ -\rho_f & 0 & \frac{-\beta_2\eta}{i\omega\kappa} & \frac{-\beta_2\gamma L\eta}{\kappa\bar{\sigma}} \\ 0 & 0 & \frac{-\beta_2\gamma L\eta}{\kappa\bar{\sigma}} & -\mu - i\frac{\omega\beta_2\gamma^2}{\bar{\sigma}} \end{bmatrix}. \tag{2.15}$$

The corresponding source vector is

$$\mathbf{S}^{(1)} = \left[0, -\tilde{F}_1 - i\omega\rho_f\frac{\kappa}{\eta}\tilde{f}_1, ik\frac{\kappa}{\eta}\tilde{f}_1, -\tilde{j}_1 - L\tilde{f}_1, -\tilde{F}_3, 0, \tilde{f}_3 - \beta_2\frac{L\eta}{\kappa\bar{\sigma}}\tilde{j}_3, -ik\frac{\beta_2}{\bar{\sigma}}\tilde{j}_3\right]^{\mathbf{T}}. \tag{2.16}$$

Once $\Phi^{(1)}$ has been determined, we may also compute the following four variables, which are dependent on System 1 only:

$$\begin{aligned} \tilde{E}_3 &= \beta_2\left(\frac{ik}{\bar{\sigma}}\tilde{H}_2 - \frac{L\eta}{\kappa\bar{\sigma}}\dot{\tilde{w}}_3 - \frac{1}{\bar{\sigma}}\tilde{j}_3\right), \\ \dot{\tilde{w}}_1 &= L\tilde{E}_1 - ik\frac{\kappa}{\eta}\tilde{p} + i\omega\rho_f\frac{\kappa}{\eta}\dot{\tilde{u}}_1 + \frac{\kappa}{\eta}\tilde{f}_1, \\ \tilde{\tau}_{11} &= \beta_1\left(-4\gamma G(C^2 - M[\lambda + G])\dot{\tilde{u}}_1 + (C^2 - \lambda M)\tilde{\tau}_{33} + 2GC\tilde{p}\right), \\ \tilde{\tau}_{22} &= \beta_1\left(-2\gamma G[C^2 - \lambda M]\dot{\tilde{u}}_1 + [C^2 - \lambda M]\tilde{\tau}_{33} + 2GC\tilde{p}\right). \end{aligned} \tag{2.17}$$

The $(2n_2 = 4)$-dimensional System 2 is equivalent to what Haartsen and Pride [7] call the SHTE system, since, as we shall see, it contains shear horizontal (SH) waves and transverse electric (TE) waves. For this system we obtain

$$\mathbf{M_1}^{(2)} = \begin{bmatrix} \frac{1}{G} & 0 \\ 0 & -\mu \end{bmatrix}, \tag{2.18}$$

$$\mathbf{M_2}^{(2)} = \begin{bmatrix} \rho - G\gamma^2 + i\omega\rho_f^2 \frac{\kappa}{\eta} & \rho_f L \\ \rho_f L & \frac{\bar{\sigma}}{i\omega} + \frac{\gamma^2}{\mu} \end{bmatrix}, \tag{2.19}$$

and source vector

$$\mathbf{S}^{(2)} = \left[0, 0, -\tilde{F}_2 - i\omega\rho_f \frac{\kappa}{\eta}\tilde{f}_2, -\tilde{j}_2 - L\tilde{f}_2\right]^{\mathbf{T}}. \tag{2.20}$$

Once $\Phi^{(2)}$ has been determined, we may also compute the following three variables, which are dependent on System 2 only:

$$\begin{aligned} \tilde{H}_3 &= \frac{\gamma}{\mu}\tilde{E}_2, \\ \dot{\tilde{w}}_2 &= L\tilde{E}_2 + i\omega\frac{\kappa}{\eta}\rho_f\dot{\tilde{u}}_2 + \frac{\kappa}{\eta}\tilde{f}_2, \\ \tilde{\tau}_{12} &= -G\gamma\dot{\tilde{u}}_2. \end{aligned} \tag{2.21}$$

To construct an algorithm for a fast computer code, we will restrict the calculations to the case where the material parameters are piecewise constant. Thus it is assumed that the material properties are constant within each layer but change discontinuously as z is varied across a layer boundary which is a horizontal interface. Then (2.11) is satisfied within each layer, where $\mathbf{M}^{(m)}$ is constant. At layer boundaries we apply Pride's interface condition, that $\mathbf{u}, p$, the normal components of $\mathbf{w}$ and τ, and the tangential components of $\mathbf{E}$ and $\mathbf{H}$ are continuous. Then it is seen that the vectors $\Phi^{(m)}$ are continuous across layer boundaries.

It remains to give boundary conditions for Systems 1 and 2 at the earth/air interface at $z = 0$. Applying Pride's interface conditions, we get that the boundary conditions for System 1 are

$$\tilde{\tau}_{13} = \tilde{\tau}_{33} = 0, \quad \tilde{p} = 0, \quad \tilde{H}_2 = -\frac{\epsilon_0}{q_0}\tilde{E}_1 \quad \text{at } z = 0. \tag{2.22}$$

In (2.22) q_0 is the vertical slowness of an EM wave in the air, that is,

$$q_0 = \sqrt{\epsilon_0\mu_0 - \gamma^2}, \tag{2.23}$$

where μ_0 is the magnetic permeability of the air so that $\epsilon_0\mu_0$ is the reciprocal of the square of the speed of light. The last of equations (2.22) is derived from the fact that there are no downgoing waves in the air (all sources are in the subsurface), and this is the relationship for an upgoing EM wave.

The boundary conditions for System 2 are

$$\tilde{\tau}_{23} = 0, \;\; \tilde{H}_1 = \frac{q_0}{\mu_0}\tilde{E}_2 \quad \text{at } z = 0. \tag{2.24}$$

The second of these relationships is again derived as the condition that there are only upgoing EM waves in the air.

Note that (2.22) gives $n_1 = 4$ conditions for System 1, which has $2n_1 = 8$ variables. Similarly, (2.24) gives $n_2 = 2$ conditions for System 2, which has $2n_2 = 4$ variables. Consequently, for each system we will need an additional n_m conditions to completely specify the solution. These relations will come from the requirement that there are no upgoing waves from $z = \infty$. The decomposition into upgoing and downgoing waves in the subsurface will be accomplished in section 4.

3. Ursin diagonalization. For completeness we give a derivation of Ursin's diagonalization procedure [23] in the form that it will be used here. We consider matrices of the form (2.12), where for simplicity we drop the superscript $^{(m)}$.

Assume that $\mathbf{M_1M_2}$ has n distinct nonzero eigenvalues $q_1^2, q_2^2, \ldots, q_n^2$ with associated eigenvectors $\mathbf{a}_1, \mathbf{a_2}, \ldots, \mathbf{a}_n$, i.e.,

$$\mathbf{M_1M_2}\mathbf{a}_m = q_m^2\mathbf{a}_m, \tag{3.1}$$

normalized by the relation

$$\mathbf{a}_m^{\mathbf{T}}\mathbf{M_2}\mathbf{a}_m = q_m, \quad m = 1, \ldots, n. \tag{3.2}$$

Here $q_m = \sqrt{q_m^2}$ with the branch of the square root chosen so that $\mathrm{Im}(q_m) \geq 0$ and $q_m > 0$ if q_m is real. (With this choice, $e^{i\omega q_m z}$ represents a downgoing wave.) We adopt this branch of the square root throughout this paper.

Define

$$\mathbf{b}_m = \frac{1}{q_m}\mathbf{M_2}\mathbf{a}_m. \tag{3.3}$$

$\mathbf{b}_m$ is an eigenvector of $\mathbf{M_2M_1}$, with eigenvalue q_m^2, as can be seen by multiplying (3.3) by $\mathbf{M_2M_1}$ and using (3.1) and (3.3). From (2.13), $\mathbf{b}_m$ is a left eigenvector of $\mathbf{M_1M_2}$.

Now

$$q_m^2\mathbf{a}_j^{\mathbf{T}}\mathbf{b}_m = \mathbf{a}_j^{\mathbf{T}}\mathbf{M_2M_1}\mathbf{b}_m = \mathbf{b}_m^{\mathbf{T}}\mathbf{M_1M_2}\mathbf{a}_j = q_j^2\mathbf{a}_j^{\mathbf{T}}\mathbf{b}_m. \tag{3.4}$$

Therefore $\mathbf{a}_j^{\mathbf{T}}\mathbf{b}_m = 0$ for $j \neq m$. Coupling this with (3.2) and (3.3) we obtain

$$\mathbf{a}_j^{\mathbf{T}}\mathbf{b}_m = \delta_{j,m}. \tag{3.5}$$

Let $\mathbf{L_1}$ be the $n \times n$ matrix whose mth column is $\mathbf{a}_m$, and let $\mathbf{L_2}$ be the $n \times n$ matrix whose mth column is $\mathbf{b}_m$. Then (3.5) implies

$$\mathbf{L_1^{-1}} = \mathbf{L_2^T}, \ \ \mathbf{L_2^{-1}} = \mathbf{L_1^T}. \tag{3.6}$$

Let Λ be the $n \times n$ diagonal matrix with entries

$$\Lambda_{j,m} = q_j\delta_{j,m}. \tag{3.7}$$

Then (3.3) implies

$$\mathbf{L_2}\Lambda = \mathbf{M_2L_1}. \tag{3.8}$$

Multiplication of (3.3) by $\mathbf{M_1}$ and use of (3.1) implies

$$\mathbf{M_1 b}_m = q_m \mathbf{a}_m, \tag{3.9}$$

and so

$$\mathbf{M_1 L_2} = \mathbf{L_1}\Lambda. \tag{3.10}$$

Now (3.10), (3.8), and (3.6) yield

$$\mathbf{M_1} = \mathbf{L_1}\Lambda\mathbf{L_1^T}, \;\; \mathbf{M_2} = \mathbf{L_2}\Lambda\mathbf{L_2^T}. \tag{3.11}$$

Let $\mathbf{L}$ be the $2n \times 2n$ matrix defined in block form by

$$\mathbf{L} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{L_1} & \mathbf{L_1} \\ \mathbf{L_2} & -\mathbf{L_2} \end{bmatrix}, \tag{3.12}$$

and let $\tilde{\Lambda}$ be the diagonal matrix

$$\tilde{\Lambda} = \begin{bmatrix} \Lambda & \mathbf{0} \\ \mathbf{0} & -\Lambda \end{bmatrix}. \tag{3.13}$$

Then from (3.6) and (3.11) it is readily verified that

$$\mathbf{L}^{-1} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{L_2^T} & \mathbf{L_1^T} \\ \mathbf{L_2^T} & -\mathbf{L_1^T} \end{bmatrix} \tag{3.14}$$

and that

$$\mathbf{M} = \mathbf{L}\tilde{\Lambda}\mathbf{L}^{-1}. \tag{3.15}$$

In the appendices we give explicit algebraic formulas for $q_m, \mathbf{a}_m, \mathbf{b}_m$ for Systems 1 and 2 as described in section 2. From these formulas $\mathbf{L_1}, \mathbf{L_2}, \mathbf{L}, \mathbf{L}^{-1}, \Lambda$ can be calculated rapidly.

4. Upgoing and downgoing waves; reflection and transmission matrices. We first consider a homogeneous, source-free region of space. Then dropping the superscript $^{(m)}$ we have a $2n$-dimensional equation of the form (2.11), with (2.12) and (2.13), and with $\mathbf{M}$ constant and $\mathbf{S} = \mathbf{0}$. Let

$$\Phi = \mathbf{L}\Psi. \tag{4.1}$$

Insertion of (4.1) into (2.11) and use of (3.15) yields

$$\frac{d}{dz}\Psi = -i\omega\tilde{\Lambda}\Psi. \tag{4.2}$$

Let

$$\Psi = \begin{bmatrix} \mathbf{U} \\ \mathbf{D} \end{bmatrix}, \tag{4.3}$$

where $\mathbf{U}, \mathbf{D}$ are n-vectors. Then from (4.2), (4.3), and (3.13)

$$\Psi(z) = e^{-i\omega\tilde{\Lambda}(z-z_0)}\Psi(z_0) = \begin{bmatrix} e^{-i\omega\Lambda(z-z_0)} & \mathbf{U}(z_0) \\ e^{i\omega\Lambda(z-z_0)} & \mathbf{D}(z_0) \end{bmatrix}, \tag{4.4}$$

where z_0 is a fixed point in the same source-free region of space as z. Here $e^{\pm i\omega\Lambda(z-z_0)}$ are diagonal matrices with mth diagonal element equal to $e^{\pm i\omega q_m(z-z_0)}$. Therefore $\mathbf{U}$ are upgoing waves and $\mathbf{D}$ are downgoing waves.

Next consider an interface at $z=\bar{z}$, where the material parameters vary discontinuously across $\bar{z}$. We denote by superscript $^+$ quantities evaluated at $\bar{z}^+$, just below the interface, while superscript $^-$ denotes quantities evaluated at $\bar{z}^-$ just above the interface. Since Φ is continuous across $\bar{z}$, we obtain from (4.1) that $\mathbf{L}^+\Psi^+ = \mathbf{L}^-\Psi^-$ and so

$$\Psi^+ = \mathbf{J}\Psi^-, \quad \Psi^- = \mathbf{J}^{-1}\Psi^+, \tag{4.5}$$

where the jump matrix is

$$\mathbf{J} = \left(\mathbf{L}^+\right)^{-1}\mathbf{L}^- = \begin{bmatrix} \mathbf{J_A} & \mathbf{J_B} \\ \mathbf{J_B} & \mathbf{J_A} \end{bmatrix} \tag{4.6}$$

and, from (3.12) and (3.14), $\mathbf{J_A}, \mathbf{J_B}$ are the $n\times n$ matrices

$$\begin{aligned} \mathbf{J_A} &= \frac{1}{2}\Big[\left(\mathbf{L_2^+}\right)^{\mathbf{T}}\mathbf{L_1^-} + \left(\mathbf{L_1^+}\right)^{\mathbf{T}}\mathbf{L_2^-}\Big], \\ \mathbf{J_B} &= \frac{1}{2}\Big[\left(\mathbf{L_2^+}\right)^{\mathbf{T}}\mathbf{L_1^-} - \left(\mathbf{L_1^+}\right)^{\mathbf{T}}\mathbf{L_2^-}\Big]. \end{aligned} \tag{4.7}$$

$\mathbf{J}^{-1}$ can be computed by interchanging the roles of $\pm$. Using this fact in (4.6) and (4.7) yields

$$\mathbf{J}^{-1} = \begin{bmatrix} \mathbf{J_A^T} & -\mathbf{J_B^T} \\ -\mathbf{J_B^T} & \mathbf{J_A^T} \end{bmatrix}. \tag{4.8}$$

Now using that $\mathbf{J}\mathbf{J}^{-1} = \mathbf{J}^{-1}\mathbf{J} = \mathbf{I}$ yields four relations:

$$\mathbf{J_A}\mathbf{J_A^T} - \mathbf{J_B}\mathbf{J_B^T} = \mathbf{I}, \tag{4.9}$$

$$\mathbf{J_A^T}\mathbf{J_A} - \mathbf{J_B^T}\mathbf{J_B} = \mathbf{I}, \tag{4.10}$$

$$\mathbf{J_A}\mathbf{J_B^T} = \mathbf{J_B}\mathbf{J_A^T}, \tag{4.11}$$

$$\mathbf{J_A^T}\mathbf{J_B} = \mathbf{J_B^T}\mathbf{J_A}. \tag{4.12}$$

We next consider a stack of layers, with layer boundaries at the interfaces at depths $0 < z_1 < z_2 < \cdots < z_N < \infty$. Homogeneous, source-free regions are assumed in (z_m, z_{m+1}). We denote by subscript m a quantity at interface z_m, with superscripts $^\pm$ as before. From (4.3) and (4.5),

$$\begin{bmatrix} \mathbf{U}_N^- \\ \mathbf{D}_N^- \end{bmatrix} = \mathbf{J}_N^{-1}\begin{bmatrix} \mathbf{0} \\ \mathbf{D}_N^+ \end{bmatrix}, \tag{4.13}$$

where we have used that there is no upgoing wave below the last interface at $z=z_N$. From (4.8) and (4.13),

$$\begin{aligned} \mathbf{U}_N^- &= \Gamma_N\mathbf{D}_N^-, \\ \mathbf{D}_N^+ &= \mathbf{T}_N\mathbf{D}_N^-, \end{aligned} \tag{4.14}$$

where

$$\begin{aligned} \Gamma_N &= -\big(\mathbf{J}^{\mathbf{T}}_{\mathbf{B},N}\big)\big(\mathbf{J}^{\mathbf{T}}_{\mathbf{A},N}\big)^{-1}, \\ \mathbf{T}_N &= \big(\mathbf{J}^{\mathbf{T}}_{\mathbf{A},N}\big)^{-1}. \end{aligned} \tag{4.15}$$

Here Γ_N is the reflection matrix from a single interface at $z = z_N$ and can be used to compute the reflected wave, i.e., the upgoing wave from the top of the interface, when the incident wave, i.e., the downgoing wave on the top of the interface, is known. Similarly $\mathbf{T}_N$ is the single interface transmission coefficient and can be used to compute the transmitted wave, i.e., the downgoing wave beneath the interface, when the incident wave is known.

Let $j < N$ and define the layer thicknesses

$$\Delta z_m = z_{m+1} - z_m, \quad m = 1, 2, \ldots, N-1. \tag{4.16}$$

Then by jumping across the layer boundary using (4.5) and propagating through the layer using (4.4) we obtain

$$\begin{aligned} \mathbf{U}^-_m &= \mathbf{J}^{\mathbf{T}}_{\mathbf{A},m} e^{i\omega\Lambda_m \Delta z_m} \mathbf{U}^-_{m+1} - \mathbf{J}^{\mathbf{T}}_{\mathbf{B},m} e^{-i\omega\Lambda_m \Delta z_m} \mathbf{D}^-_{m+1}, \\ \mathbf{D}^-_m &= -\mathbf{J}^{\mathbf{T}}_{\mathbf{B},m} e^{i\omega\Lambda_m \Delta z_m} \mathbf{U}^-_{m+1} + \mathbf{J}^{\mathbf{T}}_{\mathbf{A},m} e^{-i\omega\Lambda_m \Delta z_m} \mathbf{D}^-_{m+1}. \end{aligned} \tag{4.17}$$

Define reflection and transmission matrices $\Gamma_m, \mathbf{T}_m$ by the relations that for any incident wave $\mathbf{D}^-_m$ at the top of the stack of layers underlying $z = z_m$

$$\begin{aligned} \mathbf{U}^-_m &= \Gamma_m \mathbf{D}^-_m, \\ \mathbf{D}^+_N &= \mathbf{T}_m \mathbf{D}^-_m. \end{aligned} \tag{4.18}$$

Therefore Γ_m computes the reflected wave from the stack and $\mathbf{T}_m$ computes the transmitted wave below the stack, when the incident wave is known. From (4.17) and (4.18) we obtain by induction

$$\Gamma_m = \big(\mathbf{J}^{\mathbf{T}}_{\mathbf{A},m} \tilde{\Gamma}_{m+1} - \mathbf{J}^{\mathbf{T}}_{\mathbf{B},m}\big)\big(-\mathbf{J}^{\mathbf{T}}_{\mathbf{B},m} \tilde{\Gamma}_{m+1} + \mathbf{J}^{\mathbf{T}}_{\mathbf{A},m}\big)^{-1}, \tag{4.19}$$

$$\mathbf{T}_m = \mathbf{T}_{m+1} e^{i\omega\Lambda_m \Delta z_m} \big(-\mathbf{J}^{\mathbf{T}}_{\mathbf{B},m} \tilde{\Gamma}_{m+1} + \mathbf{J}^{\mathbf{T}}_{\mathbf{A},m}\big)^{-1}, \tag{4.20}$$

where

$$\tilde{\Gamma}_{m+1} = e^{i\omega\Lambda_m \Delta z_m} \Gamma_{m+1} e^{i\omega\Lambda_m \Delta z_m}. \tag{4.21}$$

Now all the reflection and transmission matrices can be computed by recursion using (4.19) and (4.20), starting with (4.15).

Finally, it can be shown that Γ_m is symmetric:

$$\Gamma_m = \Gamma^{\mathbf{T}}_m. \tag{4.22}$$

To see this by induction, first note that Γ_N is symmetric by using (4.11) in (4.15) to obtain that Γ_N is equal to its transpose:

$$\Gamma_N = -\left(\mathbf{J}_{\mathbf{A},N}\right)^{-1}\left(\mathbf{J}_{\mathbf{B},N}\right). \tag{4.23}$$

Next, assume that Γ_{m+1} and therefore $\tilde{\Gamma}_{m+1}$ are symmetric. From (4.9) we obtain for any symmetric $\tilde{\Gamma}$

$$\tilde{\Gamma}\mathbf{J_B}\mathbf{J_B^T} + \mathbf{J_A}\mathbf{J_A^T}\tilde{\Gamma} = \tilde{\Gamma}\mathbf{J_A}\mathbf{J_A^T} + \mathbf{J_B}\mathbf{J_B^T}\tilde{\Gamma}. \tag{4.24}$$

From (4.11) we obtain

$$\tilde{\Gamma}\mathbf{J_B}\mathbf{J_A^T}\tilde{\Gamma} + \mathbf{J_A}\mathbf{J_B^T} = \tilde{\Gamma}\mathbf{J_A}\mathbf{J_B^T}\tilde{\Gamma} + \mathbf{J_B}\mathbf{J_A^T}. \tag{4.25}$$

Subtraction of (4.25) from (4.24) and factoring the result gives

$$\big(-\tilde{\Gamma}\mathbf{J_B} + \mathbf{J_A}\big)\big(\mathbf{J_A^T}\tilde{\Gamma} - \mathbf{J_B^T}\big) = \big(\tilde{\Gamma}\mathbf{J_A} - \mathbf{J_B}\big)\big(- \mathbf{J_B^T}\tilde{\Gamma} + \mathbf{J_A^T}\big). \tag{4.26}$$

Now use of (4.26) in (4.19) gives the alternative recursion formula

$$\Gamma_m = \big(-\tilde{\Gamma}_{m+1}\mathbf{J_{B},}_m + \mathbf{J_{A},}_m\big)^{-1}\big(\tilde{\Gamma}_{m+1}\mathbf{J_{A},}_m - \mathbf{J_{B},}_m\big). \tag{4.27}$$

Comparison of (4.27) with (4.19) gives (4.22).

5. Sources and boundary conditions. We consider a $2n$-dimensional system of type (2.11) with the superscript (m) omitted. Let the source, at a depth z_s, be of the form

$$\mathbf{S} = \mathbf{S_0}\delta(z - z_s) + \mathbf{S_1}\delta'(z - z_s) \tag{5.1}$$

with $\mathbf{S_0}, \mathbf{S_1}$ independent of z. More generally, sources distributed in the depth coordinate may be synthesized by superposition of sources of this type. Let

$$\Phi_\mathbf{0} = \Phi - \mathbf{S_1}\delta(z - z_s). \tag{5.2}$$

Then from (5.1), (5.2), and (2.11),

$$\frac{d}{dz}\Phi_\mathbf{0} = -i\omega\mathbf{M}\Phi_\mathbf{0} + [\mathbf{S_0} - i\omega\mathbf{M}\mathbf{S_1}]\,\delta(z - z_s). \tag{5.3}$$

Let $\mathbf{S_A}, \mathbf{S_B}$ be n-vectors defined so that

$$\begin{bmatrix}\mathbf{S_A}\\ \mathbf{S_B}\end{bmatrix} = i\omega\mathbf{M}\mathbf{S_1} - \mathbf{S_0}. \tag{5.4}$$

Then by integrating (5.3) from just above the source at z_s^- to just below the source at z_s^+, and using (5.2) and (5.4) we obtain the jump condition across the source:

$$\Phi(z_s^-) = \Phi(z_s^+) + \begin{bmatrix}\mathbf{S_A}\\ \mathbf{S_B}\end{bmatrix}. \tag{5.5}$$

We insert a fictitious layer boundary just below the source at $z = z_s^+$ and use the methods of section 4 to compute the reflection matrix $\Gamma_\mathbf{s} \equiv \Gamma(z_s^+)$ from the top of this layer. Note that at z_s^+, $\mathbf{J_A} = \mathbf{I}, \mathbf{J_B} = \mathbf{0}$, since the material properties do not change at z_s. Then the upgoing wave $\mathbf{U_s} \equiv \mathbf{U}(z_s^+)$ is related to the downgoing wave $\mathbf{D_s} \equiv \mathbf{D}(z_s^+)$ there by (4.18), and so

$$\Psi(z_s^+) = \begin{bmatrix}\Gamma_\mathbf{s}\mathbf{D_s}\\ \mathbf{D_s}\end{bmatrix}. \tag{5.6}$$

From (5.6), (5.5), (4.1), and (3.14) we obtain

$$\Psi(z_s^-) = \begin{bmatrix} \Gamma_{\mathbf{s}}\mathbf{D_s} \\ \mathbf{D_s} \end{bmatrix} + \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{L_2^T S_A} + \mathbf{L_1^T S_B} \\ \mathbf{L_2^T S_A} - \mathbf{L_1^T S_B} \end{bmatrix}. \tag{5.7}$$

The expression (5.7) may now be propagated upwards through layers, using (4.4) and jumped upwards across layer boundaries using (4.5) until we reach the earth/air interface at $z = 0+$. Then the n boundary conditions at $z = 0$ can be used to determine the n unknowns $\mathbf{D_s}$. We will write the formulas explicitly for when z_s is in the first subsurface layer, i.e., $0 < z_s < z_1$. In this case

$$\Psi(0^+) = \begin{bmatrix} e^{i\omega\Lambda z_s}\Gamma_{\mathbf{s}}\mathbf{D_s} \\ e^{-i\omega\Lambda z_s}\mathbf{D_s} \end{bmatrix} + \frac{1}{\sqrt{2}} \begin{bmatrix} e^{i\omega\Lambda z_s}\left(\mathbf{L_2^T S_A} + \mathbf{L_1^T S_B}\right) \\ e^{-i\omega\Lambda z_s}\left(\mathbf{L_2^T S_A} - \mathbf{L_1^T S_B}\right) \end{bmatrix}. \tag{5.8}$$

We next write

$$\Phi(0^+) = \begin{bmatrix} \mathbf{G_A}\Phi_{\mathbf{g}} \\ \mathbf{G_B}\Phi_{\mathbf{g}} \end{bmatrix}, \tag{5.9}$$

where $\Phi_{\mathbf{g}}$ is an n-vector of unknowns at $z = 0$ and $\mathbf{G_A}, \mathbf{G_B}$ are $n \times n$ matrices.

For System 1, let

$$\Phi_{\mathbf{g}}{}^{(1)} = \left[\dot{\tilde{u}}_3, -\dot{\tilde{w}}_3, \dot{\tilde{u}}_1, \tilde{E}_1\right]^{\mathbf{T}}_{z=0^+},$$

$$\mathbf{G_A}{}^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -\frac{\epsilon_0}{q_0} \end{bmatrix}, \tag{5.10}$$

$$\mathbf{G_B}{}^{(1)} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Then it may be checked that (5.9) holds for System 1 with the boundary conditions given by (2.22).

For System 2 let

$$\Phi_{\mathbf{g}}{}^{(2)} = \left[\dot{\tilde{u}}_2, \tilde{E}_2\right]^{\mathbf{T}}_{z=0^+},$$

$$\mathbf{G_A}{}^{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{5.11}$$

$$\mathbf{G_B}{}^{(2)} = \begin{bmatrix} 0 & 0 \\ 0 & -\frac{q_0}{\mu_0} \end{bmatrix}.$$

Then it may be checked that (5.9) holds for System 2 with the boundary conditions given by (2.24).

Now using (4.1) we may combine equations (5.8) and (5.9) to get

$$\begin{aligned} \Phi_{\mathbf{g}} = &\left[e^{i\omega\Lambda z_s}\Gamma_{\mathbf{s}}e^{i\omega\Lambda z_s}\left(\mathbf{L_2^T G_A} - \mathbf{L_1^T G_B}\right) - \left(\mathbf{L_2^T G_A} + \mathbf{L_1^T G_B}\right)\right]^{-1} \\ &\times e^{i\omega\Lambda z_s}\left[\Gamma_{\mathbf{s}}\left(\mathbf{L_2^T S_A} - \mathbf{L_1^T S_B}\right) - \left(\mathbf{L_2^T S_A} + \mathbf{L_1^T S_B}\right)\right], \end{aligned} \tag{5.12}$$

$$\mathbf{D_s} = \frac{1}{\sqrt{2}} e^{i\omega\Lambda z_s} \left(\mathbf{L_2^T G_A} - \mathbf{L_1^T G_B}\right) \Phi_\mathbf{g} - \frac{1}{\sqrt{2}} \left(\mathbf{L_2^T S_A} - \mathbf{L_1^T S_B}\right). \tag{5.13}$$

In particular, when the source is just below the surface we get

$$\begin{aligned} \Phi_\mathbf{g} = &\left[(\Gamma_\mathbf{s} - \mathbf{I})\mathbf{L_2^T G_A} - (\Gamma_\mathbf{s} + \mathbf{I})\mathbf{L_1^T G_B}\right]^{-1} \\ &\times \left[(\Gamma_\mathbf{s} - \mathbf{I})\mathbf{L_2^T S_A} - (\Gamma_\mathbf{s} + \mathbf{I})\mathbf{L_1^T S_B}\right] \quad \text{as } z_s \downarrow 0. \end{aligned} \tag{5.14}$$

$\Phi_\mathbf{g}$ gives all of Φ at the surface $z = 0$, and $\mathbf{D_s}, \mathbf{U_s} = \Gamma_\mathbf{s}\mathbf{D_s}$ give all of Φ just below the source. Now Φ can theoretically be computed anywhere else in space by propagating through layers using (4.4) and jumping across the layer boundaries using (4.5). However, propagation of an upward-going wave in the downward direction is unstable numerically using (4.4), because the complex exponentials grow rather than decay with distance. So numerically, one must obtain $\mathbf{U}$ from $\mathbf{D}$ using Γ_m, or make use of the transmission matrices $\mathbf{T}_m$.

From the tilde variables, the lateral Fourier transforms, i.e., the hat variables, can be computed by inverting the rotation in (2.8) via the formulas

$$\begin{aligned} &\hat{\mathbf{E}} = \Omega^\mathbf{T}\tilde{\mathbf{E}}, \;\; \hat{\mathbf{H}} = \Omega^\mathbf{T}\tilde{\mathbf{H}}, \\ &\hat{\mathbf{u}} = \Omega^\mathbf{T}\tilde{\mathbf{u}}, \;\; \hat{\mathbf{w}} = \Omega^\mathbf{T}\tilde{\mathbf{w}}, \\ &\hat{\tau} = \Omega^\mathbf{T}\tilde{\tau}\Omega, \;\; \hat{p} = \tilde{p}. \end{aligned} \tag{5.15}$$

It remains to invert the lateral Fourier transforms via (2.6) to get the solution in real space. Note that the matrices for Systems 1 and 2 depend only on the magnitude, k (or equivalently the slowness γ), of the vector $(k_1, k_2)^T$ and not on its direction. However, factors of k_1 and k_2 are introduced by the transformation (5.15) and possibly by the directionality of the source. For any function $\hat{g}(k)$ let

$$\Theta_{m_1,m_2} \equiv \mathcal{F}^{-1}\left[k_1^{m_1} k_2^{m_2} \hat{g}(k)\right] = (-i)^{m_1+m_2} \left(\frac{\partial}{\partial x_1}\right)^{m_1} \left(\frac{\partial}{\partial x_2}\right)^{m_2} \mathcal{F}^{-1}\left[\hat{g}(k)\right]. \tag{5.16}$$

In cylindrical coordinates $x_1 = r\cos\theta, x_2 = r\sin\theta, x_3 = z$ these quantities may be computed as Hankel transforms. Let J_m be the Bessel functions, and for nonnegative integers m_1, m_2 let

$$\mathcal{B}_{m_1,m_2}[\hat{g}] = \frac{1}{2\pi} \int_0^\infty k^{m_1} J_{m_2}(kr)\hat{g}(k)dk. \tag{5.17}$$

Then, in particular,

$$\begin{aligned} &\Theta_{0,0} = \mathcal{B}_{1,0}, \;\; \Theta_{1,0} = i\cos\theta\mathcal{B}_{2,1}, \;\; \Theta_{0,1} = i\sin\theta\mathcal{B}_{2,1}, \\ &\Theta_{1,1} = \sin\theta\cos\theta\left[\mathcal{B}_{3,0} - \frac{2}{r}\mathcal{B}_{2,1}\right], \;\; \Theta_{2,0} = \cos^2\theta\mathcal{B}_{3,0} - \frac{\left(\cos^2\theta - \sin^2\theta\right)}{r}\mathcal{B}_{2,1}, \\ &\Theta_{0,2} = \sin^2\theta\mathcal{B}_{3,0} + \frac{\left(\cos^2\theta - \sin^2\theta\right)}{r}\mathcal{B}_{2,1}. \end{aligned} \tag{5.18}$$

6. Dynamite; hammer, weight drop, and vibroseis. For a dynamite source we take the imposed forces on the solid and the fluid to be

$$\mathbf{F}(\mathbf{x}) = \mathbf{f}(\mathbf{x}) = -\hat{h}(\omega)\nabla\delta(\mathbf{x} - \mathbf{x}_s), \tag{6.1}$$

where the position of the dynamite charge is $\mathbf{x}_s = (0, 0, z_s)^T$ and $\hat{h}(\omega)$ is the spectrum of the seismic moment $h(t)$. There are no externally applied electrical currents, so $\mathbf{j} = \mathbf{0}$. Lateral Fourier transform via (2.5) yields

$$\hat{\mathbf{F}}(k_1, k_2, z) = \hat{\mathbf{f}}(k_1, k_2, z) = -\hat{h}(\omega) \begin{bmatrix} ik_1\delta(z - z_s) \\ ik_2\delta(z - z_s) \\ \delta'(z - z_s) \end{bmatrix}, \tag{6.2}$$

and rotation by Ω via (2.8) yields

$$\tilde{\mathbf{F}}(k_1, k_2, z) = \tilde{\mathbf{f}}(k_1, k_2, z) = -\hat{h}(\omega) \begin{bmatrix} ik\delta(z - z_s) \\ 0 \\ \delta'(z - z_s) \end{bmatrix}. \tag{6.3}$$

Substitution of (6.3) into (2.16) yields the source for System 1, in the form (5.1), with

$$\begin{aligned} \mathbf{S_0}^{(1)} &= \hat{h}(\omega) \left[0, ik\left(1 + i\omega\rho_f \frac{\kappa}{\eta}\right), k^2 \frac{\kappa}{\eta}, ikL, 0, 0, 0, 0\right]^T, \\ \mathbf{S_1}^{(1)} &= \hat{h}(\omega) \left[0, 0, 0, 0, 1, 0, -1, 0\right]^T. \end{aligned} \tag{6.4}$$

Substitution of (6.3) into (2.20) shows that the source for System 2 is identically zero, and so all variables associated with System 2 are zero. This is to be expected because System 2 contains SH waves, which are not excited by dynamite.

Substitution of (6.4) into (5.4) gives

$$\begin{aligned} \mathbf{S_A}^{(1)} &= i\beta_1 \hat{h}(\omega) \left[\omega(C - M), 2kG(M - C), \omega(\lambda + 2G - C), 0\right]^T, \\ \mathbf{S_B}^{(1)} &= \mathbf{0}. \end{aligned} \tag{6.5}$$

Now (6.5) may be used in (5.12) and (5.13), or in (5.14) for a shallow source, to get all the tilde variables.

To invert the rotation Ω, using (5.15), note that from (2.10) and (2.21) and the vanishing of System 2, the following variables are identically zero: $\dot{\tilde{u}}_2, \tilde{E}_2, \tilde{\tau}_{23}, \tilde{H}_1, \tilde{H}_3$, $\dot{\tilde{w}}_2, \tilde{\tau}_{12}$. Furthermore, all the remaining tilde variables are functions of k alone, not k_1, k_2 individually. Therefore

$$\begin{aligned} \dot{\hat{u}}_1 &= \frac{k_1}{k} \dot{\tilde{u}}_1, \quad \hat{E}_1 = \frac{k_1}{k} \tilde{E}_1, \quad \hat{H}_1 = -\frac{k_2}{k} \tilde{H}_2, \\ \dot{\hat{u}}_2 &= \frac{k_2}{k} \dot{\tilde{u}}_1, \quad \hat{E}_2 = \frac{k_2}{k} \tilde{E}_1, \quad \hat{H}_2 = \frac{k_1}{k} \tilde{H}_2, \\ \dot{\hat{u}}_3 &= \dot{\tilde{u}}_3, \qquad \hat{E}_3 = \tilde{E}_3, \qquad \hat{H}_3 = 0, \end{aligned} \tag{6.6}$$

and so the transforms can be inverted in cylindrical coordinates (r, θ, z) using (5.16)–(5.18):

$$\begin{aligned} \dot{\mathbf{u}} &= \left(i\mathcal{B}_{1,1}\left[\dot{\tilde{u}}_1\right]\right) \mathbf{e_r} + \left(\mathcal{B}_{1,0}\left[\dot{\tilde{u}}_3\right]\right) \mathbf{e_z}, \\ \mathbf{E} &= \left(i\mathcal{B}_{1,1}\left[\tilde{E}_1\right]\right) \mathbf{e_r} + \left(\mathcal{B}_{1,0}\left[\tilde{E}_3\right]\right) \mathbf{e_z}, \\ \mathbf{H} &= \left(i\mathcal{B}_{1,1}\left[\tilde{H}_2\right]\right) \mathbf{e}_\theta. \end{aligned} \tag{6.7}$$

Here $\mathbf{e_r}, \mathbf{e}_\theta, \mathbf{e_z}$ are unit vectors in the r, θ, z coordinate directions, respectively.

The stress tensor may be inverted similarly. For the stresses, we get from (5.15)

$$
\begin{aligned}
&\hat{\tau}_{11} = \frac{1}{k^2}\left[k_1^2\tilde{\tau}_{11} + k_2^2\tilde{\tau}_{22}\right], \quad \hat{\tau}_{12} = \frac{k_1k_2}{k^2}\left[\tilde{\tau}_{11} - \tilde{\tau}_{22}\right], \\
&\hat{\tau}_{22} = \frac{1}{k^2}\left[k_2^2\tilde{\tau}_{11} + k_1^2\tilde{\tau}_{22}\right], \\
&\hat{\tau}_{13} = \frac{k_1\tilde{\tau}_{13}}{k}, \quad \hat{\tau}_{23} = \frac{k_2\tilde{\tau}_{13}}{k}, \quad \hat{\tau}_{33} = \tilde{\tau}_{33},
\end{aligned} \tag{6.8}
$$

and so the inverse Fourier transforms to real space are

$$
\begin{aligned}
&\tau_{11} = \Theta_{2,0}\left[\frac{1}{k^2}\tilde{\tau}_{11}\right] + \Theta_{0,2}\left[\frac{1}{k^2}\tilde{\tau}_{22}\right], \quad \tau_{12} = \Theta_{1,1}\left[\frac{1}{k^2}\left(\tilde{\tau}_{11} - \tilde{\tau}_{22}\right)\right], \\
&\tau_{22} = \Theta_{0,2}\left[\frac{1}{k^2}\tilde{\tau}_{11}\right] + \Theta_{2,0}\left[\frac{1}{k^2}\tilde{\tau}_{22}\right], \quad \tau_{13} = \Theta_{1,0}\left[\frac{\tilde{\tau}_{13}}{k}\right], \\
&\tau_{23} = \Theta_{0,1}\left[\frac{\tilde{\tau}_{13}}{k}\right], \quad \tau_{33} = \Theta_{0,0}\left[\tilde{\tau}_{33}\right].
\end{aligned} \tag{6.9}
$$

These stresses may now be computed in cylindrical coordinates from (5.18) using Hankel transforms of the appropriate tilde variables.

We next consider a source which is a vertical point force acting on the earth's surface. This models hammer, weight drop, and vibroseis (a shaking truck) sources [4]. Thus we take

$$
\mathbf{F} = \mathbf{f} = \left[0, 0, \hat{h}(\omega)\right]^T \delta(x_1)\delta(x_2)\delta(x_3 - z_s), \tag{6.10}
$$

where $\hat{h}(\omega)$ is the spectrum of the force and $z_s \downarrow 0$ puts the force on the earth's surface. By lateral Fourier transform and rotation by Ω we get

$$
\hat{\mathbf{F}} = \hat{\mathbf{f}} = \tilde{\mathbf{F}} = \tilde{\mathbf{f}} = \left[0, 0, \hat{h}(\omega)\right]^T \delta(z - z_s). \tag{6.11}
$$

From (2.16) we obtain

$$
\mathbf{S}^{(1)} = [0, 0, 0, 0, -1, 0, 1, 0]^T \hat{h}(\omega)\delta(z - z_s). \tag{6.12}
$$

From (2.20) we obtain

$$
\mathbf{S}^{(2)} = \mathbf{0}. \tag{6.13}
$$

Therefore all variables in System 2 are zero, as was the case for dynamite.

From (5.1), (6.12), and (5.4) we obtain

$$
{\mathbf{S}_\mathbf{A}}^{(1)} = \mathbf{0}, \quad {\mathbf{S}_\mathbf{B}}^{(1)} = [1, 0, -1, 0]^T \hat{h}(\omega). \tag{6.14}
$$

Now all the tilde variables at the earth's surface may be computed from equation (5.14) as $z_s \downarrow 0$ and propagated anywhere else in space. Note that ${\mathbf{S}_\mathbf{A}}^{(1)}, {\mathbf{S}_\mathbf{B}}^{(1)}$ are independent of k_1, k_2, so that the tilde variables depend only on k and not on wavenumber direction. Therefore, as for dynamite we can transform to the hat variables using (6.6) and (6.8) and transform back to real space using (6.7) and (6.9).

7. Source currents in a plane. We consider a distribution of horizontal source currents in the source plane $z = z_s$. Because of linearity and horizontal translation invariance, we need only consider a horizontal point dipole at $\mathbf{x} = (0, 0, z_s)^T$, with source current

$$\mathbf{j} = \mathbf{d}\ \delta(x_1)\delta(x_2)\delta(x_3 - z_s), \tag{7.1}$$

where

$$\mathbf{d} = (d_1, d_2, 0)^T. \tag{7.2}$$

Solutions for other horizontal sources in the plane $z = z_s$ can be synthesized by translation and superposition of sources of this type. Then by taking Fourier transforms and rotating via (2.5) and (2.8) we obtain

$$\tilde{\mathbf{j}} = \tilde{\mathbf{d}}\ \delta(z - z_s), \tag{7.3}$$

where

$$\begin{aligned} \tilde{d}_1 &= (k_1 d_1 + k_2 d_2)/k, \\ \tilde{d}_2 &= (-k_2 d_1 + k_1 d_2)/k, \\ \tilde{d}_3 &= 0. \end{aligned} \tag{7.4}$$

From (2.16) and (2.20) we obtain

$$\mathbf{S}^{(1)} = \tilde{d}_1 \bar{\mathbf{S}}^{(1)},\ \ \mathbf{S}^{(2)} = \tilde{d}_2 \bar{\mathbf{S}}^{(2)}, \tag{7.5}$$

where

$$\begin{aligned} \bar{\mathbf{S}}^{(1)} &= [0, 0, 0, -1, 0, 0, 0, 0]^T\ \ \delta(z - z_s), \\ \bar{\mathbf{S}}^{(2)} &= [0, 0, 0, -1]^T\ \ \delta(z - z_s). \end{aligned} \tag{7.6}$$

Let $\bar{\Phi}^{(1)}$ be the solution of System 1 with source $\bar{\mathbf{S}}^{(1)}$, i.e., with

$$\bar{\mathbf{S}}^{(1)}_{\mathbf{A}} = [0, 0, 0, 1]^T, \quad \bar{\mathbf{S}}^{(1)}_{\mathbf{B}} = \mathbf{0}, \tag{7.7}$$

and let $\bar{\Phi}^{(2)}$ be the solution of System 2 with source $\bar{\mathbf{S}}^{(2)}$, i.e., with

$$\bar{\mathbf{S}}^{(2)}_{\mathbf{A}} = \mathbf{0}, \quad \bar{\mathbf{S}}^{(2)}_{\mathbf{B}} = [0, 1]^T. \tag{7.8}$$

Then by linearity

$$\Phi^{(1)} = \tilde{d}_1 \bar{\Phi}^{(1)}, \quad \Phi^{(2)} = \tilde{d}_2 \bar{\Phi}^{(2)}. \tag{7.9}$$

Note that $\bar{\Phi}^{(1)}, \bar{\Phi}^{(2)}$ depend on k but not on k_1, k_2 individually. We denote the elements of these vectors, analogous to (2.10), as

$$\begin{aligned} \bar{\Phi}^{(1)} &= \left[\dot{\bar{u}}_3, \bar{\tau}_{13}, -\dot{\bar{w}}_3, \bar{H}_2, \bar{\tau}_{33}, \dot{\bar{u}}_1, \bar{p}, \bar{E}_1\right]^{\mathbf{T}}, \\ \Phi^{(2)} &= \left[\dot{\bar{u}}_2, \bar{E}_2, \bar{\tau}_{23}, -\bar{H}_1\right]^{\mathbf{T}} \end{aligned} \tag{7.10}$$

and define the corresponding auxiliary variables for System 1 with normalized sources analogously to equations (2.17), e.g.,

$$\bar{E}_3 = \beta_2 \left(\frac{ik}{\bar{\sigma}} \bar{H}_2 - \frac{L\eta}{\kappa\bar{\sigma}} \dot{\bar{w}}_3 \right), \tag{7.11}$$

and the auxiliary variables for the normalized System 2 analogously to equations (2.21), e.g.,

$$\bar{H}_3 = \frac{\gamma}{\mu}\bar{E}_2. \tag{7.12}$$

Since $\dot{\tilde{u}}_1, \dot{\tilde{u}}_3$ are in System 1, while $\dot{\tilde{u}}_2$ is in System 2,

$$\dot{\tilde{u}}_1 = \tilde{d}_1\dot{\bar{u}}_1, \quad \dot{\tilde{u}}_2 = \tilde{d}_2\dot{\bar{u}}_2, \quad \dot{\tilde{u}}_3 = \tilde{d}_1\dot{\bar{u}}_3. \tag{7.13}$$

Therefore, using (7.13), (7.4), and (5.15),

$$\begin{aligned}
\dot{\hat{u}}_1 &= (d_1k_1^2 + d_2k_1k_2)\frac{\dot{\bar{u}}_1(k,z)}{k^2} + (d_1k_2^2 - d_2k_1k_2)\frac{\dot{\bar{u}}_2(k,z)}{k^2},\\
\dot{\hat{u}}_2 &= (d_1k_1k_2 + d_2k_2^2)\frac{\dot{\bar{u}}_1(k,z)}{k^2} + (-d_1k_1k_2 + d_2k_1^2)\frac{\dot{\bar{u}}_2(k,z)}{k^2},\\
\dot{\hat{u}}_3 &= (d_1k_1 + d_2k_2)\frac{\dot{\bar{u}}_3(k,z)}{k}.
\end{aligned} \tag{7.14}$$

Inverting the Fourier transforms via (5.16) gives $\dot{\mathbf{u}}$ in real space:

$$\begin{aligned}
\dot{u}_1 &= d_1\Theta_{2,0}\left[\frac{\dot{\bar{u}}_1}{k^2}\right] + d_2\Theta_{1,1}\left[\frac{\dot{\bar{u}}_1}{k^2}\right] + d_1\Theta_{0,2}\left[\frac{\dot{\bar{u}}_2}{k^2}\right] - d_2\Theta_{1,1}\left[\frac{\dot{\bar{u}}_2}{k^2}\right],\\
\dot{u}_2 &= d_1\Theta_{1,1}\left[\frac{\dot{\bar{u}}_1}{k^2}\right] + d_2\Theta_{0,2}\left[\frac{\dot{\bar{u}}_1}{k^2}\right] - d_1\Theta_{1,1}\left[\frac{\dot{\bar{u}}_2}{k^2}\right] + d_2\Theta_{2,0}\left[\frac{\dot{\bar{u}}_2}{k^2}\right],\\
\dot{u}_3 &= d_1\Theta_{1,0}\left[\frac{\dot{\bar{u}}_3}{k}\right] + d_2\Theta_{0,1}\left[\frac{\dot{\bar{u}}_3}{k}\right].
\end{aligned} \tag{7.15}$$

Now $\dot{\mathbf{u}}$ can be written in cylindrical coordinates using the Hankel transform relations (5.18).

Next, note that $\tilde{E}_1$ is in System 1, $\tilde{E}_2$ is in System 2, and $\tilde{E}_3$ is an auxiliary variable in System 1. Therefore

$$\tilde{E}_1 = \tilde{d}_1\bar{E}_1, \quad \tilde{E}_2 = \tilde{d}_2\bar{E}_2, \quad \tilde{E}_3 = \tilde{d}_1\bar{E}_3. \tag{7.16}$$

Comparison of (7.16) with (7.13) shows that $\mathbf{E}$ may be obtained in real space with (7.15) by replacing $\dot{\mathbf{u}}$ with $\mathbf{E}$ and $\dot{\bar{\mathbf{u}}}$ with $\bar{\mathbf{E}}$.

For the magnetic field, note that $\tilde{H}_1, \tilde{H}_3$ are in System 2, while $\tilde{H}_2$ is in System 1. Therefore

$$\tilde{H}_1 = \tilde{d}_2\bar{H}_1, \quad \tilde{H}_2 = \tilde{d}_1\bar{H}_2, \quad \tilde{H}_3 = \tilde{d}_2\bar{H}_3. \tag{7.17}$$

Similar to the procedure for $\dot{\mathbf{u}}$, we first use (7.17) to write $\hat{\mathbf{H}}$ and then invert the Fourier transform using (5.16) to get $\mathbf{H}$ in real space:

$$\begin{aligned}
H_1 &= -d_1\Theta_{1,1}\left[\frac{\bar{H}_1}{k^2}\right] + d_2\Theta_{2,0}\left[\frac{\bar{H}_1}{k^2}\right] - d_1\Theta_{1,1}\left[\frac{\bar{H}_2}{k^2}\right] - d_2\Theta_{0,2}\left[\frac{\bar{H}_2}{k^2}\right],\\
H_2 &= -d_1\Theta_{0,2}\left[\frac{\bar{H}_1}{k^2}\right] + d_2\Theta_{1,1}\left[\frac{\bar{H}_1}{k^2}\right] + d_1\Theta_{2,0}\left[\frac{\bar{H}_2}{k^2}\right] + d_2\Theta_{1,1}\left[\frac{\bar{H}_2}{k^2}\right],\\
H_3 &= -d_1\Theta_{0,1}\left[\frac{\bar{H}_3}{k}\right] + d_2\Theta_{1,0}\left[\frac{\bar{H}_3}{k}\right].
\end{aligned} \tag{7.18}$$

Again $\mathbf{H}$ can be written in cylindrical coordinates using the Hankel transform relations (5.18).

A similar treatment can be given for the other variables. However, we will focus on $\dot{u}_3$, i.e., the vertical velocity of the ground, since at the surface $z = 0$ this is the response of a conventional geophone. So the behavior of $\dot{u}_3$ gives the seismic response in an ES land survey. From (7.15) and (5.18) the seismic response for a horizontal dipole is given in cylindrical coordinates by the order 1 Hankel transform

$$\dot{u}_3(r,\theta,z) = i(d_1\cos\theta + d_2\sin\theta)\mathcal{B}_{1,1}[\dot{u}_3(k,z)]. \tag{7.19}$$

Alternatively, from (7.15), (5.16), and (5.18),

$$\dot{u}_3(r,\theta,z) = -\mathbf{d}\cdot\nabla R(r,z), \tag{7.20}$$

where the response function R is the order 0 Hankel transform

$$R(r,z) = i\mathcal{B}_{0,0}[\dot{u}_3(k,z)]. \tag{7.21}$$

Note that for the vertical component of the geophone response, only $\dot{u}_3$ need be computed, so it is not necessary to solve System 2.

The preceding results are for a point dipole at $\mathbf{x} = (0,0,z_s)^T$. Now consider a source which is a current sheet in a horizontal plane, so that

$$\begin{aligned} \mathbf{j} &= \mathbf{C}(x_1,x_2)\,\delta(x_3 - z_s), \\ \mathbf{C} &= [C_1, C_2, 0]^T. \end{aligned} \tag{7.22}$$

The vertical velocity for this source may be computed by superposition,

(7.23)

$$\begin{aligned} \dot{u}_3(r,\theta,z) &= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{C}(x_1',x_2')\cdot\nabla R\Big(\sqrt{(x_1-x_1')^2+(x_2-x_2')^2}, z\Big)\,dx_1'dx_2' \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (-\nabla'\cdot\mathbf{C}(x_1',x_2'))\,R\Big(\sqrt{(x_1-x_1')^2+(x_2-x_2')^2}, z\Big)\,dx_1'dx_2', \end{aligned}$$

where the second expression comes from the divergence theorem, assuming that $\mathbf{C}$ is smooth and vanishes sufficiently rapidly at ∞. Therefore the geophone response is a superposition of cylindrically symmetric scalar response functions R weighted at each point by the current leakage $-\nabla\cdot\mathbf{C}$. In particular, only the leakage of current from the source plane contributes to the geophone response.

We may obtain a related result for the response to a leaky wire in the plane $z = z_s$, when the wire follows the path $(\bar{x}_1(\alpha), \bar{x}_2(\alpha), z_s)$, where α is arclength along the path and $0 \le \alpha \le l$. If $I(\alpha)$ is the current in the wire at position α, then the source current is

$$\mathbf{j} = \int_0^l I(\alpha) \begin{bmatrix} \frac{d\bar{x}_1}{d\alpha} \\ \frac{d\bar{x}_2}{d\alpha} \\ 0 \end{bmatrix} \delta\left(x_1 - \bar{x}_1(\alpha)\right)\ \ \delta\left(x_2 - \bar{x}_2(\alpha)\right)\delta\left(x_3 - z_s\right)d\alpha. \tag{7.24}$$

Let

$$\bar{r}(\alpha, x_1, x_2) = \sqrt{(x_1 - \bar{x}_1(\alpha))^2 + (x_2 - \bar{x}_2(\alpha))^2}. \tag{7.25}$$

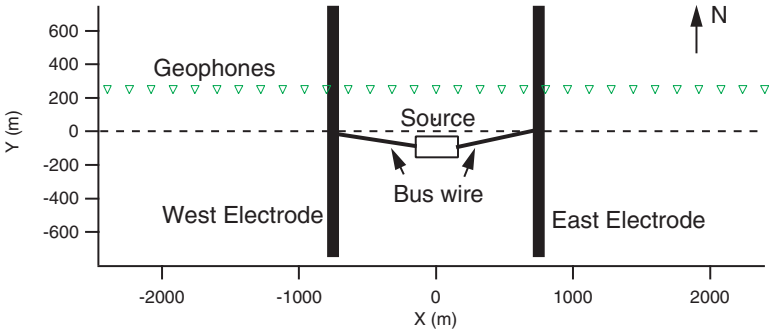


FIG. 1. *Configuration of electrodes and geophones.*

Then superposition of response functions gives the expression

(7.26)

$$\dot{u}_3 = -\int_0^l I(\alpha) \begin{bmatrix} \frac{d\bar{x}_1}{d\alpha} \\ \frac{d\bar{x}_2}{d\alpha} \\ 0 \end{bmatrix} \cdot \nabla R\left(\bar{r}(\alpha, x_1, x_2), z\right) d\alpha$$

$$= I(l)R(\bar{r}(l, x_1, x_2), z) - I(0)R(\bar{r}(0, x_1, x_2), z) + \int_0^l \left(-\frac{dI}{d\alpha}\right) R(\bar{r}(\alpha, x_1, x_2), z) d\alpha.$$

Again, only the leakage, both at the end points and along the length of the wire, contributes to the geophone response.

In particular, for an insulated wire I is constant. Then we have only the end-point contributions, and the path of the wire between its end points is immaterial. If the insulated wire is a closed loop, then the end points cancel and the geophone response vanishes. This fact was derived by Haartsen and Pride [7] for the special case of a circular current loop.

8. Electroseismic prospecting. A computer code based on the theory of this paper was written and used for planning and interpretation of several field tests [21, 20, 18, 10, 9, 8]. In these tests, an EM source was used, and geophones on the surface of the earth recorded the vertical velocity of the ground.

Figure 1 describes the "railroad track" electrode design used for these tests. The electrodes are two parallel transmission lines buried in the shallow subsurface. The length and separation of the electrodes are comparable to the desired depth of investigation, which for the deeper tests is on the order of a kilometer. A power source is located midway between the two tracks, and insulated bus wires feed current from the power source to the centers of the two electrodes.

From the center of the west electrode, current from the bus wire flows outward toward the electrode ends, leaking into the ground all along the electrode's length. Because of this leakage the current carried by the electrode decreases linearly from a maximum at the bus wire feed point to zero at each end. The current behavior in the east electrode is similar to that described for the west electrode, but with sign reversed.

A number of factors influenced this electrode design. Horizontal sources were chosen because they allow a large scale geometry that would be impractical with vertical structures. However, as shown in section 7, current leakage is essential for horizontal sources if we are to obtain a seismic response from a conventional geophone.

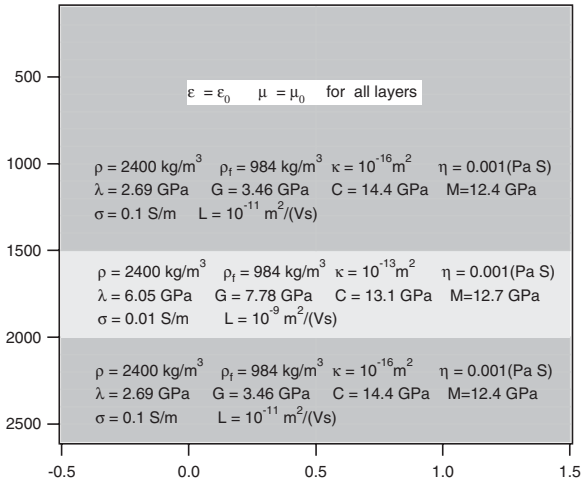


FIG. 2. *Three-layer model.*

For example, as shown in section 7, a source consisting of an insulated wire loop, as is common in EM prospecting [14, 25], would give no seismic response at all in a layered earth. Furthermore, maximizing the current leakage is essential for enhancing the magnitude of the seismic signal. Also, because Pride's equations are linear, the seismic response for any electrode configuration is directly proportional to the source current; so the greatest practically obtainable current level is desired in order to produce the greatest possible seismic response.

To model the ES response to these electrodes over a layered earth, we first obtain for the specified earth model the horizontal dipole response function $R(r, 0)$ on the surface of the earth $z = 0$ using (7.21). The electrodes and bus wires are described mathematically as segments of curves in the plane, carrying varying amounts of current at each point along the length. That is, I is constant in the bus wires, and $dI/d\alpha$ is constant along each half of each electrode. Then the second expression in (7.26) is used to calculate the geophone response. Note that in this calculation, the path of the bus wires is immaterial, because they are insulated.

In modeling the current in the electrodes as linearly decreasing from the bus wire feed to the end points, we have neglected phase changes in $I(\alpha)$ caused by the inductance of the earth. Calculation of the inductance is beyond the scope of the present paper.

Figure 2 shows a layered earth model with three layers, where the free space ϵ_0, μ_0 are assumed in all layers. Figure 3 shows a typical trace, as recorded at a geophone at coordinates $x = 1000\,\mathrm{m}$, $y = 250\,\mathrm{m}$. In transforming the computer calculations to the time domain, a 15 Hz Ricker wavelet pulse form was used. In the actual field tests pulses are not used; instead a periodic source function is generated and repeated many times, after which the received seismic signal is averaged over a period [8, 21, 20, 10, 9]. This methodology allows a great amount of EM energy to be put into the ground to enhance the received seismic response. The resulting data can then be Fourier transformed to the frequency domain and a pulse wavelet inserted before inverse transforming back to the time domain. In this way a trace of the type of Figure 3, analogous to the geophone responses used in conventional seismic interpretation, can be constructed.

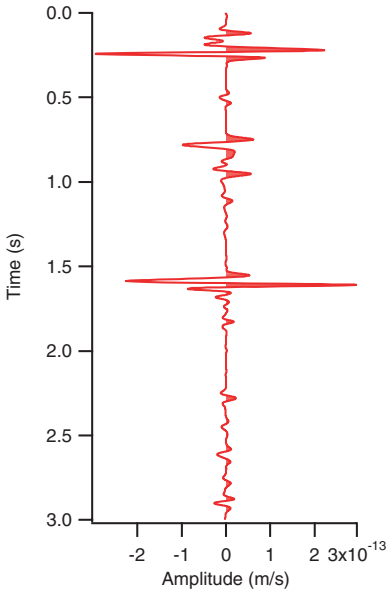
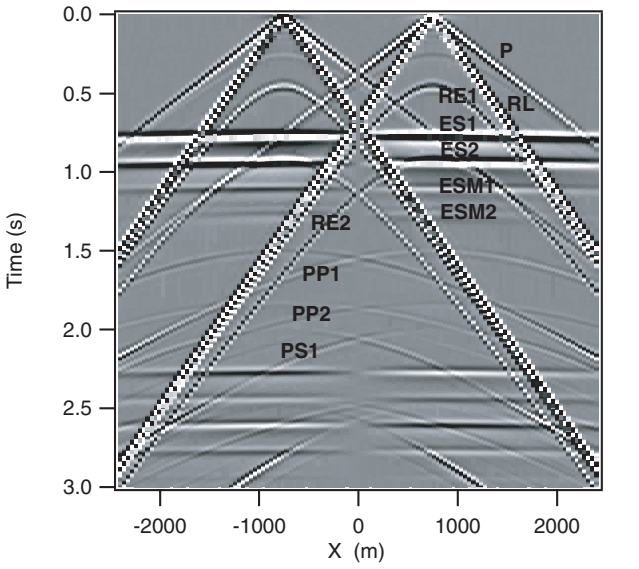


FIG. 3. *Trace at* $x = 1000$ m.

Note that the scale in Figure 3 predicts that very small signals will be received. These signals are within the sensitivity of a geophone, but smaller than many noise sources that can be identified in the field, so considerable signal averaging will be necessary for detection to be at all feasible. In the computation, we have assumed a current leakage of 1 amp/m of electrode length, for a total leakage of 1500 amps. All the results scale linearly with the current, so the small signals predicted can be enhanced by increasing the current.

Figure 4 shows the resulting electroseismogram, analogous to a conventional seismic gather, for a line of geophones orthogonal to the two electrodes and offset 250 m from the center. We have labeled a number of the events in Figure 4. Event ES1 results from energy that has been converted from EM to seismic at the first interface and has subsequently propagated upward to the surface. Note that this event is flat, that is, the arrival times are the same at all the receivers. This behavior is in contrast to that of a conventional seismic reflection, which has "moveout" caused by the geometry of the ray paths, i.e., a hyperbolic dependence of arrival time with offset of the receiver from the source. In contrast, the asymptotic theory of ES conversion [24] shows that the seismic rays resulting from an incident EM wave leave the interface normal to it. For a layered earth, this means that all the converted seismic rays travel vertically. Since the propagation of EM energy is virtually instantaneous on the seismic time scale, the converted seismic signals propagating vertically from each point on the interface arrive at the surface simultaneously. Another way to understand this phenomenon intuitively is to note that the EM wave speed is much greater than the seismic wave speed. Therefore Snell's law predicts that the converted seismic waves leave the interface in approximately the normal direction.

Note also that the amplitudes of event ES1 go to zero at a point midway between the electrodes, with a sign reversal as this midpoint is crossed. This is a result of symmetry of the problem. The asymptotic theory [24] shows that the P-wave con-



P : Direct arrival of P wave
RL : Rayleigh wave from east line electrode
RE1 : Rayleigh wave from north end of east electrode
RE2 : Rayleigh wave from south end of east electrode
ES1 : ES conversion at first interface
ES2 : ES conversion at second interface
ESM1 : Mutiple of ES conversion at first interface
ESM2 : Mutiple of ES conversion at second interface
PP1 : P wave reflection at first interface
PP2 : P wave reflection at second interface
PS1 : PS conversion or/and SP conversion at first interface

FIG. 4. *Electroseismogram* (*three-layer model*).

version is dependent on the electric field normal to the interface, which for a layered earth is in the vertical direction. From symmetries of the electrode, the vertical **E** field changes sign on a line midway between the electrodes, and so the ES conversion to P-waves changes sign there also.

Moreover, only the P-wave response of the energy converted at an interface contributes to the vertical velocity of the ground. This is because the converted S-waves, which likewise travel vertically, have a particle motion orthogonal to their direction of travel, i.e., they contribute a purely horizontal component to the solid velocity. Furthermore the Biot slow wave is undetectable at the surface. Although Biot slow waves are generated at the interfaces by the EM field, their rapid decay (typically on a scale of inches!) makes their amplitudes transcendentally small at the surface.

The amplitude of ES1 peaks at a position 160 m from each electrode, outside of the area bounded by the two electrodes. This is a general property of conversions from a single interface, and is used in planning the placement of the geophones. Typically geophones are placed to cover an area where significant signals are expected. The geophone line shown here is for illustrative purposes only.

Returning to Figure 4, other flat events, e.g., events ES2, ESM1, and ESM2, can be identified by comparing their times of arrival with the P-wave velocities in the

layers. Event ES2 results from the primary conversion to seismic propagating upward from the second interface. Events ESM1 and ESM2 are multiples: Event ESM1 is energy converted to seismic at the first interface, directed downward where it is reflected upward as seismic at the second interface and finally received at the geophones. Event ESM2 is energy converted at the second interface, with seismic reverberation between the first and second interfaces before being received at the surface. As is usual, amplitudes of the multiples are considerably less than those of the primaries.

Other events, which do not have simultaneous arrival times, result from conversion of EM to seismic energy directly where the electrode contacts the ground. This produces a small, order of L, seismic survey along with the ES survey. Of course the ES conversions at the interfaces, which are the signals we seek, are also of order L. So conversion at the electrodes cannot be ignored.

Event P is the direct arrival of the P-wave generated at the electrodes, traveling just below the surface. Events PP1 and PP2 are reflections of the P-wave generated at the electrodes from the first and second interfaces, respectively. Event PS1 is a combination of P to S and S to P conversions at the second interface.

Event RL is a Rayleigh wave, or "ground roll" traveling along the surface and generated at all points of the east electrode. It is identified by the Rayleigh wave speed and linear dependence of arrival time with distance from the source. In the layer code, the Rayleigh wave manifests itself as a pole in k space, and care must be taken in numerical integration of the Hankel transforms as the path of integration nears this pole. Since the pole is near the real axis, numerical stability is enhanced by giving the frequency a small positive imaginary part, that is, replacing $\omega \to \omega + i\delta$. The effects of complex frequency can then be removed in the time domain by multiplication by a factor of $e^{\delta t}$. Also, to enhance numerical stability, the wavenumber can be given a small negative imaginary part.

Event RL is approximately what would be obtained if the electrode were an infinite line source. However, because the actual electrode is of finite length, there are end-point contributions RE1 and RE2 from the north and south ends of the east electrode, and these manifest as separate events.

Many more events can be identified in Figure 4, and interpretation becomes increasingly complicated as the number of layers increases to model realistic exploration geometries.

The conductivity in layer 2 is representative of that of an oil reservoir. In Figure 5 we compare this model with that of a similar model where layer 2 conductivity is an order of magnitude higher, indicating that there is no oil. Shown is the amplitude versus offset for the ES conversion at the first interface. As expected, the presence of oil enhances the signal considerably. In contrast, changing the permeability by an order of magnitude has virtually no effect, as is shown in Figure 6. These results are qualitatively consistent with conclusions based on the asymptotic theory [24].

In Figure 7 we show that in some carefully chosen cases, the presence of very small layers can have a dramatic effect on the amplitudes of the received signals. We compare the model of Figure 2 with a similar model that has a 10 cm gas sand inserted just above the reservoir, i.e., between layers 1 and 2. The result is that the ES conversion is more than doubled.

This effect can be understood qualitatively as arising from the much greater compressibility of a gas compared to a liquid; the enhanced compressibility provided by even a small layer of gas allows for much more movement of the solid matrix. Mathematically, a small layer can have an effect only if the exponential matrices $e^{i\omega\Lambda_m\Delta z_m}$ that occur in (4.21) differ substantially from the identity matrix. Of course these

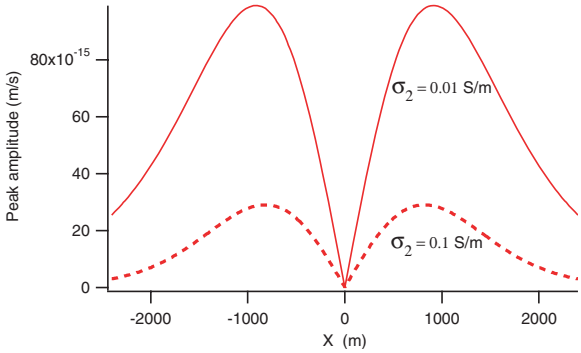


FIG. 5. *Conductivity dependence of ES conversion.*

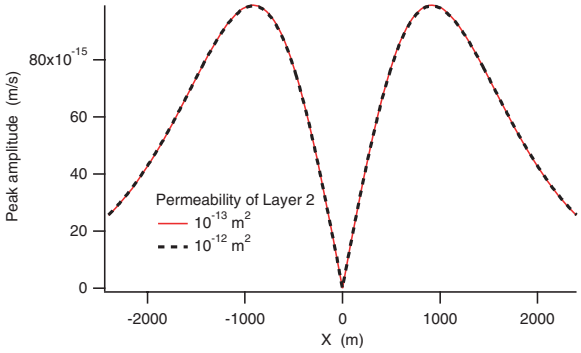


FIG. 6. *Permeability dependence of ES conversion.*

matrices must approach the identity as the layer thickness Δz_m becomes sufficiently small. However, the Biot slow wave has a very large slowness, as is discussed asymptotically in [24]. The complex Biot slow wave slowness is one of the eigenvalues of Λ_m, namely, q_2 in (A.1), and has been computed for the parameters of Figure 2 (and horizontal slowness $\gamma = 0$) as $q_2 = .348(1 + i)$. Therefore the matrices in question have a second diagonal element of $e^{3.28(-1+i)}$, which differs substantially from unity. In general it is the potentially large slowness of the Biot slow wave, i.e., its short wavelength, that allows the possibility that a small layer might have an order one effect.

A small layer can also have a large effect by substantially changing the local electric field, as is illustrated in Figure 8. For the base case in this figure, we consider the same model parameters as in Figure 2, except that the reservoir thickness is reduced to 10 m. The result of this model may be compared with case 1, in which we insert a thin, 2 cm, highly resistive layer just below the top of the reservoir, with a varying conductivity profile as shown in Figure 9. This continuously variable conductivity profile is simulated by discretizing it over 200 layers, each with a thickness of .1 mm. The result is a dramatic rise of the amplitude of the ES conversion by more than an order of magnitude, as compared with the base case. The much higher ES conversion may also be achieved with a constant resistivity layer at the top of the reservoir, as is illustrated in case 2. For this case, the highly resistive thin layer has a constant resistivity of $2500\,\Omega - m$ and a thickness of 5 cm.

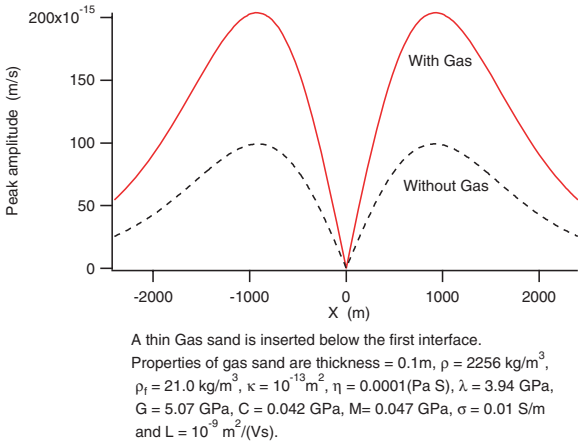


FIG. 7. *ES conversion enhancement due to gas presence.*

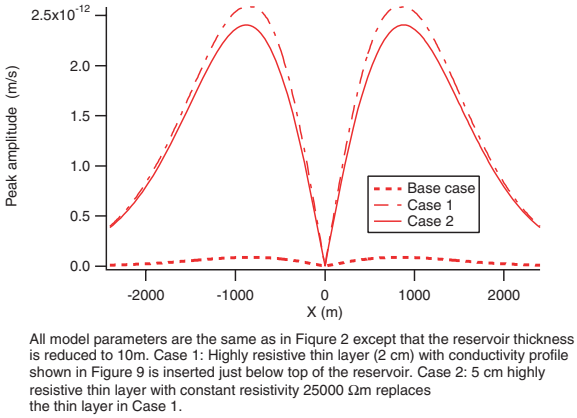


FIG. 8. *Effect of a high resistivity thin layer on ES conversion.*

The normal electric field for case 2 of Figure 8 is plotted in Figure 10. It is relatively constant over the thin resistive layer, with a discontinuity at the interface proportional to the ratio of resistivities, as is inherent in Maxwell's equations. In the asymptotic theory [24] the size of the discontinuity of the normal electric field is a major factor in the magnitude of the ES conversion at an interface. Although the layer in this example is too small for the strict validity of the asymptotic theory, the qualitative dependence of ES conversion on the normal electric field is seen to be similar here.

In Figure 11 we demonstrate how fine structure in a reservoir can also raise the amplitude of the ES conversion. In this figure, we consider a 20 m thick reservoir with a periodic structure of 67 gas-oil-shale layers, each of thickness 1 cm. The result is a substantial increase in converted wave amplitude, as compared with the case of a homogeneous reservoir.

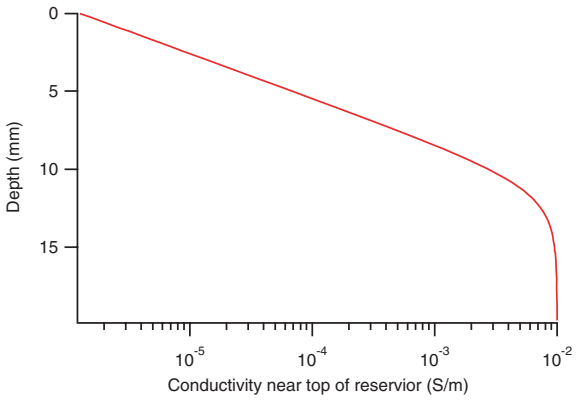


FIG. 9. *Profile of conductivity near top of reservoir.*

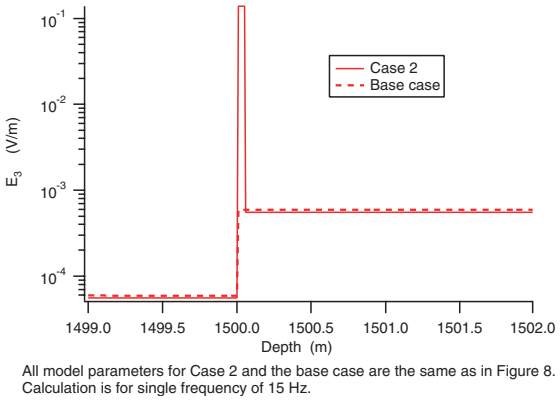


FIG. 10. *Local normal E field near a highly resistive thin layer.*

9. Conclusions. We have shown how the equations of electrokinetics can be put into Ursin's form in a plane-layered medium. Using this form we have derived explicit formulas that can be used as the basis of an efficient layer code, and have shown numerical results for spatially extended electrode sources that have been used in field tests.

More generally, the methods developed are applicable to any system that can be put into Ursin's form. In particular, the code that was written for ES waves can be modified to separately compute seismic waves, EM waves, or the waves of Biot theory, which are all included in Pride's equations: For seismic waves, isotropic elasticity is recovered from the third and fifth equations of (2.1) by simply taking $\rho_f = 0, C = 0$; then $\mathbf{E}, \mathbf{H}, \mathbf{w}, p$ vanish, so their components may be deleted to reduce the dimensionality of the systems. Similarly, in the limit $L \to 0$ we recover either Maxwell's equations or Biot's equations separately.

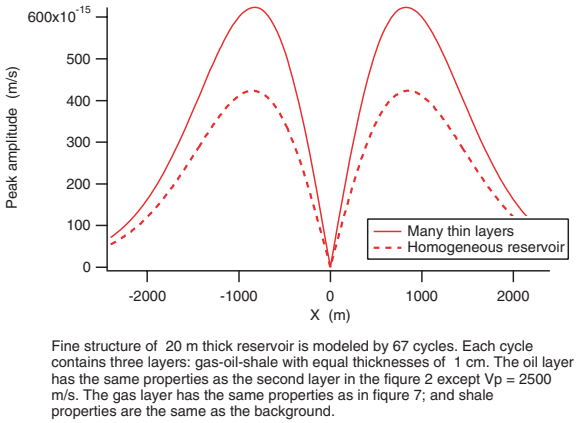


FIG. 11. *Effect of fine structure on ES conversion.*

As expected, the higher electrical resistivity of hydrocarbons does considerably enhance the ES waves converted at reservoir boundaries. For shallow reservoirs, ES signals should be detectable. However, our calculations show that there are significant technological challenges to make ES a reliable tool for detecting deep oil reservoirs because the received signals from depth are likely to be small unless a great amount of EM energy can be put into the ground. Effective signal processing techniques will be necessary to separate the signals from noise. Also, desired signals originating from deep interfaces may be obscured by conversions from shallow interfaces because the EM field is likely to be large at shallow depths where it has not attenuated much.

In comparison to purely EM prospecting, the EM waves used in ES can be of lower frequency to give the same wavelength of the returned (i.e., seismic) signal. Also, in ES prospecting the EM waves only go one way, i.e., down to the reservoir, as opposed to the round-trip taken by the EM waves in an EM survey. Since EM waves attenuate with propagation distance, and with shortened wavelength, these two effects favor ES. However, these effects are counterbalanced by the fact that in ES only a small fraction of the EM energy is converted to seismic at an interface, giving a corresponding reduction in the amplitude of the ES signal.

The ES conversions at interfaces are generally consistent with the asymptotic theory of [24], which gives three-dimensional effects not included in layered earth modeling. However, the layer code predicts a surprising result that is not computable with the asymptotic theory: the possibility that layers very much smaller than a seismic wavelength can have a large effect on the amplitude of the received ES signal. This was demonstrated by the model of a thin gas sand overlying an oil reservoir, by the model of a thin but highly resistive layer, and by a cyclical model of reservoir fine structure. The possibility of signal enhancement through this mechanism must be considered in analyzing the amplitudes of electroseismograms.

Another challenge is the interpretation of the many different types of events seen in electroseismograms. For example, we have shown that ES conversions where the electrode contacts the ground produce a multitude of signals; these signals are a sort of source-generated noise, which can obscure the returns from EM waves converted to seismic at reservoir boundaries. Modeling of the type we have shown here is essential for proper interpretation and classification of all these signals.

Appendix A. Eigenvectors for System 1. See also [17, 7]. The four modes for system 1 are

- $m = 1$: fast compressional wave (P-wave);
- $m = 2$: Biot slow wave;
- $m = 3$: vertical shear wave (SV-wave);
- $m = 4$: transverse magnetic wave (TM-wave).

$m = 1, 2$ are longitudinal waves. For these modes

(A.1)
$$q_m^2 = -\gamma^2 + \beta_1\Bigg(C\rho_f - \frac{1}{2}M\rho - \frac{i}{2}(\lambda+2G)\left(\frac{\eta}{\kappa\omega}\right)\beta_2 \pm\frac{1}{2}\sqrt{\left(i(\lambda+2G)\left(\frac{\eta}{\kappa\omega}\right)\beta_2 - M\rho\right)^2 - 4\left(M\rho_f - i\left(\frac{\eta}{\omega\kappa}\right)C\beta_2\right)\left(C\rho - (\lambda+2G)\rho_f\right)}\Bigg), \quad m = 1, 2.$$

In (A.1) the plus sign is for $m = 1$ (P-waves) and the minus sign is for $m = 2$ (Biot slow waves). Then for $m = 1, 2$

$$\text{(A.2)}\qquad \mathbf{a}_m = \bar{a}_m\begin{bmatrix} -1 \\ 2\gamma G \\ \xi_m \\ 0 \end{bmatrix}, \quad \mathbf{b}_m = \frac{\bar{a}_m}{q_m}\begin{bmatrix} 2\gamma^2 G - \rho - \rho_f\xi_m \\ \gamma \\ \rho_f + i\left(\frac{\eta}{\omega\kappa}\right)\beta_2\xi_m \\ -\gamma\left(\frac{\eta L}{\kappa\bar{\sigma}}\right)\beta_2\xi_m \end{bmatrix}, \quad m = 1, 2,$$

where

(A.3)
$$\xi_m = \frac{\left(C\rho_f - M\rho - \frac{\left(q_m^2+\gamma^2\right)}{\beta_1}\right)}{\left(M\rho_f - i\left(\frac{\eta}{\omega\kappa}\right)C\beta_2\right)} = \frac{(C\rho - (\lambda+2G)\rho_f)}{\left(\frac{\left(q_m^2+\gamma^2\right)}{\beta_1} - C\rho_f + i(\lambda+2G)\left(\frac{\eta}{\omega\kappa}\right)\beta_2\right)}, \quad m = 1, 2,$$

and

$$\text{(A.4)}\qquad \bar{a}_m = \sqrt{\frac{q_m}{\rho + 2\rho_f\xi_m + i\left(\frac{\eta}{\omega\kappa}\right)\beta_2\xi_m^2}}, \quad m = 1, 2.$$

$m = 3, 4$ are transverse waves. For these modes

$$\text{(A.5)}\qquad q_m^2 = -\gamma^2 + \frac{1}{2G}\left(\rho + i\rho_f^2\left(\frac{\omega\kappa}{\eta}\right) + i\frac{\bar{\sigma}\mu G}{\omega}\right) \mp \frac{1}{2G}\sqrt{\left(i\frac{\bar{\sigma}\mu G}{\omega} - \rho - i\rho_f^2\left(\frac{\omega\kappa}{\eta}\right)\right)^2 - 4L^2\rho_f^2\mu G}, \quad m = 3, 4.$$

In (A.5) the minus sign is for $m = 3$ (SV-waves) and the plus sign is for $m = 4$ (TM-waves). Then

$$\text{(A.6)}\qquad \mathbf{a}_m = -\frac{\bar{b}_m}{q_m}\begin{bmatrix} -\gamma \\ G(\gamma^2 - q^2) \\ i\gamma\rho_f\left(\frac{\omega\kappa}{\eta}\right) + \gamma L\xi_m \\ -L\rho_f + i\frac{\bar{\sigma}}{\omega}\xi_m \end{bmatrix}, \quad \mathbf{b}_m = \bar{b}_m\begin{bmatrix} 2\gamma G \\ 1 \\ 0 \\ \xi_m \end{bmatrix}, \quad m = 3, 4,$$

where

$$\xi_m = \frac{\rho_f L\mu}{\left(\frac{i\bar{\sigma}\mu}{\omega}\right) - q_m^2 - \gamma^2} = \frac{G}{\rho_f L}\left(q_m^2 + \gamma^2 - \frac{\rho}{G} - i\left(\frac{\rho_f^2}{G}\right)\left(\frac{\omega\kappa}{\eta}\right)\right), \quad m = 3, 4, \tag{A.7}$$

and

$$\bar{b}_m = \sqrt{\frac{q_m}{G\left(q_m^2 + \gamma^2\right) + \rho_f L\xi_m - i\left(\frac{\bar{\sigma}}{\omega}\right)\xi_m^2}}, \quad m = 3, 4. \tag{A.8}$$

Appendix B. Eigenvectors for System 2. See also [17, 7]. The two modes for system 1 are

- $m = 1$: horizontal shear wave (SH-wave);
- $m = 2$: transverse electric wave (TE-wave).

The eigenvalues are the same as for the transverse modes of System 1, i.e.,

$$\begin{aligned} q_m^2 = &-\gamma^2 + \frac{1}{2G}\left(\rho + i\rho_f^2\left(\frac{\omega\kappa}{\eta}\right) + i\frac{\bar{\sigma}\mu G}{\omega}\right) \\ &\mp \frac{1}{2G}\sqrt{\left(i\frac{\bar{\sigma}\mu G}{\omega} - \rho - i\rho_f^2\left(\frac{\omega\kappa}{\eta}\right)\right)^2 - 4L^2\rho_f^2\mu G}, \quad m = 1, 2. \end{aligned} \tag{B.1}$$

In (B.1) the minus sign is for $m = 1$ (SH-waves) and the plus sign is for $m = 2$ (TE-waves). The eigenvectors are of the form

$$\mathbf{a}_m = \bar{a}_m \begin{bmatrix} 1 \\ \xi_m \end{bmatrix}, \quad \mathbf{b}_m = \frac{\bar{a}_m}{q_m}\begin{bmatrix} Gq_m^2 \\ \rho_f L + \xi_m\left(\frac{\gamma^2}{\mu} - \frac{i\bar{\sigma}}{\omega}\right) \end{bmatrix}, \quad m = 1, 2, \tag{B.2}$$

where

$$\begin{aligned} \xi_m = &\frac{G}{2\rho_f L}\left[\left(\frac{i\bar{\sigma}\mu}{\omega}\right) - \left(\frac{\rho}{G} + i\omega\frac{\rho_f^2\kappa}{G\eta}\right)\right] \\ &\mp \frac{G}{2\rho_f L}\sqrt{\left[\left(\frac{i\bar{\sigma}\mu}{\omega}\right) - \left(\frac{\rho}{G} + i\omega\frac{\rho_f^2\kappa}{G\eta}\right)\right]^2 - 4\frac{\mu\rho_f^2 L^2}{G}}, \quad m = 1, 2, \end{aligned} \tag{B.3}$$

and

$$\bar{a}_m = \sqrt{\frac{q_m}{Gq_m^2 + \rho_f L\xi_m + \xi_m^2\left(\frac{\gamma^2}{\mu} - i\frac{\bar{\sigma}}{\omega}\right)}}. \tag{B.4}$$

Again, in (B.3) the minus sign is for $m = 1$ (SH-waves) and the plus sign is for $m = 2$ (TE-waves).

Acknowledgment. We thank Max Deffenbaugh for his tests of the computer code.

REFERENCES

[1] M. A. Biot, *Theory of propagation of elastic waves in a fluid-saturated solid* I. *Low frequency range*, J. Acoust. Soc. Amer., 28 (1956), pp. 168–178.

[2] M. A. Biot, *Theory of propagation of elastic waves in a fluid-saturated solid* II. *Higher frequency range*, J. Acoust. Soc. Amer., 28 (1956), pp. 179–191.
[3] K. E. Butler, R. D. Russell, A. W. Kepic, and M. Maxwell, *Measurement of the seismoelectric response from a shallow boundary*, Geophys., 61 (1996), pp. 1769–1778.
[4] M. B. Dobrin, *Introduction to Geophysical Prospecting*, 4th ed., McGraw-Hill, New York, 1988.
[5] S. Garambois and M. Dietrich, *Seismoelectric wave conversions in porous media: Field measurements and transfer function analysis*, Geophys., 66 (2001), pp. 1417–1430.
[6] S. Garambois and M. Dietrich, *Full waveform numerical simulations of seismoelectromagnetic wave conversions in fluid saturated stratified porous media*, J. Geophys. Res., 107 (2002), pp. 40–58.
[7] M. W. Haartsen and S. R. Pride, *Electroseismic waves from point sources in layered media*, J. Geophys. Res., 102 (1997), pp. 24745–24796.
[8] S. C. Hornbostel and A. H. Thompson, *Source Waveforms for Electroseismic Exploration*, U.S. Patent 6,477,113 B2, issued Nov. 5, 2002.
[9] S. C. Hornbostel and A. H. Thompson, *Waveform design for electroseismic exploration*, 75th Annual Meeting of the Society of Exploration Geophysicists, Expanded Abstracts, Tulsa, OK, 2005.
[10] S. C. Hornbostel and A. H. Thompson, *Waveform design for electroseismic exploration*, Geophys., submitted.
[11] B. L. N. Kennett and N. J. Kerry, *Seismic waves in a stratified half space*, Geophys. J. R. Astron. Soc., 57 (1979), pp. 557–583.
[12] O. V. Mikhailov, M. W. Haartsen, and N. Toksoz, *Electroseismic investigation of the shallow subsurface: Field measurements and numerical modeling*, Geophys., 62 (1997), pp. 97–105.
[13] O. V. Mikhailov, J. Queen, and N. Toksoz, *Using borehole electroseismic measurements to detect and characterize fractured (permeable) zones*, Geophys., 65 (2000), pp. 1098–1112.
[14] M. N. Nabighian, ed., *Electromagnetic Methods in Applied Geophysics*, Vols. 1,2, Investigations in Geophysics 3, Society of Exploration Geophysicists, Tulsa, OK, 1987.
[15] S. R. Pride, *Governing equations for the coupled electromagnetics and acoustics of porous media*, Phys. Rev. B, 50 (1994), pp. 15678–15696.
[16] S. R. Pride, A. F. Gangi, and F. D. Morgan, *Deriving the equations of motion for porous isotropic media*, J. Acoust. Soc. Amer., 6 (1992), pp. 3278–3290.
[17] S. R. Pride and M. W. Haartsen, *Electroseismic wave properties*, J. Acoust. Soc. Amer., 100 (1996), pp. 1301–1315.
[18] A. H. Thompson, *Electromagnetic-to-seismic conversion: Successful developments suggest viable applications in exploration and production*, 75th Annual Meeting of the Society of Exploration Geophysicists, Expanded Abstracts, Tulsa, OK, 2005.
[19] A. H. Thompson and G. A. Gist, *Geophysical applications of electrokinetic conversion*, Leading Edge, 1993, pp. 1169–1173.
[20] A. H. Thompson, S. C. Hornbostel, J. S. Burns, T. J. Murray, R. A. Raschke, J. C. Wride, P. Z. McCammon, J. R. Sumner, G. H. Haake, M. S. Bixby, W. S. Ross, B. S. White, M. Zhou, and P. K. Peczak, *Field tests of electroseismic hydrocarbon detection*, 75th Annual Meeting of the Society of Exploration Geophysicists, Expanded Abstracts, Tulsa, OK, 2005.
[21] A. H. Thompson, S. C. Hornbostel, J. S. Burns, T. J. Murray, R. A. Raschke, J. C. Wride, P. Z. McCammon, J. R. Sumner, G. H. Haake, M. S. Bixby, W. S. Ross, B. S. White, M. Zhou, and P. K. Peczak, *Field tests of electroseismic hydrocarbon detection*, Geophys., in press.
[22] R. R. Thompson, *The seismic electric effect*, Geophys., 1 (1936), pp. 327–335.
[23] B. Ursin, *Review of elastic and electromagnetic wave propagation in horizontally layered media*, Geophys., 48 (1983), pp. 1063–1081.
[24] B. S. White, *Asymptotic theory of electroseismic prospecting*, SIAM J. App. Math, 65 (2005), pp. 1443–1462.
[25] M. S. Zhdanov and G. V. Keller, *The Geoelectrical Methods in Geophysical Exploration*, Methods in Geochemistry and Geophysics 31, Elsevier, New York, 1994.

SIAM J. APPL. MATH.
Vol. 67, No. 1, pp. 99–115

A QUEUE-LENGTH CUTOFF MODEL FOR A PREEMPTIVE TWO-PRIORITY $M/M/1$ SYSTEM*

QIANG GONG† AND RAJAN BATTA‡

Abstract. We consider a two-priority, preemptive, single-server queueing model. Each customer is classified into either a high-priority class or a low-priority class. The arrivals of the two-priority classes follow independent Poisson processes, and service time is assumed to be exponentially distributed. A queue-length cutoff method is considered. Under this discipline the server responds only to high-priority customers until the queue length of the other class exceeds a threshold L. After that the server switches to handle only the low-priority queue. Steady-state balance equations are established for this system. Then we introduce two-dimensional generating functions to obtain the average number of customers for each priority class. We then focus on the preemptive resume case while allowing for weights associated with both priority class queues. We develop methodologies to obtain the optimal cutoffs for the situation when the weights of both queues are constant (i.e., not a function of queue length) and the situation when the weights change linearly with the queue lengths. It is important to point out that our method does not lead to a closed-form exact solution, but rather to a numerical approximation, from which cutoff policies are analyzed.

Key words. priority queue, queue-length cutoff, generating function

AMS subject classifications. 90-08, 60G99, 60-02

DOI. 10.1137/050648146

1. Introduction and literature review. Our research is primarily motivated by a disaster-relief project which deals with how to rescue casualties after a disaster occurs. We consider a dynamic disaster environment (e.g., earthquake), in which thousands of casualties need to be treated. The casualties in such a disaster setting are usually placed into four levels (see the description of HAZUS, a GIS-enabled software used by the Federal Emergency Management Agency (FEMA) for the purpose of earthquake loss estimation, in the paper by Al-Momani and Harrald [1]):

1. Severity level 1: injuries will require medical attention, but hospitalization is not needed.
2. Severity level 2: injuries will require hospitalization but are not considered life threatening.
3. Severity level 3: injuries will require hospitalization and can become life threatening if not promptly treated.
4. Severity level 4: victims are killed by the earthquake.

In an earthquake disaster-relief setting (e.g., the one that occurred in Northridge, CA, in 1994) severity level 1 and 4 calls are initially not responded to. Thus the system operates as a two-priority queue, with severity level 3 being priority 1 and severity level 2 being priority 2. Since injuries can rapidly deteriorate when unattended, it is possible that severity level 2 injuries that are left unattended for a long period of time can become even more critical than a typical severity level 3 injury. Thus operating in

*Received by the editors December 21, 2005; accepted for publication (in revised form) July 21, 2006; published electronically November 14, 2006. This paper was supported by the Air Force Office of Scientific Research through grant F49620-01-1-0371.
http://www.siam.org/journals/siap/67-1/64814.html
†Enterprise Optimization, United Airlines, 1200 East Algonquin, Elk Grove Township, IL 60007 (qiang.gong@united.com).
‡Industrial and Systems Engineering, University at Buffalo (SUNY), 438 Bell Hall, Buffalo, NY 14260 (batta@eng.buffalo.edu).

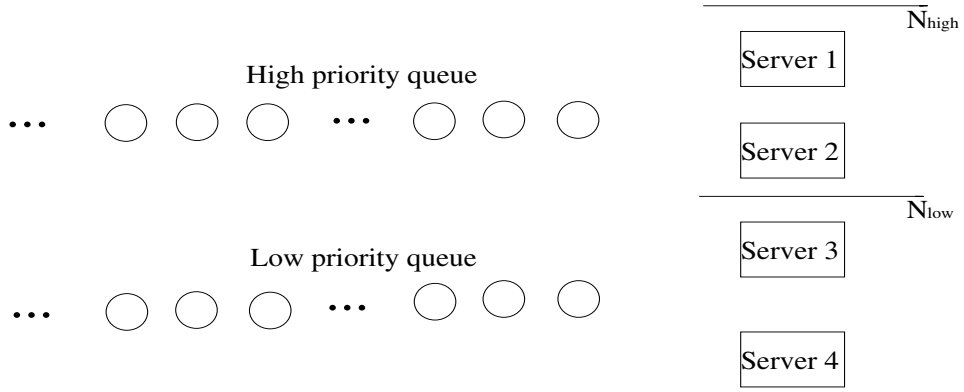


FIG. 1. *Server cutoff model.*

a strict priority queue model in a heavy-traffic situation would be detrimental. This provides the motivation to study a two-priority queueing system with a queue-length cutoff. This cutoff model is being implemented in the software for disaster relief being developed at the Center for Multisource Information Fusion at the University at Buffalo (SUNY). Details of its effectiveness via case studies developed for an earthquake scenario in Northridge, CA, will be presented in a later paper.

There are other applications of this queue-length cutoff model. For example, telecommunication in ATM (asynchronous transfer mode) networks also has this flavor. Voice data must flow through the network without noticeable distortion or delay. Losing a chunk of voice data isn't a problem, but a delay or receiving data out of order is. So voice is "delay sensitive, loss insensitive." On the contrary, computer data are "delay insensitive, loss sensitive," since individual chunks are not of much use until they are all received, but in many cases data delay in transmission is often acceptable. Based on the characteristics of both types of data, voice is classified as a high-priority class, computer data as a low-priority class. Again, computer data cannot be indefinitely delayed, so it makes sense to have a queue-length cutoff model in such a situation.

Previous research in the area of priority queueing models may be categorized as either server cutoff or queue-length cutoff. Figures 1 and 2 illustrate two straightforward examples for both types of models, respectively.

Depending upon the number of available servers, server cutoff discipline determines which classes of patients are qualified for service. The example shown in Figure 1 has two cutoffs, N_{high} and N_{low}. Obviously, N_{high} is equal to the total number of servers. Low-priority customers enter service only if fewer than $N_{low}(\leq N_{high})$ servers are busy. The purpose of this method is to reserve servers for high priorities. Taylor and Templeton [2] studied two variants of a simple two-priority server cutoff model: one assumes high-priority customers backlogged in the queue, while the other assumes they are lost if all servers are busy. Schaack and Larson [3] extended the two-priority case to the T-priority problem ($T \geq 3$). In a subsequent paper, Schaack and Larson [4] derived waiting time distribution of each class for an extension of this model, which assumes that customers require a random number of servers for service.

The queue-length cutoff priority queueing model can be regarded as the dual problem of the server cutoff model. Instead of considering the number of available servers, it manipulates the system based on the queue lengths. In the example shown

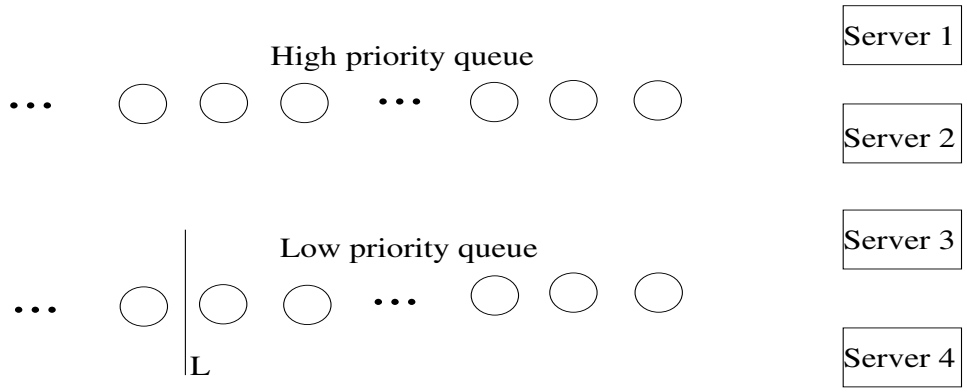


FIG. 2. *Queue-length cutoff model.*

in Figure 2, a cutoff number L is set on the low-priority queue. The servers process the high-priority queue only if the low-priority queue length is less than or equal to L. Once the threshold L is exceeded, part of or all of the servers go to serve the low-priority queue. Gross and Harris [5] published solutions of expected queue length and expected waiting time for a special two-priority model, which assumes the head-of-the-line discipline (i.e., $L = \infty$). Miller [6] obtained the steady-state probabilities by a matrix-geometric method for the same model. Recently, Knessl, Choi, and Tier [7] derived the joint queue-length distribution as an integral for their dynamic two-priority queue-length cutoff model. Our work builds upon their research by developing methodologies to obtain the desired queue-length cutoff L in the preemptive resume case for the situation when the weights associated with customers in both queues are constant and the situation when these weights change linearly with the queue lengths. In a disaster setting the weight signifies the importance associated with timely medical treatment of the patient.

2. Model formulation. Customers are designated one of two priority classes which are numbered as class-1 and class-2 so that the smaller the number, the higher the priority. The arrivals follow independent Poisson processes with rates λ_1 and λ_2, respectively. A single server processes both types of the customers with a mean rate μ. In order to make the system stable, we assume the stability condition as $\rho_1 + \rho_2 < 1$, where $\rho_1 = \frac{\lambda_1}{\mu}$ and $\rho_2 = \frac{\lambda_2}{\mu}$.

Let $X(t)$ and $Y(t)$ be the number of class-1 and class-2 customers in the system at time t, respectively. We consider the bivariate process $\{(X(t), Y(t)), t \geq 0\}$ with state space $S = \{(i, j) : i, j = 0, 1, 2, \dots\}$. The steady-state probabilities are defined as $p_{ij} = Pr\{\text{in steady-state } i \text{ class-1 customers and } j \text{ class-2 customers in the system}\}$.

The service discipline is controlled by the queue-length cutoff policy. A cutoff number L is set on the lower priority class. If the number of customers in the lower priority queue is less than or equal to L, only class-1 customers are served. Once the threshold L is exceeded, the server preempts the customer of class-1 currently in service. The server keeps serving class-2 customers until the queue length of class-2 is shortened to L. Then the server preempts the class-2 customer who is being processed and switches back to service class-1 customers. The sequence within each class is ordered on a first come, first served basis. When there is an empty queue, the server only processes the other queue regardless of the threshold L.

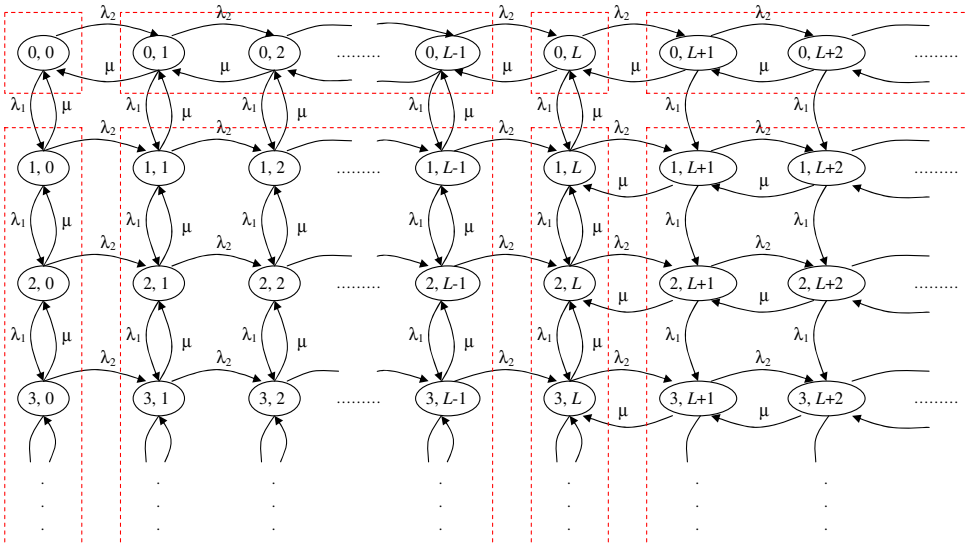


FIG. 3. *Rate transition diagram.*

This problem, in summary, is a Poisson-arrival, exponential-service, single-server, two-priority queue with the preemptive queue-length cutoff discipline.

3. Balance equations and generating functions. Our model is the same as that in [7]. The rate transition diagram is shown in Figure 3. The system is separated into two main parts by the threshold L. The first one gives class-1 customers higher priority, while class-2 customers receive higher priority in the second one.

Equating flow in to flow out, we get the balance equations for all sets of states in the dashed boxes in Figure 3 as follows:
$i = 0$ and $j = 0$:

$$(\lambda_1 + \lambda_2)p_0 = \mu p_{01} + \mu p_{10}. \tag{3.1}$$

$i = 0$ and $1 \le j \le L - 1$:

$$(\lambda_1 + \lambda_2 + \mu)p_{0j} = \lambda_2 p_{0,j-1} + \mu p_{0,j+1} + \mu p_{1j}. \tag{3.2}$$

$i = 0$ and $j = L$:

$$(\lambda_1 + \lambda_2 + \mu)p_{0L} = \lambda_2 p_{0,L-1} + \mu p_{0,L+1} + \mu p_{1L}. \tag{3.3}$$

$i = 0$ and $j \ge L + 1$:

$$(\lambda_1 + \lambda_2 + \mu)p_{0j} = \lambda_2 p_{0,j-1} + \mu p_{0,j+1}. \tag{3.4}$$

$i \ge 1$ and $j = 0$:

$$(\lambda_1 + \lambda_2 + \mu)p_{i0} = \mu p_{i+1,0} + \lambda_1 p_{i-1,0}. \tag{3.5}$$

$i \ge 1$ and $1 \le j \le L - 1$:

$$(\lambda_1 + \lambda_2 + \mu)p_{ij} = \lambda_1 p_{i-1,j} + \lambda_2 p_{i,j-1} + \mu p_{i+1,j}. \tag{3.6}$$

$i \geq 1$ and $j = L$:

$$(\lambda_1 + \lambda_2 + \mu)p_{iL} = \lambda_1 p_{i-1,L} + \lambda_2 p_{i,L-1} + \mu p_{i+1,L} + \mu p_{i,L+1}. \tag{3.7}$$

$i \geq 1$ and $j \geq L + 1$:

$$(\lambda_1 + \lambda_2 + \mu)p_{ij} = \lambda_1 p_{i-1,j} + \lambda_2 p_{i,j-1} + \mu p_{i,j+1}. \tag{3.8}$$

In view of the difficulty in obtaining the solutions from the recursive method, Knessl, Choi, and Tier [7] first derived the generating functions from the balance equations, then got the probabilities by inverting the generating function. Of interest in most applications are the measured system performances such as the expected number of class-1 customers N_1 and the expected number of class-2 customers N_2 in the system. However, their joint queue length is given by an integral which makes it difficult to calculate or even estimate N_1 and N_2.

We calculate N_1 and N_2 by computing the first moment of the generating function. To facilitate this we introduce the two-dimensional generating functions as follows:

$$H_j(w) = \sum_{i=0}^{\infty} p_{ij} w^i, \ 0 \leq j \leq L - 1, \tag{3.9}$$

$$H(w, z) = \sum_{i=0}^{\infty} \sum_{j=0}^{L-1} p_{ij} w^i z^j = \sum_{j=0}^{L-1} z^j H_j(w), \tag{3.10}$$

$$G(w, z) = \sum_{i=0}^{\infty} \sum_{j=L}^{\infty} p_{ij} w^i z^j, \tag{3.11}$$

$$F(w, z) = H(w, z) + G(w, z). \tag{3.12}$$

4. Expressions for generating functions. Since the threshold L divides the system into two parts, we need to calculate $H(w, z)$ and $L(w, z)$ separately to obtain the generating function $F(w, z)$ for the whole system.

We first consider $H(w, z)$. From (3.1), (3.2), (3.5), and (3.6), it is found that

$$H(w, z) = \frac{\left(\frac{\mu}{w} - \mu\right) p_0 + \lambda_2 z^L H_{L-1}(w) + \left(\frac{\mu}{w} - \frac{\mu}{z}\right) \sum_{j=1}^{L-1} p_{0j} z^j - \mu z^{L-1} p_{0L}}{\lambda_1 w + \lambda_2 z - (\lambda_1 + \lambda_2 + \mu) + \frac{\mu}{w}}. \tag{4.1}$$

Details of this derivation are shown in Appendix A.

Similarly, (3.3), (3.4), (3.7), and (3.8) yield

$$\begin{aligned} &\left[\lambda_1 w + \lambda_2 z - (\lambda_1 + \lambda_2 + \mu) + \frac{\mu}{z}\right] G(w, z) \\ &\quad = z^L \left[-\lambda_2 H_{L-1}(w) + \mu \left(\frac{1}{z} - \frac{1}{w}\right) H_L(w) + \frac{\mu}{w} p_{0L}\right]. \end{aligned} \tag{4.2}$$

Details of this derivation are shown in Appendix B. By the method presented by Knessl, Choi, and Tier [7], the left-hand side of (4.2) can be rewritten as

$$\frac{\lambda_2}{z} [(z - z_-(w))(z - z_+(w))] G(w, z), \tag{4.3}$$

where

$$z_-(w) = \frac{\mu + \lambda_1 + \lambda_2 - \lambda_1 w - \sqrt{(\mu + \lambda_1 + \lambda_2 - \lambda_1 w)^2 - 4\lambda_2 \mu}}{2\lambda_2}, \tag{4.4}$$

$$z_+(w) = \frac{\mu + \lambda_1 + \lambda_2 - \lambda_1 w + \sqrt{(\mu + \lambda_1 + \lambda_2 - \lambda_1 w)^2 - 4\lambda_2 \mu}}{2\lambda_2}. \tag{4.5}$$

By setting $z = z_-(w)$, we can get $H_L(w)$ in terms of $H_{L-1}(w)$ and p_{0L} as

$$H_L(w) = \frac{z_-(w)\left[\rho_2 w H_{L-1}(w) - p_{0L}\right]}{w - z_-(w)}. \tag{4.6}$$

Substituting (4.6) into (4.2) gives

$$G(w,z) = \left(\frac{p_{0L} - \rho_2 w H_{L-1}(w)}{\rho_2 (w - z_-(w))}\right)\left(\frac{z^L}{z - z_+(w)}\right). \tag{4.7}$$

It follows that

$$F(w,z) = \frac{\left(\frac{\mu}{w} - \mu\right) p_0 + \lambda_2 z^L H_{L-1}(w) + \left(\frac{\mu}{w} - \frac{\mu}{z}\right)\sum_{j=1}^{L-1} p_{0j} z^j - \mu z^{L-1} p_{0L}}{\lambda_1 w + \lambda_2 z - (\lambda_1 + \lambda_2 + \mu) + \frac{\mu}{w}} + \left[\frac{p_{0L} - \rho_2 w H_{L-1}(w)}{\rho_2 (w - z_-(w))}\right]\left[\frac{z^L}{z - z_+(w)}\right]. \tag{4.8}$$

In order to evaluate the expression for $F(w,z)$ given in (4.8) p_0, p_{0j} $(j = 1, \ldots, L-1)$, $H_{L-1}(w)$, and p_{0L} have to be determined.

We first focus on finding the initial state probability p_0. Intuitively, for our problem p_0 is solely determined by μ, λ_1, and λ_2 and is not affected by the ordering of service. Thus, the probability of idleness should be the same as the one in the $M/M/1$ model with two input streams. We formally establish this result in Proposition 4.1.

PROPOSITION 4.1. *The idle probability is given by* $p_0 = 1 - \rho_1 - \rho_2$.

Proof. Setting $z = 1$ in (4.1) and (4.7), we have

$$H(w,1) = \frac{\lambda_2 H_{L-1}(w) + \mu\left(\frac{1}{w} - 1\right)\sum_{j=1}^{L-1} p_{0j} + \mu\left(\frac{1}{w} - 1\right) p_0 + \mu p_{0L}}{\lambda_1 w - (\lambda_1 + \mu) + \frac{\mu}{w}}$$

and

$$G(w,1) = \left(\frac{p_{0L} - \rho_2 w H_{L-1}(w)}{\rho_2 (w - z_-(w))}\right)\left(\frac{1}{1 - z_+(w)}\right).$$

Then we set $w = 1$ in the equations above and use l'Hôpital's rule to get

$$H(1,1) = \frac{\mu}{\lambda_2}\left(\sum_{j=1}^{L-1} p_{0j} + p_{0L}\right)$$

and

$$G(1,1) = -\frac{\mu}{\lambda_2}\left(\sum_{j=1}^{L-1} p_{0j} + p_{0L}\right) + \left(\frac{\mu}{\mu - \lambda_1 - \lambda_2}\right) p_0.$$

By employing the condition that $F(1,1) = 1$, we find that $p_0 = 1 - \rho_1 - \rho_2$. ∎

Now we are going to describe how to calculate p_{0j} $(j = 1, \ldots, L-1)$. Define p'_{ij} $(i, j = 1, 2, \ldots)$ to be the state probabilities in the head-of-the-line case. Miller [6] presented a series of recursive formulas for calculating p'_{ij}. Knessl, Choi, and Tier [7] explained that p_{ij} in our model are the same as the corresponding p'_{ij} for all i and $0 \le j \le L-1$. Therefore, Miller's method [6] can be directly used for our problem to obtain p_{0j} $(j = 1, \ldots, L-1)$.

Our next focus is on deriving the expression for $H_{L-1}(w)$. Equations (3.1) and (3.5) yield

$$\left[\lambda_1 w - (\lambda_1 + \lambda_2 + \mu) + \frac{\mu}{w}\right] H_0(w) = -\mu p_{01} + \mu\left(\frac{1-w}{w}\right) p_0. \tag{4.9}$$

From (3.2) and (3.6), the relationship between $H_j(w)$ and $H_{j-1}(w)$ is found as

$$\begin{aligned}\left[\lambda_1 w - (\lambda_1 + \lambda_2 + \mu) + \frac{\mu}{w}\right] H_j(w) + \lambda_2 H_{j-1}(w) \\ = -\mu p_{0,j+1} + \frac{\mu}{w} p_{0j},\ 1 \le j \le L-1.\end{aligned} \tag{4.10}$$

By setting $A(w) = \lambda_1 w - (\lambda_1 + \lambda_2 + \mu) + \frac{\mu}{w}$ and solving (4.9) and (4.10) recursively, we can establish the following result (presented without proof).

PROPOSITION 4.2. *The general form of* $H_j(w)$ $(1 \le j \le L-1)$ *is*

$$\begin{aligned}H_j(w) = \mu \Bigg[-\frac{1}{A(w)} p_{0,j+1} + \left(\frac{A(w) + w\lambda_2}{w\,(A(w))^{j+1}}\right) \left(\sum_{k=0}^{j-1} (A(w))^{j-k-1}\, p_{0,j-k} (-\lambda_2)^k\right) \\ + (-\lambda_2)^j \left(\frac{1-w}{w\,(A(w))^{j+1}}\right) p_0 \Bigg], \quad 1 \le j \le L-1.\end{aligned} \tag{4.11}$$

By setting $j = L-1$ and $w = 1$ in (4.11), we get p_{0L} as

$$p_{0L} = \rho_2 H_{L-1}(1). \tag{4.12}$$

Knessl, Choi, and Tier [7] presented an exact formula of $H_{L-1}(1)$ as an integral. However, as they noticed, for $L > 30$ the calculation becomes intractable. Thus we use an approximate method to calculate $H_{L-1}(1)$. From (3.9) we know that $H_{L-1}(1) = \sum_{i=0}^{\infty} p_{i,L-1}$. Thus $H_{L-1}(1)$ is approximated by $\sum_{i=0}^{M} p_{i,L-1}$, where M is a sufficiently large number—in particular, we will later see in section 7 that using $M = 5L$ works well in numerical tests.

5. Derivation of expected numbers in system. Armed with an expression of the generating function $F(w, z)$, we proceed to calculate L_1 and L_2. We take the partial derivatives of $F(w, z)$ in terms of both w and z and evaluate at (1,1) to get the results:

$$\begin{aligned}N_1 = {} & \frac{2\mu\lambda_1 \left(\sum_{j=1}^{L-1} p_{0j} + p_0\right) - 2\mu\lambda_2 H'_{L-1}(1) - \lambda_2(\mu - \lambda_1) H''_{L-1}(1)}{2(\mu - \lambda_1)^2} \\ & + \frac{2H'_{L-1}(1) + H''_{L-1}(1)}{2(z'_-(1) - 1)(1 - z_+(1))} - \frac{(H_{L-1}(1) + H'_{L-1}(1))(z''_-(1))}{2(1 - z'_-(1))^2 (1 - z_+(1))} \\ & - \frac{(H_{L-1}(1) + H'_{L-1}(1))(z'_+(1))}{(1 - z'_-(1))(1 - z_+(1))^2}\end{aligned} \tag{5.1}$$

and

$$N_2 = (L-1)H_{L-1}(1) + \frac{1}{\rho_2}\sum_{j=1}^{L-1}(j-1)p_{0j} + \left[\frac{H_{L-1}(1)+H'_{L-1}(1)}{z'_-(1)-1}\right]\left[\frac{L(1-z_+(1))-1}{(1-z_+(1))^2}\right]. \tag{5.2}$$

To evaluate N_1 and N_2, we observe that we further need to know the values of $z'_-(1)$, $z''_-(1)$, $z_+(1)$, $z'_+(1)$, $H'_{L-1}(1)$, and $H''_{L-1}(1)$. These are as follows:

$$z'_-(1) = \frac{\lambda_1}{\mu-\lambda_2}, \qquad z''_-(1) = \frac{2\mu\lambda_1^2}{(\mu-\lambda_2)^3},$$
$$z_+(1) = \frac{\mu}{\lambda_2}, \qquad z'_+(1) = -\frac{\mu\lambda_1}{(\mu-\lambda_2)\lambda_2},$$
$$H'_{L-1}(1) = \mu\left[\left(\frac{\lambda_1-\mu}{\lambda_2^2}\right)p_{0L} + \left(\frac{\lambda_1+\lambda_2-\mu}{\lambda_2^2}\right)\left(\sum_{j=1}^{L-1}p_{0j}\right) + \frac{p_0}{\lambda_2}\right],$$

and

$$H''_{L-1}(1) = \mu\left[\left(\frac{2\mu\lambda_2+2(\mu-\lambda_1)^2}{\lambda_2^3}\right)p_{0L} + \left(\frac{2\mu\lambda_2+2(\mu-\lambda_1-\lambda_2)(L\mu-L\lambda_1+\lambda_2)}{\lambda_2^3}\right)\left(\sum_{j=1}^{L-1}p_{0j}\right) + \left(\frac{2(\mu-\lambda_1-\lambda_2)(\lambda_1-\mu)}{\lambda_2^3}\right)\left(\sum_{j=2}^{L-1}(j-1)p_{0j}\right) - \left(\frac{2(L\mu-L\lambda_1+\lambda_2)}{\lambda_2^2}\right)p_0\right].$$

6. Properties. Having studied the generating functions and derived the formulas for N_1 and N_2, we are ready to discuss some important properties of this queueing system.

As mentioned previously in section 1, our queueing model is a generalization of the head-of-the-line model. The first two properties are straightforward to establish. The reader is referred to [8] for detailed proofs.

PROPERTY 6.1. *When $L = \infty$, the queue-length cutoff model is reduced to the head-of-the-line model.*

The next property has been discovered through intuitive observation. The point here is to investigate the mean number of customers (including both class-1 and class-2) in the system. If we consider the two classes as a whole, it is instructive to point out that changing the value of L only changes the order of the service and never changes the mean number of customers in the system. Clearly, the mean number of customers in our problem is the same as the one in the head-of-the-line model, or even the same as the one in the nonpriority $M/M/1$ model. It needs to be noted that the service rates of the two classes have been assumed to be equal and the classes have the same weight—hence the class-1 and class-2 jobs are indistinguishable.

PROPERTY 6.2. *Independent of the queue-length cutoff L, the mean number of customers $N_1 + N_2$ is a constant, which is given by*

$$N_1 + N_2 = \frac{\lambda_1+\lambda_2}{\mu-\lambda_1-\lambda_2}. \tag{6.1}$$

Although the mean total number of customers in the system is a constant, it is quite natural to see that N_1 and N_2 do change as L changes. Consider the example of increasing the value of L. It is intuitively clear that the server spends more time on the high-priority queue than before. Thus N_1 decreases as L increases. Conversely, N_2 is an increasing function in terms of L.

Basically, there are two different preemptive priority disciplines: preemptive resume and preemptive repeat. Preemptive resume allows preempted customers to continue their service where they left off when they reenter service, while preemptive repeat requires preemptive customers to pick up a new value of service time from the service-time distribution whenever they reenter service. The following property is presented and proved under the first case only, i.e., preemptive resume.

PROPERTY 6.3. *Under the preemptive resume priority discipline, suppose there are two queue-length cutoffs L and L', where $L < L'$. Then the following statements are true:*

$$N_1(L) \geq N_1(L') \tag{6.2}$$

and

$$N_2(L) \leq N_2(L'). \tag{6.3}$$

Proof. Since it is not even clear how $H_{L-1}(1)$ and $\sum_{j=1}^{L-1} p_{0j}$ behave as L changes, it seems impossible to prove this property directly by (5.1) and (5.2). The remarkable difficulty makes us resort to the following method.

We consider an arbitrary busy period. Obviously, when the priority discipline is preemptive resume, the total service time of a class-1 or class-2 customer is in no way affected by the number of times he/she is preempted. That is, changing the value of L only changes the order of service, while the duration of any busy period is always equivalent to the total service time of the customers in that period.

A typical example is shown in Figure 4, where the queue-length cutoff $L = 4$. It is instructive to see that N_1 and N_2 in a busy period can be calculated as

$$N_i = \frac{\text{Area}(N_i^L)}{D}, \quad i = 1, 2, \tag{6.4}$$

where

$$\begin{aligned} \text{Area}(N_i^L) &= \text{area covered by class-}i\text{ customers given that the queue-length} \\ &\qquad \text{cutoff is } L, \text{and} \\ D &= \text{duration of that busy period.} \end{aligned}$$

In this proof, we only focus on (6.2). The assertion of (6.3) can be derived in a similar manner. Suppose the current queue-length cutoff is L. If there is no preemption for class-1 customers in this period, it is easy to see that the number of preemptions for class-1 customers is still zero if L is increased to L'.

Consider now that there is at least one preemption for the high-priority queue. We pick up an arbitrary preemption to study. An example is shown in Figure 5. We can see that the preempted time point and the resume time point have been shifted from P^L and R^L to $P^{L'}$ and $R^{L'}$, respectively. Clearly, this shifting does not affect the area before time point P^L. Changes only occur after that time point. Since the interarrival time between two customers and the service time of a customer cannot

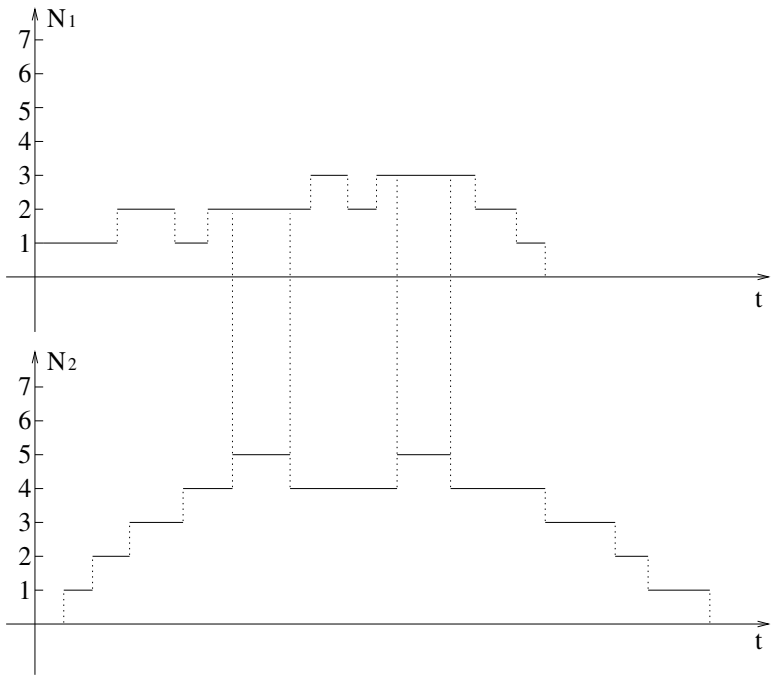


FIG. 4. *An example of a busy period.*

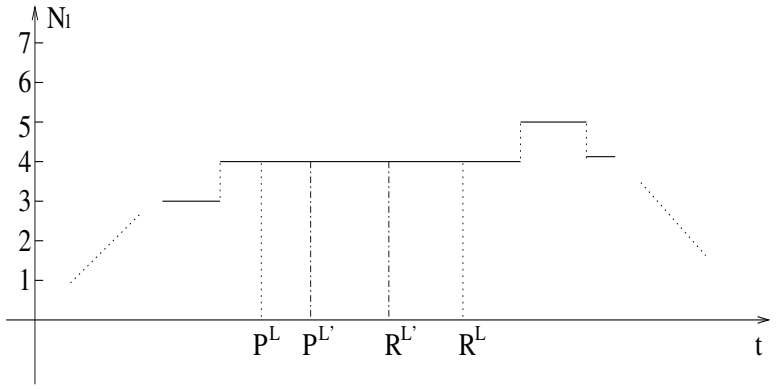


FIG. 5. *An example of a preemption.*

be zero, the preemption time $R^{L'} - P^{L'}$ is strictly less than the one $R^L - P^L$. This result leads to another conclusion: at any time t ($\geq P^L$), $X(t)$ in the second case is less than or equal to the corresponding quantity in the first case. Thus, when L is increased to L', the area after time point P^L is strictly decreased. Combining the two areas together, we can conclude that Area(N_1) is a decreasing function of L. This yields the final result that $N_1(L) \geq N_1(L')$ when $L < L'$. The property follows. □

Property 6.2 tells us that the total number of customers in the system is constant. However, let us consider an example. Suppose that there are 8 priority-1 customers and 2 priority-2 customers in case 1 and that there are 2 priority-1 customers and 8 priority-2 customers in case 2. Although the total number of customers is 10 for

both cases, it is obvious that case 1 is much worse than case 2. The reason is that we usually assign a higher weight, π_1, to priority-1 customers and a lower weight, π_2, to priority-2 customers. In consideration of the weighted number of customers in the system, Properties 6.2 and 6.3 lead us to the following result for minimizing the function $\pi_1 N_1 + \pi_2 N_2$ for certain choices of weights.

PROPERTY 6.4. *In the preemptive resume model, suppose that π_1 and π_2 are constant, where $\pi_1 > \pi_2 > 0$ and $\pi_1 + \pi_2 = 1$. The optimal queue-length cutoff is given by $L^* = \infty$.*

Proof. We consider two queue-length cutoffs L and L' with $L < L'$. Suppose that the average number of priority-1 customers is N_1 and that the average number of priority-2 customers is N_2 if the cutoff is L. Thus the weighted number of customers for this case is calculated as

$$\pi_1 N_1 + \pi_2 N_2. \tag{6.5}$$

When L is increased to L', Property 6.3 tells us that N_1 decreases. Suppose the number of priority-1 customers changes to $N_1 - \epsilon$, where $\epsilon > 0$. Then Property 6.2 shows that the number of priority-2 customers changes to $N_2 + \epsilon$. The weighted number of customers is given by

$$\pi_1 N_1 + \pi_2 N_2 + (\pi_2 - \pi_1)\epsilon. \tag{6.6}$$

It is easy to see that the value of (6.6) is smaller than the value of (6.5). Thus we conclude that the optimal cutoff is given by $L^* = \infty$. □

We now address the more interesting case—which is particularly relevant to the disaster-relief application—where the weight of a priority class may vary as the queue length changes. Generally speaking, the weight increases (decreases) as the queue length increases (decreases). Since π_1 and π_2 correlate each other ($\pi_1 = 1 - \pi_2$), we only need to specify one of them, e.g., π_2. We consider the case when the weight is a linear function of the queue length, i.e., the function can be expressed as $\pi_2 = KN_2 + C$. The function is shown in Figure 6, where $N_{2L_{min}}$ and $N_{2L_{max}}$ stand for the numbers of priority-2 customers under the minimal cutoff ($L = 3$) and the maximal cutoff ($L = \infty$) cases, respectively. The weights, $\pi_{2L_{min}}$ and $\pi_{2L_{max}}$, for these two extreme cases are assumed to be given. The parameters K and C are then determined uniquely by the two points $(N_{2L_{min}}, \pi_{2L_{min}})$ and $(N_{2L_{max}}, \pi_{2L_{max}})$ as follows:

$$K = \frac{\pi_{2L_{max}} - \pi_{2L_{min}}}{N_{2L_{max}} - N_{2L_{min}}} \tag{6.7}$$

and

$$C = \pi_{2L_{max}} - \frac{(\pi_{2L_{max}} - \pi_{2L_{min}})N_{2L_{max}}}{N_{2L_{max}} - N_{2L_{min}}}. \tag{6.8}$$

The weighted number of customers for the minimal cutoff case is given by

$$\pi_{1L_{min}} N_{1L_{min}} + \pi_{2L_{min}} N_{2L_{min}}. \tag{6.9}$$

Assume that L' is an arbitrary cutoff that is larger than L_{min}. Define δ and Δ to be the values increased from $\pi_{2L_{min}}$ to $\pi_{2L'}$ and from $N_{2L_{min}}$ to $N_{2L'}$, respectively. We can verify that $\delta = K\Delta$. Then the weighted number of customers for this case is calculated as

$$(\pi_{1L_{min}} - \delta)(N_{1L_{min}} - \Delta) + (\pi_{2L_{min}} + \delta)(N_{2L_{min}} + \Delta), \tag{6.10}$$

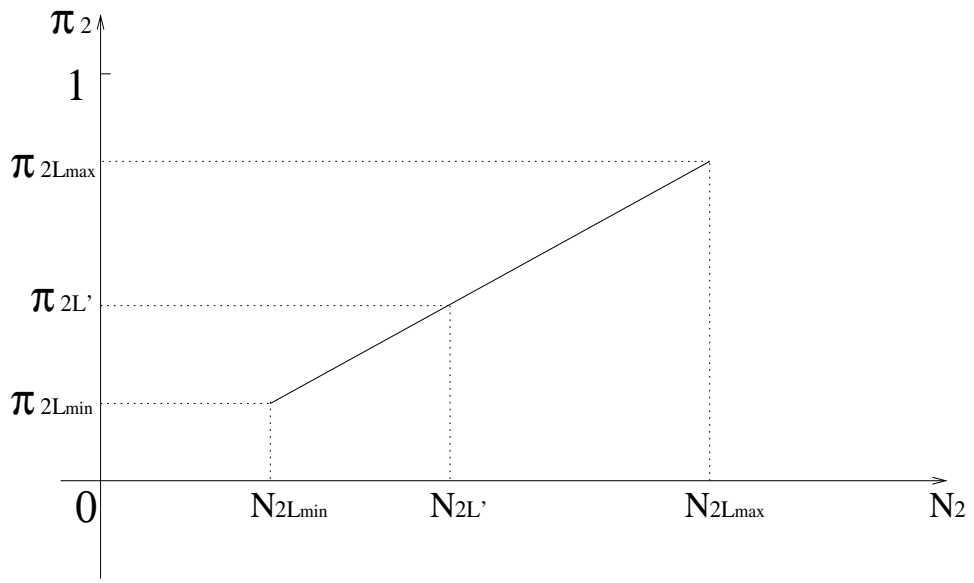


FIG. 6. *A linear weight function.*

which is equivalent to

$$(\pi_{1L_{min}} N_{1L_{min}} + \pi_{2L_{min}} N_{2L_{min}}) + \{2K\Delta^2 + [(\pi_{2L_{min}} - \pi_{1L_{min}}) + (N_{2L_{min}} - N_{1L_{min}})]\Delta\}. \tag{6.11}$$

Comparing (6.11) with (6.9), we see that the optimal cutoff is determined by the discrete function

$$f(\Delta) = 2K\Delta^2 + [(\pi_{2L_{min}} - \pi_{1L_{min}}) + (N_{2L_{min}} - N_{1L_{min}})]\Delta. \tag{6.12}$$

Since $K > 0$, the value Δ^*, which minimizes the *continuous* equation (6.12), is

(6.13)

$$\Delta^* = \frac{(1 - 2\pi_{2L_{min}})(N_{2L_{max}} - N_{2L_{min}}) + (\pi_{2L_{max}} - \pi_{2L_{min}})(N_{1L_{min}} - N_{2L_{min}})}{4(\pi_{2L_{max}} - \pi_{2L_{min}})}.$$

However, considering that $0 \le \Delta \le N_{2L_{max}} - N_{2L_{min}}$, we can identify the following three cases:

Case 1: $\Delta^* \le 0$. In this case, $L^* = L_{min}$.

Case 2: $\Delta^* \ge N_{2L_{max}} - N_{2L_{min}}$. In this case, $L^* = L_{max}$.

Case 3: $0 < \Delta^* < N_{2L_{max}} - N_{2L_{min}}$. The function $f(\Delta)$ in our research is discrete. Usually the optimal Δ^* does not correspond to points in this discrete set. In this case, we only need to identify two points as follows:

$$\Delta_1 = \min\{\Delta : f(\Delta) \le f(\Delta^*)\}$$

and

$$\Delta_2 = \min\{\Delta : f(\Delta) > f(\Delta^*)\}.$$

TABLE 1
An example of the experiments with $\lambda_1 = 0.3$, $\lambda_2 = 0.2$, *and* $\mu = 1$.

L	Exact method	Approximate method	Error
3	0.078103088	0.0781031	1.49418E-07
4	0.031085012	0.031085	3.89899E-07
5	0.012962869	0.0129629	2.39916E-06
6	0.005570972	0.00557097	3.50747E-07
7	0.002444427	0.00244443	1.22033E-06
8	0.001088903	0.0010889	2.50803E-06
9	0.000490679	0.000490678	1.47469E-06
10	0.000223131	0.000223118	5.81318E-05

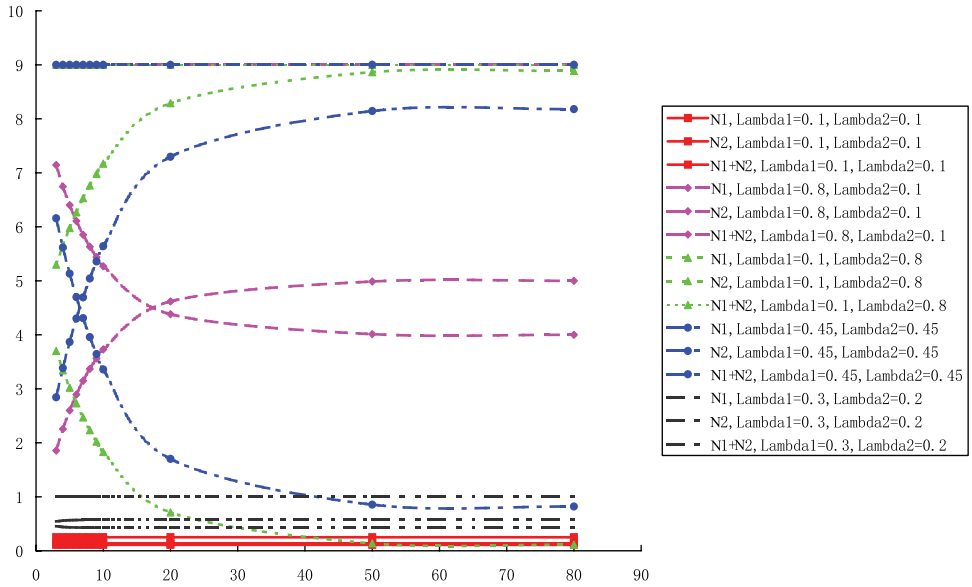


FIG. 7. *Computational results of* N_1, N_2, *and* $N_1 + N_2$.

The optimal value of Δ, denoted as Δ^{**}, is given by

$$\Delta^{**} = \text{argmin}\{f(\Delta_1), f(\Delta_2)\}.$$

The cutoff L^* which corresponds to Δ^{**} is the optimal solution.

7. Computational results. Before proceeding with the numerical results, we first investigate the approximate method of calculating $H_{L-1}(1)$. As discussed in section 4, for $L > 30$ an appropriate value of M needs to be used in order to make $\sum_{i=0}^{M} p_{i,L-1}$ as a good estimation of $H_{L-1}(1)$. We conduct a series of numerical experiments using various combinations of λ_1 and λ_2. We employ eight different values of L (from 3 to 10) in each experiment. The exact results calculated by the method in [7] are used as benchmarks. After some trial runs, we find that $\sum_{i=0}^{M} p_{i,L-1}$ can provide a good approximation of $H_{L-1}(1)$ if $M = 5L$. In most of the cases, the errors are within 0.1%. Table 1 shows a sample of results from these experiments.

Next we focus on calculating N_1 and N_2. The results are shown in Figure 7. We can see that for all cases our model approaches the head-of-the-line case as L increases. The total number of customers in the system is a constant, while N_1 and N_2 decrease and increase, respectively.

TABLE 2
Optimal cutoffs.

	L^*	Weighted no. of customers
$\lambda_1 = 0.1, \lambda_2 = 0.1$	3	0.118114
$\lambda_1 = 0.1, \lambda_2 = 0.8$	5	3.9845
$\lambda_1 = 0.45, \lambda_2 = 0.45$	9	4.39441
$\lambda_1 = 0.8, \lambda_2 = 0.1$	13	4.47104
$\lambda_1 = 0.1, \lambda_2 = 0.88$	37	24.0178
$\lambda_1 = 0.49, \lambda_2 = 0.49$	42	24.2835
$\lambda_1 = 0.88, \lambda_2 = 0.1$	46	24.4654

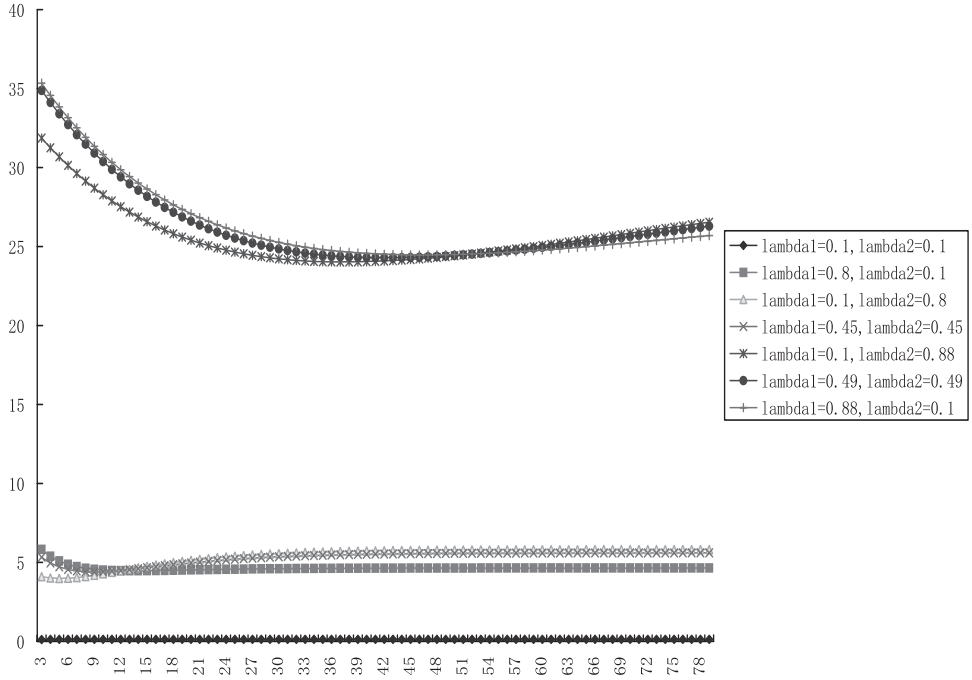


FIG. 8. *Computational results of weighted number of customers.*

We now present an example which calculates the weighted number of customers under the case of a linear weight function. We set $\pi_{2L_{min}} = 0.45$ and $\pi_{2L_{max}} = 0.65$. The results are shown in Table 2. We consider seven different combinations of λ_1 and λ_2. In the case of $\lambda_1 = \lambda_2 = 0.1$, since $\Delta^*(= -0.006849)$ is less than zero, we have $L^* = L_{min}$. The values of Δ^* in all the other cases satisfy the condition $0 < \Delta^* < N_{2L_{max}} - N_{2L_{min}}$, and thus different finite optimal cutoffs are obtained as shown in Table 2.

Figure 8 also illustrates the detailed results. When both λ_1 and λ_2 are small ($\lambda_1 = \lambda_2 = 0.1$), the change of the average weighted number of customers is negligible with a change in the value of L. This is because the system is in an *unsaturated* status, which leads to very little change of N_1 and N_2 values. When either λ_1 or λ_2 (or both) increases, the average number of customers becomes much larger than in the *unsaturated* case. We focus on the three cases ($\lambda_1 = 0.1, \lambda_2 = 0.8$; $\lambda_1 = 0.45, \lambda_2 = 0.45$; $\lambda_1 = 0.8, \lambda_2 = 0.1$) in which the total arrival rates are the same. We can see that the optimal cutoff increases as λ_1 increases. Given the condition that $\lambda_1 + \lambda_2$ is a constant, Property 6.2 shows that $N_1 + N_2$ is a constant no matter

what the specific values of λ_1 and λ_2 are. Thus the increase of λ_1 causes an increase in N_1 and a decrease in N_2. More priority-1 customers lead to an increase in the optimal cutoff value. As the system approaches the *saturated* status, both N_1 and N_2 increase dramatically. Consequently, the average number of customers also increases significantly. This causes a sharp increase in the optimal cutoff value.

8. Summary and future work. A two-priority, preemptive, single-server system with a queue-length cutoff queueing discipline has been studied in this paper. This is a generic problem for various applications such as disaster relief and telecommunication. Expressions for calculating the number of class-1 customers and the number of class-2 customers are developed based on a generating function approach. The method we present does not lead to a closed-form solution, but rather to an effective numerical approximation.

We have shown that our model reduces to the head-of-the-line model if $L = \infty$. The total number of customers in the system is shown to be constant with respect to L. Then we focus on the preemptive resume case, in which N_1 and N_2 are decreasing and increasing functions of L, respectively. The weighted average number of customers is first analyzed for the case where the weights for both queues are constant. We prove that the optimal policy is to set $L^* = \infty$. Then the case where the weights change linearly with the queue lengths is analyzed and a procedure is developed to find the optimal cutoff. Numerical results illustrate the properties and other results.

There are several possible directions for future work:

(1) In our model, the moment that the number of low-priority jobs hits threshold L, the server stops working on high-priority jobs entirely. An alternative threshold policy in which the server is shared when the threshold is reached should be studied.
(2) For analytical tractability we assumed that the service rate for the high- and low-priority jobs is the same. The more realistic case where the service rates are class dependent should be studied.
(3) Another direction of future work is to consider the use of an alternate solution method, namely, dimensionality reduction for Markov chains. The work of Osogami, Harchol-Balter, and Scheller-Wolf [9] serves as a useful starting point.
(4) A further opportunity is in analyzing the multiserver version of our model, which is closer to reality for a disaster-relief application.
(5) By applying the memoryless property of the exponential distribution it may be possible to establish Property 6.3 for the preemptive repeat case.

Appendix A. Derivation of (4.1). From (3.10), we get

$$\begin{aligned}&(\lambda_1+\lambda_2+\mu)H(w,z)\\ &\quad=(\lambda_1+\lambda_2+\mu)\left(p_0+\sum_{j=1}^{L-1}p_{0j}z^j+\sum_{i=1}^{\infty}p_{i0}w^i+\sum_{i=1}^{\infty}\sum_{j=1}^{L-1}p_{ij}w^iz^j\right).\end{aligned}\tag{A.1}$$

From (3.1), (3.2), (3.5), and (3.6), the right-hand side of (A.1) can be written as

$$\begin{aligned}&(\lambda_1+\lambda_2+\mu)p_0+\sum_{j=1}^{L-1}(\lambda_2p_{0,j-1}+\mu p_{0,j+1}+\mu p_{1j})z^j+\sum_{i=1}^{\infty}(\mu p_{i+1,0}+\lambda_1p_{i-1,0})w^i\\ &+\sum_{i=1}^{\infty}\sum_{j=1}^{L-1}(\lambda_1p_{i-1,j}+\lambda_2p_{i,j-1}+\mu p_{i+1,j})w^iz^j.\end{aligned}\tag{A.2}$$

Regrouping the terms in (A.2) according to λ_1, λ_2, and μ, we have

$$
\begin{aligned}
&(\lambda_1+\lambda_2+\mu)H(w,z)\\
&\quad=\lambda_1\left(p_0+\sum_{i=1}^{\infty}p_{i-1,0}w^i+\sum_{i=1}^{\infty}\sum_{j=1}^{L-1}p_{i-1,j}w^iz^j\right)\\
&\quad\quad+\lambda_2\left(p_0+\sum_{j=1}^{L-1}p_{0,j-1}z^j+\sum_{i=1}^{\infty}\sum_{j=1}^{L-1}p_{i,j-1}w^iz^j\right)\\
&\quad\quad+\mu\left[p_0+\sum_{j=1}^{L-1}(p_{0,j+1}+p_{1j})z^j+\sum_{i=1}^{\infty}p_{i+1,0}w^i+\sum_{i=1}^{\infty}\sum_{j=1}^{L-1}p_{i+1,j}w^iz^j\right].
\end{aligned}
\tag{A.3}
$$

Then we arrange the terms on the right-hand side in (A.3) and obtain

$$
\begin{aligned}
&(\lambda_1+\lambda_2+\mu)H(w,z)\\
&\quad=\lambda_1wH(w,z)+\lambda_2(zH(w,z)-z^LH_{L-1}(w))\\
&\quad\quad+\frac{\mu}{w}H(w,z)+\left(\frac{\mu}{z}-\frac{\mu}{w}\right)\sum_{j=1}^{L-1}p_{0j}z^j+\mu p_0-\frac{\mu}{w}p_0+\mu z^{L-1}p_{0L}.
\end{aligned}
\tag{A.4}
$$

Equation (A.4) immediately yields

$$
H(w,z)=\frac{\left(\frac{\mu}{w}-\mu\right)p_0+\lambda_2z^LH_{L-1}(w)+\left(\frac{\mu}{w}-\frac{\mu}{z}\right)\sum_{j=1}^{L-1}p_{0j}z^j-\mu z^{L-1}p_{0L}}{\lambda_1w+\lambda_2z-(\lambda_1+\lambda_2+\mu)+\frac{\mu}{w}}.
\tag{A.5}
$$

Appendix B. Derivation of (4.2). From (3.11), we get

$$
\begin{aligned}
&(\lambda_1+\lambda_2+\mu)G(w,z)\\
&\quad=(\lambda_1+\lambda_2+\mu)\left(p_{0L}z^L+\sum_{i=1}^{\infty}p_{iL}w^iz^L+\sum_{j=L+1}^{\infty}p_{0j}z^j+\sum_{i=1}^{\infty}\sum_{j=L+1}^{\infty}p_{ij}w^iz^j\right).
\end{aligned}
\tag{B.6}
$$

From (3.3), (3.4), (3.7), and (3.8), the right-hand side of (B.6) can be written as

$$
\begin{aligned}
&z^L(\lambda_2p_{0,L-1}+\mu p_{0,L+1}+\mu p_{1L})+z^Ls\sum_{i=1}^{\infty}(\lambda_1p_{i-1,L}+\lambda_2p_{i,L-1}+\mu p_{i+1,L}+\mu p_{i,L+1})w^i\\
&\quad+\sum_{j=L+1}^{\infty}(\lambda_2p_{0,j-1}+\mu p_{0,j+1})z^j+\sum_{i=1}^{\infty}\sum_{j=L+1}^{\infty}(\lambda_1p_{i-1,j}+\lambda_2p_{i,j-1}+\mu p_{i,j+1})w^iz^j.
\end{aligned}
\tag{B.7}
$$

Regrouping the terms in (B.7) according to λ_1, λ_2, and μ, we have

$$
\begin{aligned}
&(\lambda_1+\lambda_2+\mu)G(w,z)\\
&\quad=\lambda_1 w\left(z^L\sum_{i=1}^{\infty}p_{i-1,L}w^{i-1}+\sum_{i=1}^{\infty}\sum_{j=L+1}^{\infty}p_{i-1,j}w^{i-1}z^j\right)\\
&\qquad+\lambda_2 z\left(z^{L-1}p_{0,L-1}+\sum_{j=L+1}^{\infty}p_{0,j-1}z^{j-1}+z^{L-1}\sum_{i=1}^{\infty}p_{i,L-1}w^i\right.\\
&\qquad\left.+\sum_{i=1}^{\infty}\sum_{j=L+1}^{\infty}p_{i,j-1}w^iz^{j-1}\right)+\frac{\mu}{z}\left[z^{L+1}(p_{0,L+1}+p_{1L})+\sum_{j=L+1}^{\infty}p_{0,j+1}z^{j+1}\right.\\
&\qquad\left.+z^{L+1}\sum_{i=1}^{\infty}(p_{i+1,L}+p_{i,L+1})w^i+\sum_{i=1}^{\infty}\sum_{j=L+1}^{\infty}p_{i,j+1}w^iz^{j+1}\right].
\end{aligned}
\tag{B.8}
$$

Then we arrange the terms on the right-hand side in (B.8) and obtain

$$
\begin{aligned}
&(\lambda_1+\lambda_2+\mu)G(w,z)\\
&\quad=\lambda_1 wG(w,z)+\lambda_2 zG(w,z)+\lambda_2 z^LH_{L-1}(w)\\
&\qquad+\frac{\mu}{z}G(w,z)+z^L\left[\mu\left(-\frac{1}{z}H_L(w)+\frac{1}{w}H_L(w)\right)-\frac{\mu}{w}p_{0L}\right].
\end{aligned}
\tag{B.9}
$$

Equation (B.9) immediately yields

$$
\begin{aligned}
&\left[\lambda_1 w+\lambda_2 z-(\lambda_1+\lambda_2+\mu)+\frac{\mu}{z}\right]G(w,z)\\
&\quad=z^L\left[-\lambda_2H_{L-1}(w)+\mu\left(\frac{1}{z}-\frac{1}{w}\right)H_L(w)+\frac{\mu}{w}p_{0L}\right].
\end{aligned}
\tag{B.10}
$$

Acknowledgment. The authors are grateful for constructive comments from two anonymous referees that led to a much tighter presentation of the results.

REFERENCES

[1] N. M. Al-Momani and J. R. Harrald, *Sensitivity of earthquake loss estimation model: How useful are the predictions?*, Internat. J. Risk Assess. Management, 4 (2003), pp. 1–19.

[2] I. D. S. Taylor and J. G. C. Templeton, *Waiting time in a multi-server cutoff-priority queue, and its application to an urban ambulance service*, Oper. Res., 28 (1980), pp. 1168–1188.

[3] C. Schaack and R. C. Larson, *An N-server cutoff priority queue*, Oper. Res., 34 (1986), pp. 257–266.

[4] C. Schaack and R. C. Larson, *An N server cutoff priority queue where arriving customers request a random number of servers*, Management Sci., 35 (1989), pp. 614–634.

[5] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, Wiley, New York, 1998.

[6] D. R. Miller, *Computation of steady-state probabilities for M/M/1 priority queues*, Oper. Res., 29 (1981), pp. 945–958.

[7] C. Knessl, D. I. Choi, and C. Tier, *A dynamic priority queue model for simultaneous service of two traffic types*, SIAM J. Appl. Math., 63 (2002), pp. 398–422.

[8] Q. Gong, *Responding to Casualties in a Disaster Relief Operation: Initial Ambulance Allocation and Reallocation, and Switching of Casualty Priorities*, Ph.D. dissertation, Department of Industrial and Systems Engineering, University at Buffalo (SUNY), Buffalo, NY, 2005.

[9] T. Osogami, M. Harchol-Balter, and A. Scheller-Wolf, *Analysis of cycle stealing with switching cost*, in Proceedings of the 2003 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, San Diego, CA, 2003, pp. 184–195.

SIAM J. APPL. MATH.
Vol. 67, No. 1, pp. 116–137

FLAME BALLS FOR A FREE BOUNDARY COMBUSTION MODEL WITH RADIATIVE TRANSFER*

JAN BOUWE VAN DEN BERG†, VINCENT GUYONNE‡, AND JOSEPHUS HULSHOF†

Abstract. We study radial flame ball solutions of a three-dimensional free boundary problem (FBP), which models combustion of a gaseous mixture with dust in a microgravity environment. The model combines diffusion of mass and temperature with reaction at the flame front, the reaction rate being temperature dependent. The radiative flux due to the presence of dust enters the equation for the temperature in the form of a divergence term. This flux is modeled by Eddington's radiative transfer equation. The main parameters are the dimensionless opacity and the ratio of radiative and thermal fluxes. We prove existence of spherical flame ball solutions for the FBP. Bifurcation diagrams are obtained, exhibiting the multiplicity of solutions. Singular limit cases of the parameter values are also discussed.

Key words. flame balls, radiative transfer, free boundary problem, existence

AMS subject classifications. 80A25, 35J60, 35R35

DOI. 10.1137/050636516

1. Introduction.

Combustion processes in gaseous mixtures exhibit a variety of phenomena, such as propagating flame fronts, and, in zero- or microgravity situations, flame balls. The latter are perhaps harder to observe, but the advantage is that they are stationary. From a mathematical point of view they are easier to understand, namely as equilibria rather than traveling wave solutions of the mathematical models used to describe the combustion processes. From a physical point of view, because of the force and speed of the reaction, it is hard to do controlled experiments on flame fronts, whereas the combustion is much less violent in flame balls, which can be observed for prolonged periods of time at the cost of having to transfer the experiment to a microgravity environment. In any case, the high costs and experimental difficulties in combustion research highlight the need for a thorough understanding of the mathematical models.

Since the work of Zeldovich et al. [1], flame balls are known to exist for models of combustion with simple chemistry, such as a one step reaction in which a gaseous reactant is converted into a gaseous product. Figure 1 is a sketch of a flame ball in the nonradiative case. Note that, in this particular situation, the temperature θ_b in the burnt region is constant inside the ball. In this model, commonly referred to as the adiabatic case, flame balls are linearly unstable, in apparent agreement with the absence of experimentally observed flame balls. That was, until 1984, when Ronney discovered, by surprise, the existence during drop tower experiments of physical flame balls, later confirmed by experiments in the space shuttle [2, 3]. Since then, several effects have been taken into account in combustion models to explain stabilization of

*Received by the editors July 21, 2005; accepted for publication (in revised form) July 24, 2006; published electronically November 14, 2006. This work was supported by a CNRS/NWO grant and the RTN network Front-Singularities, HPRN-CT-2002-00274.
http://www.siam.org/journals/siap/67-1/63651.html
†Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands (janbouwe@few.vu.nl, vincent@few.vu.nl, jhulshof@few.vu.nl). The work of the first author was partly supported by an NWO VENI grant. The work of the third author was also supported by the CWI in Amsterdam.
‡Université Bordeaux 1, Mathématiques Appliquées de Bordeaux, 33405 Talence cedex, France (vincent.guyonne@math.u-bordeaux1.fr).

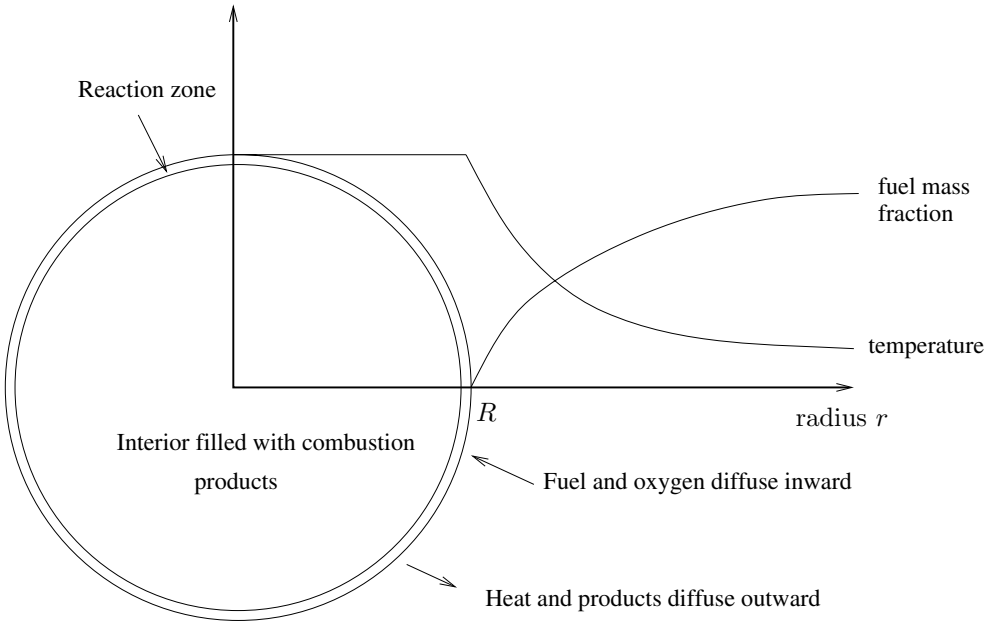


FIG. 1. *Profile of the temperature and the mass fraction variables in the adiabatic case. The radius of the flame ball is denoted by R, corresponding to the flame front.*

flame balls, particularly (radiative) heat losses from the combustion products inside the flame ball. We refer to [3] and references therein; see also the SOFBALL (structure of flame balls at low lewis number) home page [4].

In fact, the radiative transfer of heat in combustion processes taking place in inert not fully transparent media (e.g., dust, porous media, ...) involves both emission and absorption of radiation and may significantly influence the flame temperature (see Figure 2), its propagation speed, and the flammability of the medium itself. This occurs, for instance, in forest fires and fires in confined spaces such as tunnels, and the importance of radiative transfer has been noted and stressed in [5, 6, 7]. In this paper, we concentrate on the effects of radiative transfer on flame balls.

There are two common formulations for modeling combustion processes: the reaction-diffusion and the free boundary formulation. Although both formulations are widely used in the combustion literature, the relation between the two approaches has so far largely been based on numerical simulation and heuristic arguments.

The basic thermo-diffusive model of combustion with simple chemistry is a reaction-diffusion system (RDS) that is written as

$$Y_t = \frac{1}{\mathrm{Le}}\Delta Y - YF(\theta), \tag{1a}$$

$$\theta_t = \Delta\theta + YF(\theta), \tag{1b}$$

where Y denotes the mass fraction of the reactant, θ the temperature, and Le the Lewis number (ratio between conductivity and diffusivity). The function F is an Arrhenius-type reaction rate involving a small parameter ε which is the inverse of the activation energy. The Arrhenius law is often modified by the choice of an ignition temperature, below which the reaction rate is taken to be zero. In this framework, (linearly) unstable flame balls are known to exist. For Lewis number close to unity,

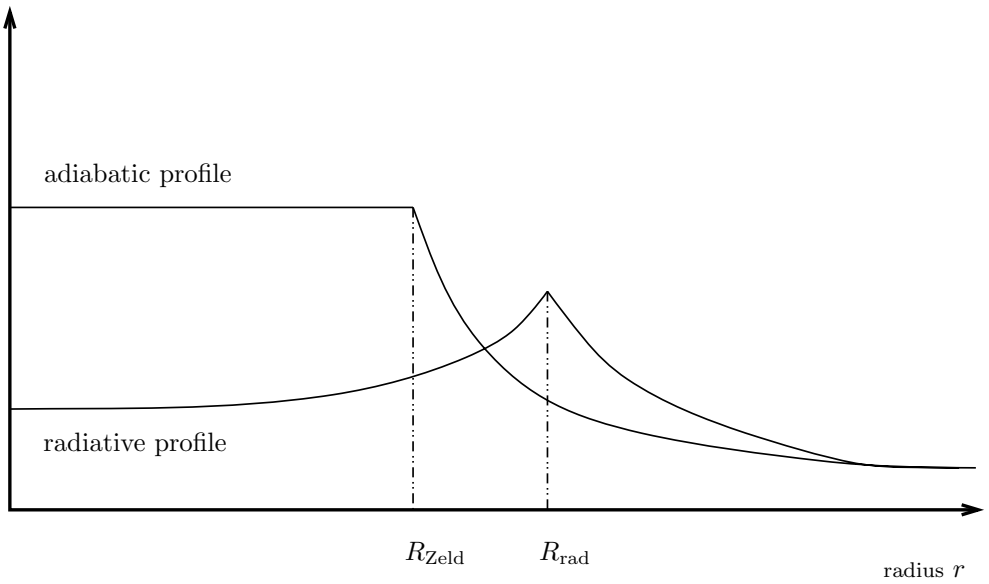


FIG. 2. *Difference of temperature profiles in the adiabatic and in the radiative case.*

the growth of the radius has been described using an integro-differential equation derived formally by Buckmaster, Joulin, and Ronney [8] and rigorously validated by Lederman, Roquejoffre, and Wolansky [9].

When one assumes that the flame occurs in a very thin region, it is quite natural to define a free boundary problem (FBP). Its derivation from the RDS formulation has been justified formally in [10] under the assumption of high activation energy. Its validity is also confirmed by numerical simulations on the RDS, and its great advantage is that several analytical aspects are simpler to treat. The FBP reads as follows:

$$Y_t = \frac{1}{\text{Le}}\Delta Y \qquad \text{for } x \notin R(t), \tag{2a}$$

$$\theta_t = \Delta\theta \qquad \text{for } x \notin R(t), \tag{2b}$$

with

$$[\theta] = Y = 0, \quad -[\theta_n] = \frac{1}{\text{Le}}[Y_n] = F(\theta), \qquad \text{for } x \in R(t), \tag{2c}$$

where $R(t)$ represents the location of the free boundary (the flame front) and brackets denote jumps across the free boundary (in the direction of the normal n). The mass flux into the flame is balanced by (reaction) heat flux coming out of the flame, with a (predominantly) temperature dependent reaction rate. Note that at the flame front we impose the condition that $Y = 0$. Usually, one imposes only that the jump $[Y] = 0$, silently assuming that $Y \equiv 0$ on the burnt side of the flame front. Without such an assumption, the FBP formulation with $[Y] = 0$ instead of $Y = 0$ is underdetermined. As an FBP this model should not be confused with the well-studied model for nearly equidiffusional flames (NEF), which was derived by Sivashinsky by means of an asymptotic analysis, in which he coupled the deviation of the Lewis number from unity to ε, the inverse of the activation energy (see [11, 12]), and derived what is now known as the Kuramoto–Sivashinsky equation.

The remaining step is to incorporate a model for the radiative effects. The first models developed by Buckmaster, Joulin, and Ronney [8, 13] and Shah, Thachter, and Dold [10], for example, are based on rather ad hoc heat loss assumptions (whereas in the present paper we will consider a more thorough radiative transfer model). We would like to recall at this point some results obtained in this context. In [8], heat losses are assumed to occur in the volume enclosed by the flame sheet only. In this framework two stationary solution branches exist, corresponding to small and large radius. Concerning the ensuing stability issues, the authors showed that, provided that the Lewis number is less than unity, all small flames are unstable to one-dimensional (radial) perturbations. Large flames are unstable to three-dimensional perturbations, but only if they have a radius greater than some critical value. Thus there is a band of large flames, lying between the quenching point and unstable flames, that are stable. In [13], the authors extend this result by including the effects of heat loss in the far field (unburned gas), and they conclude that far field losses do not qualitatively change the (stability) properties of the solutions. Finally, in [10], flame balls are studied in a porous medium that serves to exchange heat with the gas, and two heat loss models are considered. One of these treats the heat loss as being constant in the burnt region and linear in the unburned region. The other does not distinguish between burnt and unburnt gas and is based on a (nonlinear) Stefan's law. For both heat loss models, the authors find, again, two branches of solutions of small and large flame balls, respectively. For Lewis number greater than unity the solutions are unstable, while at Lewis number less than unity part of the branch of large flame balls becomes stable, solutions with the nonlinear radiative law being stable over a smaller range of parameters. The stable parameter region increases when the heat capacity of the porous medium is increased. It is clear from the considerations in [8, 10, 13] that the stability properties depend strongly on the Lewis number. More details and a comparison with our model can be found in section 4.

In this paper, we would like to go one step further in the description of the radiative effects and introduce a physically more realistic radiative transfer model. Let us start with a microscopic description of the radiative transfer, which is given by the equation

$$\partial_t I + \Omega \cdot \nabla I = \sigma(B(\nu, \theta) - I),$$

where $I = I(x, t, \Omega, \nu)$ is a total radiative intensity, x the position, t the time, Ω the direction of emission vector, ν the frequency, σ the opacity of the medium, and $B(\nu, \theta)$ the Planck distribution: $B(\nu, \theta) = \frac{2h\nu^3}{c^2}(\exp(\frac{h\nu}{k\theta}) - 1)^{-1}$. Since numerical simulations of this model are very cumbersome, radiation is most commonly described by simplified models, such as the (Milne–)Eddington diffusion equations, valid in the limit of isotropic radiation; the Rosseland model, valid for high opacity media; or the optically thin model, valid for nonabsorbent media [14, 15].

In this paper we adopt the Eddington diffusion model [14, 15, 16, 17, 18, 19], namely,

$$-\nabla(\nabla \cdot q) + 3\alpha^2 q = -\alpha \nabla \theta^4, \tag{3}$$

where q is the radiative flux. Thus, the radiative effects are a direct consequence of temperature variations. Following Joulin and Buckmaster and coworkers [5, 6, 7], these radiative effects couple back to the temperature equation, in which the divergence of the radiative flux appears with coupling constant β, the Boltzmann constant.

Thus β is a measure of the ratio between the radiative and the diffusive flux. For flame fronts, this extended model was proposed and studied in [5, 6, 7], and in [20, 21].

In this paper we study equilibria of the resulting FBP in the radially symmetric case, i.e., steady spherically symmetric flame balls. If we set $r = |x|$, we may thus write the Laplacian operator as $\Delta = \partial_{rr} + \frac{2}{r}\partial_r$, so that the problem can be viewed as a system of ordinary differential equations. Hence, throughout the paper all functions depend on the radial coordinate r only, and they all have zero derivative at $r = 0$. To make the mathematical analysis easier, we do not use the vector equation (3) but work with the scalar equation

$$-\Delta u + 3\alpha^2 u - \alpha\Delta\theta^4 = 0,$$

where $-u = \nabla \cdot q$, the divergence of the radiation flux as it appears in the modified temperature equation. This equation is nonlinear and therefore does not allow us to compute explicit solutions for the full problem defined below. On the other hand, if one considers θ instead of θ^4, i.e., the "linear" problem, one can write down explicitly the solution, and in section 4 we compare the solutions of the linear and nonlinear equations. The FBP reads

$$\frac{1}{\mathrm{Le}}\Delta Y = 0 \qquad \text{for } r \neq R, \tag{4a}$$

$$-\Delta\theta - \beta u = 0 \qquad \text{for } r \neq R, \tag{4b}$$

$$-\Delta u + 3\alpha^2 u - \alpha\Delta\theta^4 = 0. \tag{4c}$$

Equation (4c) is satisfied in the whole space in the sense of the distributions (and classically for $r \neq R$). The jump conditions at $r = R$ are

$$[\theta] = Y = 0, \qquad -[\theta_r] = \frac{1}{\mathrm{Le}}[Y_r] = F(\theta(R)), \tag{4d}$$

with u being continuous, while the size of the jump in u_r follows automatically from (4c) and (4d). The asymptotic boundary conditions are

$$Y \to Y_f, \quad \theta \to \theta_f, \quad u \to 0 \quad \text{as } r \to \infty. \tag{4e}$$

The parameters θ_f and Y_f denote the temperature and the mass fraction far away in the fresh region. We recall that R is the free boundary variable corresponding to the flame front, and that $F(\theta(R))$ is the reaction rate evaluated at $r = R$. Note that we will not specify the reaction rate and work only with general reaction rates F. The reason is that, to prove existence properties, we need to know only that F is a positive function of the temperature at the flame front. The main result of this paper is the following.

THEOREM 1 (existence). *Let $\alpha \geq 0$, $\beta \geq 0$, let F be continuous and positive, and let $\theta_f > 0$, $Y_f > 0$. Then there exists a radial solution $(\theta(r), Y(r), u(r), R)$ to (4). Moreover, for generic choices of the parameters the number of solutions is odd.*

Let us briefly outline the method of the proof. We first observe that the FBP formulation, with the Arrhenius law acting only on the flame front, allows us to decouple (4a) for Y from the two others, (4b) and (4c). The only bounded function Y which solves (4a) and satisfies $Y = 0$ at $r = R$ and $Y \to Y_f$ as $r \to \infty$ is given by

$$Y(r) = \begin{cases} 0 & \text{for } r \leq R, \\ Y_f\left(1 - \frac{R}{r}\right) & \text{for } r > R. \end{cases} \tag{5}$$

Here R is still unknown. We now drop one of the free boundary conditions, namely the last equality in (4d), and solve the problem with R as a parameter. In other words, we drop the reaction rate and fix R. The next theorem provides us with a unique solution of the resulting reduced problem, parameterized by the now prescribed flame ball radius R.

THEOREM 2 (uniqueness and existence for R fixed). *Fix $R > 0$ and let $\alpha \geq 0$, $\beta \geq 0$, $\theta_f > 0$, $Y_f > 0$. Then there exists a unique solution $(\theta_R(r), Y_R(r), u_R(r))$ to (4), with $\theta > 0$.*

To prove this theorem, we will first decompose the temperature as $\theta = \theta_h + w$, where θ_h is an adiabatic profile with an arbitrary fixed radius R. Because we seek radial solutions, we can explicitly compute θ_h, namely,

$$\theta_h = \theta_f + \frac{Y_f}{\mathrm{Le}} \min\left(1, \frac{R}{r}\right). \tag{6}$$

Then we show that w satisfies a nonlinear elliptic equation defined on all $\mathbb{R}$. Thus θ_h is the temperature component of the solution of the reduced problem with given R, in the case that $\beta = 0$. The subscript h stands for "homogeneous," because θ_h is the solution of the homogeneous part of (4b) which satisfies the jump condition. The other part w in the splitting will then be the solution of the full inhomogeneous equation (4b), which is smooth (i.e., $[w] = [w_r] = 0$) across $r = R$. Hence w satisfies the equation $-\Delta w = u$ globally, just as u solves (4c) globally, in the sense of the distributions. To solve this equation, we consider the problem on a bounded domain, more precisely on a ball $B_\rho = B(0, \rho) \subset \mathbb{R}^3$, with $\rho > R$ large. Using sub- and supersolution arguments, one obtains a solution on the bounded domain. Then we let $\rho \to \infty$, and, by a diagonal process, this leads to a solution on $\mathbb{R}^3$. Uniqueness is proved using classical arguments (see section 2 for details).

Remark 1. We consider only positive θ; the solution $\theta_R(r)$ depends continuously on R, and θ_R is bounded between θ_f and $\theta_f + \frac{Y_f}{\mathrm{Le}}$.

Going back to the proof of Theorem 1, we need to find a value of R for which $\theta_R(r)$ satisfies the final free boundary condition in (4d). As we know Y explicitly, we are left with one "algebraic" equation,

$$\frac{Y_f}{\mathrm{Le}} \frac{1}{R} = F(\theta_R(R)). \tag{7}$$

Thus the reaction rate F plays a role in the analysis only at this final stage. From Figure 3 we can easily see that (7) has at least one solution (see section 2 for more details). This ends the proof of Theorem 1.

Remark 2. When solving (7), one can easily see from Remark 1 and Figure 3 that the radiative radius R_{rad} is bounded between two values. If the reaction rate F is an increasing function of the temperature, as is usually the case, the lower bound on the flame radius is given by the adiabatic or Zeldovich radius $R_{\mathrm{Zeld}} = \frac{Y_f}{\mathrm{Le}} \frac{1}{F(\theta_f + Y_f/\mathrm{Le})}$ (i.e., the radius in the absence of radiative effects; see section 2 for more details), whereas the upper bound is $\frac{Y_f}{\mathrm{Le}} \frac{1}{F(\theta_f)}$.

In section 3 we examine limit cases of problem (4). The cases $\alpha \to \infty$ with β fixed, and $\alpha \to 0$ (or transparent limit) lead to the adiabatic case and are the easiest to justify. A more subtle analysis is needed to treat the cases $\beta \to \infty$ (large Boltzmann limit) and $\alpha \to 0$ supposing $\alpha\beta = \chi$ fixed (transparent limit combined with large Boltzmann numbers). In the large Boltzmann limit, we prove that the temperature profile converges to a constant profile, namely to the fresh temperature θ_f. On the

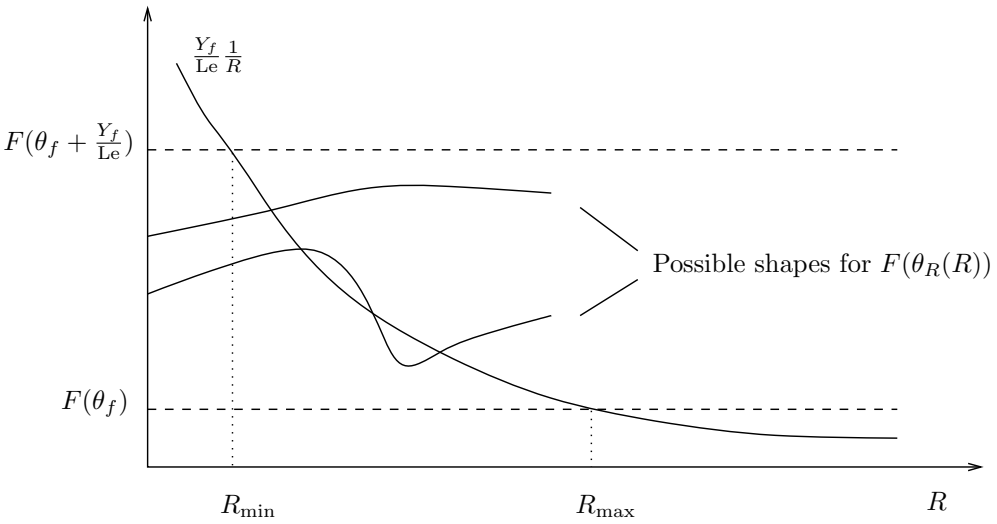


FIG. 3. *Sketch of the graphs occurring in* (7). *The dashed lines are the bounds on* $F(\theta_R(R))$*; they are depicted here for the case of increasing* F.

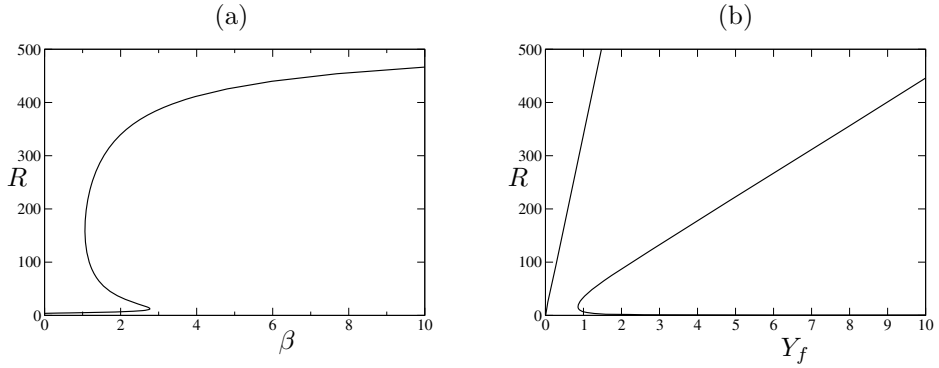


FIG. 4. *Bifurcation curves exhibiting turning points:* (a) *with* β *used as bifurcation parameter;* (b) *with* Y_f *as the parameter.*

other hand, in the transparent limit combined with large Boltzmann numbers, the temperature profile does not converge to a constant profile, but to a radiative one (cf. Figure 2).

Finally, in section 4 we compare the analytic expressions in the asymptotic limits to numerical computations for the full problem. We also make a comparison with analytic calculations for a "linearized" system; see section 4.1 for details. As an example, in Figure 4(a) we depict a typical bifurcation diagram, where β is used as the bifurcation parameter. For a range of parameter values there are three distinct flame ball solutions (for the adiabatic (nonradiative) problem there is always only one solution). Examining the corresponding solution profiles, the upper branch turns out to be physically irrelevant, since the temperature profile is almost identically equal to θ_f. In Figure 4(b) the fuel mass fraction Y_f in the fresh region is used as a bifurcation parameter. Again, multiple solutions are obtained, on two disconnected branches.

2. Existence of solutions. In this section we sketch the arguments that lead to Theorem 2 and subsequently to Theorem 1. All remaining details of the proofs are provided in the appendix. We recall that in order to fix R, we consider problem (4), and we drop the equation involving the reaction rate F in (4d). The expression for Y is of course given by (5). Theorem 1 follows immediately from Theorem 2 when we combine it with the fact that the algebraic equation (7) has a solution.

To begin with, we reduce (4b) and (4c) to one elliptic equation. To do this, we first need to introduce a splitting of the solution θ that we are looking for, writing

$$\theta = \theta_h^R + w. \tag{8}$$

Here θ_h^R is the solution of

$$-\Delta\theta_h^R = 0 \qquad \text{for } r \neq R, \tag{9a}$$

with jump conditions

$$[\theta_h^R] = 0, \quad -\left[\frac{\partial\theta_h^R}{\partial r}\right] = \frac{1}{\mathrm{Le}}\left[\frac{\partial Y}{\partial r}\right] \quad \text{at } r = R, \tag{9b}$$

and the asymptotic boundary condition

$$\theta_h^R \to \theta_f \quad \text{as } r \to \infty. \tag{9c}$$

We note that (9) can be solved explicitly, where θ_h^R is given by (6). The advantage of the splitting (8) is that w must have zero jumps,

$$[w] = [w_r] = 0,$$

and $w \to 0$ as $r \to \infty$. Hence it must be a solution of

$$-\Delta w = \beta u \tag{10}$$

on the *whole space* in the sense of the distributions.

Next we observe that (4c) implies that

$$u = \alpha(3\alpha^2 - \Delta)^{-1}\Delta\theta^4,$$

which expresses u in terms of θ^4 by means of the bounded operator

$$\alpha(3\alpha^2 - \Delta)^{-1}\Delta = \alpha\Delta(3\alpha^2 - \Delta)^{-1},$$

which operates from $L^\infty \to L^\infty$. Note that the Laplacian and its resolvent commute because $3\alpha^2 > 0$. Combining this with (10), it follows that

$$\Delta(w + \alpha\beta(3\alpha^2 - \Delta)^{-1}\theta^4) = 0,$$

whence, since both w and θ^4 are bounded, $w + \alpha\beta(3\alpha^2 - \Delta)^{-1}\theta^4$ must be a constant:

$$w + \alpha\beta(3\alpha^2 - \Delta)^{-1}\theta^4 = C.$$

Subtracting θ_f^4 from θ^4 only changes the constant. Moreover, $\theta^4 - \theta_f^4$ has zero limit at infinity ($r \to \infty$), a property which is preserved by the resolvent $(3\alpha^2 - \Delta)^{-1}$, and

also $w \to 0$ as $r \to \infty$. Thus $w + \alpha\beta(3\alpha^2 - \Delta)^{-1}(\theta^4 - \theta_f^4) = 0$. Applying $(3\alpha^2 - \Delta)^{-1}$ to both sides, we arrive at

$$(3\alpha^2 - \Delta)w + \alpha\beta\left((w + \theta_h^R)^4 - \theta_f^4\right) = 0, \tag{11a}$$

which again should hold globally, with asymptotic boundary condition

$$w \to 0 \quad \text{as } r \to \infty. \tag{11b}$$

We note that $w(r)$ is a solution of a second order ordinary differential equation (globally). Thus it has zero jumps $[w]$ and $[w_r]$ at $r = R$. We have split the problem for θ, which was inhomogeneous because of the jump in $r = R$ and the nonzero limit as $r \to \infty$, on the one hand, and the presence of βu in (4b), on the other, into two parts. The first part, θ_h^R, takes care of the jumps and limits, while the second, w, corresponds to the inhomogeneous term βu in (4b).

The crucial idea in the existence proof is the above reduction of the system of two equations (4b) and (4c) to one elliptic equation (11a). Therefore, existence of a solution pair (θ, u) for (4b) and (4c) is equivalent to the existence of a solution w for problem (11). In order to solve this problem, we first consider (11a) on a ball $B_\rho \subset \mathbb{R}^3$, with the boundary condition (11b) being replaced by $w = 0$ on ∂B_ρ. Then, using classical monotone iteration methods, we can prove existence on this finite domain. Finally, we take the limit $\rho \to \infty$ and arrive at the following result.

LEMMA 3. *For R fixed, there exists a unique solution w_R of problem* (11) *satisfying the bound*

$$-\frac{Y_f}{\mathrm{Le}} \min\left(1, \frac{R}{|x|}\right) \le w \le 0. \tag{12}$$

The solution w_R is $C^2(\mathbb{R})$, radially symmetric, monotonically increasing in $|x|$, and depends continuously on R.

Thus, this proves Theorem 2 and shows that, omitting the reaction rate from the problem formulation, there exists for every $R > 0$ a unique solution triple (θ, Y, u) with $\theta > 0$. It remains, in order to prove Theorem 1, to solve (7) with $\theta_R(R)$ given by Theorem 2. Lemma 3 shows that $\theta_R(R)$ depends continuously on R. Moreover, in view of estimate (12), $\theta_f \le \theta_R(R) \le \theta_f + \frac{Y_f}{\mathrm{Le}}$. Hence, Theorem 1 is an easy consequence of the intermediate value theorem applied to (7). All details of the proof, as well as additional estimates, can be found in the appendix.

3. Limit cases of the radiative parameters. In this section we examine some singular limit cases. We recall that we introduced the splitting $\theta = \theta_h^R + w$. Throughout this section, we consider a pair $(\theta_{\mathrm{par}}, R_{\mathrm{par}})$ depending on some parameters, and we seek a limit. Let us start by noting the following.

Remark 3. As R_{par} lies in a compact set (see Remark 9), one can extract a subsequence converging to a limit, called R. Along the subsequence, $\theta_h^{R_{\mathrm{par}}}$ converges to θ_h^R (uniformly).

3.1. The limit case $\alpha \to \infty$ with β fixed. The limit $\alpha \to \infty$, β fixed is usually called the optically thick limit for an opaque medium. In this limit the effect of the radiation is lost. Indeed, we have the following claim.

LEMMA 4. *The solution w of problem* (11) *converges to zero uniformly as $\alpha/\beta \to \infty$.*

As a consequence of this lemma, and in view of (19), the flame ball solution has a temperature profile that converges to the Zeldovich solution, and also the flame ball radius converges to the Zeldovich radius as $\alpha/\beta \to \infty$.

Proof. We simply modify the subsolution in the proof of Lemma 9 in such a way that it pushes the solution obtained in Lemma 3, and thereby w itself, to zero. A negative constant w is a subsolution, provided

$$3\alpha^2 w + \beta\alpha((\theta_h^R + w)^4 - \theta_f^4) \le 0.$$

This is certainly the case if

$$\left(\theta_f + \frac{Y_f}{\mathrm{Le}} + w\right)^4 - \theta_f^4 = -\frac{3\alpha}{\beta} w,$$

which has a unique solution $w \in (-\frac{Y_f}{\mathrm{Le}}, 0)$, which is easily seen to converge to zero as $\alpha/\beta \to \infty$. This completes the proof. ∎

Remark 4. Note that the limit is the same as the one for α fixed and $\beta \to 0$, i.e., radiative flux negligible with respect to convective flux.

3.2. The transparent limit $\alpha \to 0$ with β fixed. Surprisingly, as opposed to the traveling wave case (see [20]), this limit also reproduces the adiabatic (Zeldovich) flames. As in the previous section we have the next claim.

LEMMA 5. *The solution w of problem* (11) *converges to zero uniformly if $\alpha \to 0$ with β fixed.*

Proof. We have, in view of (12),

$$-\Delta w = -3\alpha^2 w - \alpha\beta\left((\theta_h^R + w)^4 - \theta_f^4\right) \to 0$$

uniformly, as $\alpha \to 0$ and $\alpha\beta \to 0$. Also, again because of (12), w is uniformly small for large r. By the maximum principle for the Laplacian, this implies that $w \to 0$ uniformly as $\alpha \to 0$ and $\alpha\beta \to 0$. ∎

3.3. Large Boltzmann numbers $\beta \to \infty$ with α fixed. With large Boltzmann numbers the solution loses its physical meaning because the temperature profile becomes flat. We have the following result.

LEMMA 6. *For α fixed and $\beta \to \infty$ the temperature profile θ converges to θ_f uniformly.*

Proof. Let us set $w_n = w_{\beta_n}$, with $\beta_n \to \infty$ as $n \to \infty$. We are looking for a limit of the problem

$$-\Delta w_n = -3\alpha^2 w_n - \alpha\beta((\theta_h^{R_n} + w_n)^4 - \theta_f^4), \tag{13}$$

with asymptotic boundary condition $w_n \to 0$ as $|x| \to \infty$.

Writing the weak formulation of (13) and dividing by β_n, we find that, for any test function $\varphi \in C_c^\infty([0,\infty))$, in view of (12),

$$\int \left((\theta_h^{R_n} + w_n)^4 - \theta_f^4\right)\varphi = -\frac{1}{\beta_n}\int 3\alpha^2 w_n\varphi + \frac{1}{\alpha\beta_n}\int w_n\Delta\varphi \to 0 \qquad \text{as} \quad n \to \infty.$$

By the bound (12), the functions

$$(\theta_h^{R_n} + w_n)^4 - \theta_f^4 \tag{14}$$

are nonnegative. Thus we may conclude that they converge to the zero function in L^1_{loc} strongly. Next, we rewrite (14) as

$$G(Z_{R_n} + w_n),$$

where $G(\xi) = (\theta_f + \xi)^4 - \theta_f^4$ and $Z_R(r) = \frac{Y_f}{\text{Le}} \min\left(1, \frac{R}{r}\right)$.

Again in view of (12), the variable $\xi = Z_R + w$ ranges between 0 and Z_R. In this range G' is positive and bounded away from zero and infinity. Consequently, the functions $Z_{R_n} + w_n$ also converge strongly to zero in L^1_{loc}. However, Z_{R_n} converges if we restrict to a further subsequence, along which R_n converges, not only in L^1_{loc} but also in L^∞.

We claim that for any sequence R_n bounded away from zero and infinity, and for any sequence $\beta_n \to \infty$, the corresponding solutions w_n of (11) have the property that $\theta_n = \theta_h^{R_n} + w_n \to \theta_f$ uniformly on $[0, \infty)$. To prove this, we apply the following simple lemma.

LEMMA 7. *Let f_n and g_n be functions on $\mathbb{R}_+$ such that*

- *$f_n + g_n \geq 0$,*
- *$f_n + g_n \to 0$ in $L^1(0, \rho)$ for all $\rho > 0$,*
- *$f_n' \geq -C$ in a weak sense,*
- *$g_n' \geq 0$;*

then $f_n + g_n \to 0$ in $L^\infty(0, \rho)$ for all $\rho > 0$.

Proof. The proof is immediate from the estimate

$$f_n(r) + g_n(r) \geq f_n(r_0) + g_n(r_0) - C(r - r_0)$$

if $r > r_0 > 0$. ∎

This lemma applies to $f_n = Z_{R_n}$ and $g_n = w_n$, which is monotone by Lemma 3. As before, we conclude that $\theta_n - \theta_f \to 0$ in $L^\infty(\mathbb{R})$. ∎

3.4. The transparent limit combined with large Boltzmann numbers: $\alpha \to 0$ with $\alpha\beta = \chi$ fixed. Finally, we consider the limit $\alpha \to 0$, $\alpha\beta = \chi > 0$ fixed, which was also treated in the traveling wave context; see [20, 21]. We show that in this limit solutions of the radiative transfer problem converge to solutions of a radiative heat loss problem, where θ solves

$$\Delta\theta - \chi(\theta^4 - \theta_f^4) = 0, \qquad r \neq R,$$

and R is the flame radius of the limit solution. This will follow along the same lines as in the previous sections from the following.

LEMMA 8. *In the limit $\alpha \to 0$ with $\alpha\beta = \chi > 0$ fixed, the solution w of (11) converges along subsequences to a solution of*

$$-\Delta w + \chi((\theta_h^R + w)^4 - \theta_f^4) = 0, \tag{15}$$

with $w \to 0$ as $r \to \infty$.

Proof. In view of the a priori bounds on w and on R, and in view of Remark 7, we know that w, w', and w'' are (uniformly) equicontinuous on bounded balls. This suffices again to conclude that, as $\alpha \to 0$, a subsequence converges in $C^2(\overline{B_\rho})$, for any $\rho > 0$, to a solution of (15). As before, a diagonal process finishes the proof. ∎

Remark 5. In this limit w remains nontrivial in the sense that it does not coincide with one of the bounds in (12). Thus, in the limit we will have a bifurcation diagram given by

$$\frac{Y_f}{\text{Le}R} = F\left(\theta_f + \frac{Y_f}{\text{Le}} + w(R)\right),$$

and the right-hand side truly depends on R.

4. Numerical calculations. In this section we examine the flame balls numerically. We will compare the outcome of the computations with analytic formulas for the "linearized" problem, which we present below.

4.1. Analytic solutions for the linear case. In this first part, we derive a bifurcation diagram equation for the linear case. Namely, we still consider problem (4), except that (4c) is replaced by the linear equation

$$-\Delta u + 3\alpha^2 u - \alpha\Delta\theta = 0. \tag{16}$$

We can compute explicit formulas for the temperature θ and the variable u. To simplify the notation we introduce

$$\mu = \mu_{\alpha\beta} = \sqrt{3\alpha^2 + \alpha\beta}.$$

Then

$$\theta(r) = \begin{cases} \dfrac{B_1}{r}\sinh(\mu r) + B_3 + \theta_f & \text{for } r \le R, \\ \dfrac{B_2}{r}\exp(-\mu r) + \dfrac{B_3 R}{r} + \theta_f & \text{for } r > R, \end{cases}$$

where the constants are given by

$$B_1 = \frac{\alpha\beta Y_f}{\mathrm{Le}\mu^3}\exp(-\mu R), \qquad B_2 = \frac{\alpha\beta Y_f}{\mathrm{Le}\mu^3}\sinh(\mu R), \qquad B_3 = \frac{3\alpha^2 Y_f}{\mathrm{Le}\mu^2}.$$

The expression for u is

$$u(r) = \begin{cases} -\dfrac{B_1\mu^2}{\beta r}\sinh(\mu r) & \text{for } r \le R, \\ -\dfrac{B_2\mu^2}{\beta r}\exp(-\mu r) & \text{for } r > R. \end{cases}$$

Finally, the equation that fixes the flame radius R, and that determines the bifurcation diagrams, reads

$$F\left(\frac{\alpha\beta Y_f}{2\mu^3 \mathrm{Le} R}\left[1 - 2\mu R - e^{-2\mu R}\right] + \frac{Y_f}{\mathrm{Le}} + \theta_f\right) = \frac{Y_f}{\mathrm{Le} R}. \tag{17}$$

4.2. Bifurcation diagrams. Let us turn to the numerical investigation of the problem. Since we know from Theorem 1 that a solution is uniquely determined by its flame radius R, we exhibit diagrams in which the flame ball is represented by R along the vertical axis, and the horizontal axis is reserved for a control parameter such as Y_f or one of the radiative parameters α or β.

We can do numerical simulations only on bounded domains, so we choose a large ball B_ρ on which we impose Dirichlet boundary conditions, as used in the existence proof. From the proof of Lemma 3 we know that the solution on the bounded ball B_ρ approaches the solution on $\mathbb{R}^3$ as $\rho \to \infty$, and in the numerical calculations we always make sure that $\rho \gg R$. Since the flame balls are radially symmetric, the problem is thus reduced to a boundary value problem for an ordinary differential equation, and we use the continuation software [22] to compute the bifurcation diagrams.

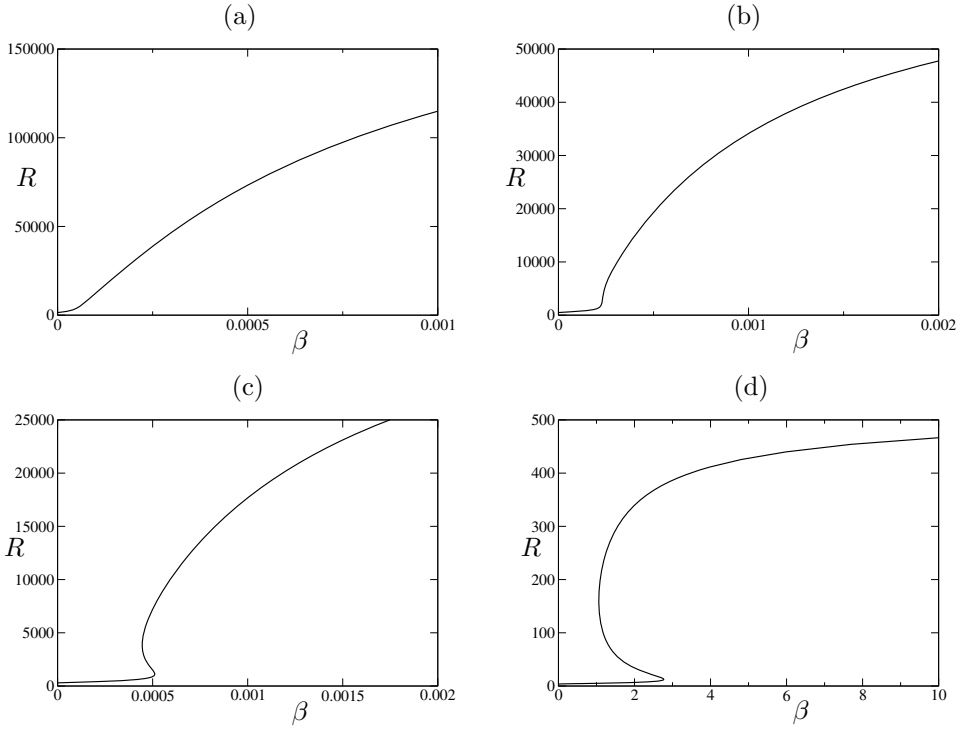


FIG. 5. *Bifurcation diagrams with β as the bifurcation parameter for* (a) $A = 0.1$, (b) $A = 0.3$, (c) $A = 0.5$, (d) $A = 40$.

We need an explicit expression for the reaction rate. Following the literature, e.g., [23, 24], we choose a simple Arrhenius law

$$F(\theta(R)) = A \exp\left(-\frac{1}{\varepsilon\theta(R)}\right), \tag{18}$$

where ε is a normalized inverse activation energy and $A > 0$ is the pre-exponential factor. Next we must choose values for the parameters. Unless mentioned otherwise, in all computations we take

$$\theta_f = 1, \quad Y_f = 1, \quad \mathrm{Le} = 1, \quad \varepsilon = 0.1, \quad A = 40, \quad \alpha = 10^{-4}, \quad \beta = 2.$$

In fact, the parameters Y_f and Le appear in the stationary problem only in the combination Y_f/Le, so we will use this ratio as a parameter in what follows.

In Figure 5 bifurcation diagrams are shown with β as the bifurcation parameter, for various values of the pre-exponential constant A. We see that a turning point appears in the bifurcation diagram as we increase A. Hence, for A sufficiently large there is a range of values of the Boltzmann number β for which there exist multiple stationary flame balls. Increasing A corresponds to making the function in the Arrhenius law (18) steeper. In the context of traveling wave solutions (moving flame fronts) it was already observed (and extensively analyzed) that a steeper Arrhenius law may lead to turning points in bifurcation diagrams; see [21]. We note that the presence of turning points is due to the radiative effects being incorporated in the model, since

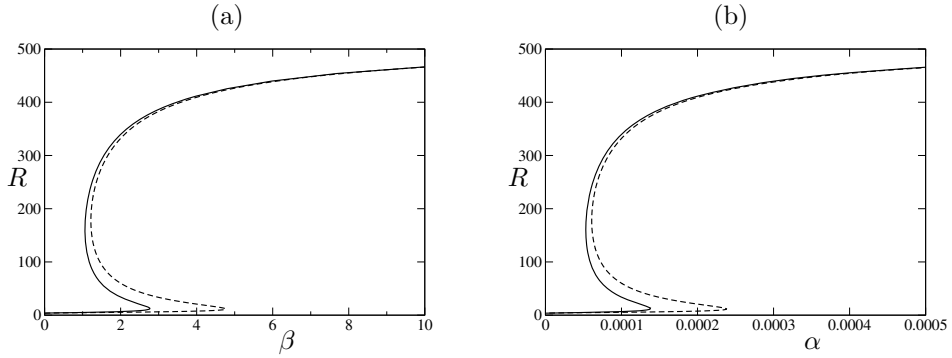


FIG. 6. *Comparison between the nonlinear problem (solid line) and the linearized one (dashed line) with* (a) β *and* (b) α *as the bifurcation parameter.*

uniqueness of the adiabatic flame ball implies the absence of turning points in the adiabatic problem.

Figure 5 also corroborates the study of the limit cases in section 3. In the limit $\beta \to 0$ the flame radius R converges to the Zeldovich radius (the minimal possible radius); see Remarks 2 and 4. On the other hand, as proved in Lemma 6, in the limit $\beta \to \infty$ the temperature profile converges to θ_f, which corresponds to the maximal radius (see again Remark 2).

To make a useful comparison between the full, nonlinear problem and the "linearized" equation (16) from section 4.1, we need to linearize the term θ^4 around some *characteristic temperature* θ_c: $\theta^4 \approx \theta_c^4 + 4\theta_c^3(\theta - \theta_c)$. Introducing the rescaled variable $\tilde{u} = \beta u$, we then arrive at the system

$$\begin{cases} \Delta\theta + \tilde{u} = 0, \\ \Delta\tilde{u} - 3\alpha^2\tilde{u} + 4\alpha\beta\theta_c^3\Delta\theta = 0. \end{cases}$$

Therefore, solutions of the full problem should be compared to solutions of the linearized problem for $\tilde{\beta} = 4\beta\theta_c^3$. Hence, in all figures, for the (dashed) curves representing the analytic expression (17) for the linearized problem, the scaling factor $4\theta_c^3$ is taken into account. As the characteristic temperature we simply adopt $\theta_c = \theta_f$ throughout.

In Figure 6 we compare the outcome of the numerical computations on the nonlinear problem with the analytic expression for the linearized one, using both α and β as bifurcation parameters. In Figure 6(a) we see that the nonlinear and linear problems are qualitatively very similar. In the limit $\beta \to \infty$ we know from Lemma 6 that $\theta \to \theta_f$ uniformly, so our choice of $\theta_c = \theta_f$ leads to quantitative agreement for large β. In the adiabatic limit, i.e., $\beta \to 0$, the solution becomes independent of the radiative effect, irrespective of the equations being linear or not.

Figure 6(b) is, up to a scaling in the horizontal direction, the same as Figure 6(a). The reason is that α is so small that α^2 is negligible compared to $\alpha\beta$, so that to good approximation the solution in this parameter regime depends only on the combination $\alpha\beta$.

From Lemma 4 we know that for large α the solution converges to the adiabatic one, and the radius decreases towards the Zeldovich radius. Indeed, when we continue the bifurcation curve of Figure 6(b) for larger values of α we obtain Figure 7, where we need three different scales to be able to see the full picture. In accordance with

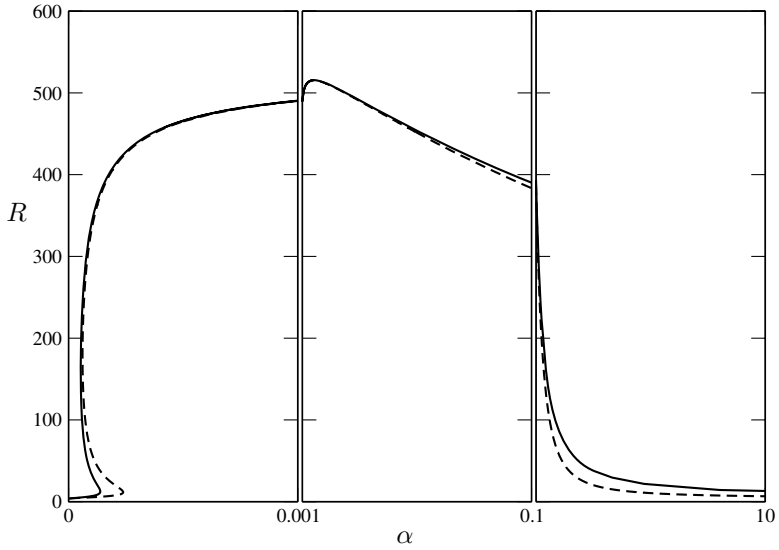


FIG. 7. *The complete* (α, R) *bifurcation diagram, on three different scales.*

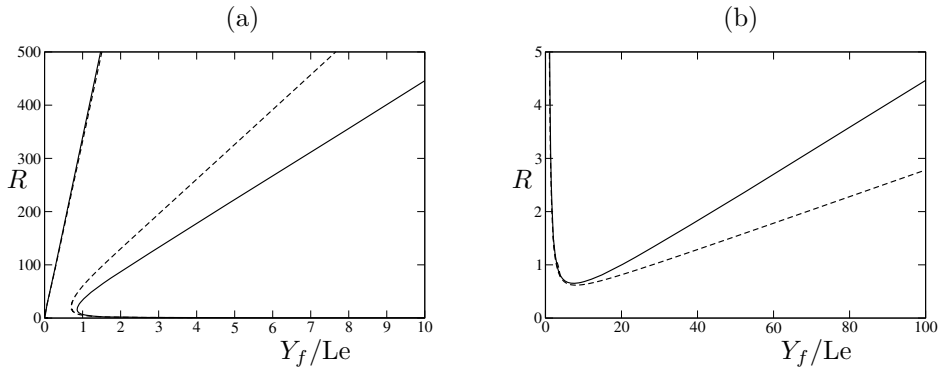


FIG. 8. *Comparison between the nonlinear problem* (*solid line*) *and the linearized one* (*dashed line*) *with* Y_f/Le *as bifurcation parameter, depicted at two different scales.*

Lemmas 4 and 5, the flame radius R tends to the $\frac{Y_f}{\mathrm{Le}} \frac{1}{F(\theta_f + Y_f/\mathrm{Le})}$ in both limits $\alpha \to 0$ and $\alpha \to \infty$, while it makes an excursion near $\frac{Y_f}{\mathrm{Le}} \frac{1}{F(\theta_f)}$ in between.

In Figure 8 we employ Y_f/Le as the bifurcation parameter. For Y_f/Le sufficiently large there are again three solutions, and we need to examine two different scales to see them. The linearized problem does not mimic the nonlinear one too closely, since for large values of Y_f/Le the temperature varies too much to be adequately represented by the characteristic temperature θ_c.

The linear behavior of the curves in Figure 8 can be understood from the fact that α is chosen very small. In this asymptotic regime it is not hard to calculate the slopes for the linearized problem. In fact $R \sim C_i \frac{Y_f}{\mathrm{Le}}$, where the two slopes $C_{1,2}$ in Figure 8(a)

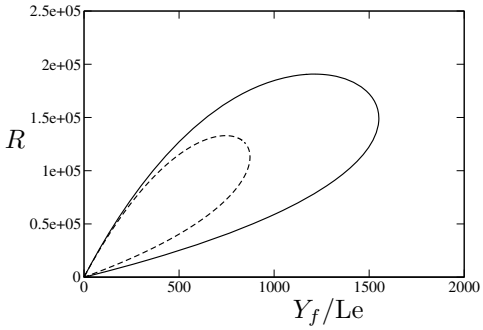


FIG. 9. *The global picture of the* $(Y_f/\mathrm{Le}, R)$ *bifurcation diagram. Note that there is a branch of solutions* (*almost*) *coinciding with the horizontal axis.*

are approximately given by the two largest solutions of $F(\frac{1}{2\sqrt{\alpha\beta}}C^{-1}+\theta_f) = C^{-1}$, while the slope C_3 in Figure 8(b) is approximately equal to A^{-1}. Of course, the reason we can determine these slopes is that we have the explicit expression (17) for the bifurcation curve in the linearized problem. For the nonlinear problem, determining the slopes is an exercise in asymptotic analysis that falls outside the scope of this paper. Note that near the origin the slope is given by $F(\theta_f)^{-1}$ for both the linearized and the nonlinear problem.

Finally, while Figure 8(a) suggests that there are two disconnected solution branches, the global bifurcation diagram depicted in Figure 9 shows that these branches are in fact connected to each other for large values of R and Y_f/Le.

The multiplicity of flame ball solutions, also found in heat loss models [8, 10, 13], leads to questions about stability, which we intend to study in future work. As in the heat loss case, it is in these stability issues that the Lewis number, which plays a somewhat subdued role in the analysis of the stationary problem, will be crucial.

Multiplicity of solutions and stability in the heat loss case are discussed in the introduction. At this point we would like to make some comparisons with our results using the Eddington equation for radiative transfer. One may notice that for some range of parameter values there are three branches of stationary solutions, compared to two in the heat loss models; see, for example, Figure 5. Here one needs to keep in mind that for large R (i.e., the extreme part of the upper branch) the temperature profile is almost flat and therefore does not correspond to a physical flame ball. On the other hand, near the turning point, there are truly three stationary flame balls, of which we expect the *middle* one (see also Figure 6) to be stable, at least in some range of the parameters. In the bifurcation diagram in Figure 8 we may also expect stability of the middle branch. Detailed analysis of stability is the subject of current research. Some preliminary *in*stability results have been obtained in [25].

5. Conclusion. Radiation can significantly influence combustion processes. In this paper we investigate a free boundary model for combustion in a gaseous mixture, where we couple the usual diffusion equations to the radiation field. The radiation itself is described by the Eddington equation, which models radiative transfer in a dusty medium under (near) isotropic conditions. This model thus incorporates both emission and absorption of radiation, in contrast to the usual simplified heat loss models; cf. [5, 10]. Mathematically, this leads to the addition of an elliptic equation

describing the radiation field, which is coupled to the (parabolic) diffusion equations.

In this context we prove the existence of radially symmetric stationary solutions, or *flame balls*, which are physically observed in microgravity environments [3]. We find that a solution exists for any combination of the parameters in the model. Since we consider a free boundary model, determining the radius of the flame is part of the problem. Our strategy is to split the analysis into two parts. First we fix the free boundary and solve an elliptic problem on a fixed domain. Subsequently, we solve the remaining algebraic equation to select the correct flame radius.

Having proved the existence of stationary flame balls, we then turn our attention to asymptotic regimes of the radiative parameters, namely the opacity α of the medium and the Boltzmann number β. In both the limits $\alpha \to 0$ and $\alpha \to \infty$ we recover the adiabatic (nonradiative or "Zeldovich") flames. The same limit is obtained in the limit $\beta \to 0$, whereas when $\beta \to \infty$ the temperature profile becomes flat. The limit $\alpha \to 0$, $\beta \to \infty$ with fixed $\alpha\beta$, leads to a nontrivial limit problem with a truly radiative asymptotic profile.

Finally, by using numerical computations and by examining analytically a "linearized" problem, we investigate the multiplicity of solutions (for fixed parameter values). We find large parameter regimes where multiple stationary flame balls exist. This of course raises interesting stability questions, which we plan to address in a forthcoming paper (extending the work of Buckmaster, Joulin, and Ronney [8] on the heat loss case). We expect the Lewis number Le, which is of minor importance for the stationary problem, to play a crucial role in the stability issues for radiative flame balls.

Appendix. Existence proof. In this appendix we collect the details of the proofs of the statements in section 2, particularly Lemma 3. We also provide additional uniform estimates on the function w.

A.1. Existence on a bounded domain. Let us consider (11) on a ball $B_\rho = B(0,\rho) \subset \mathbb{R}^3$, the boundary condition (11b) being replaced by $w = 0$ on ∂B_ρ. We assume $\rho > R$.

LEMMA 9. *For fixed $0 < R < \rho$, there is a unique solution w of* (11a) *with $\theta = \theta_h^R + w \geq 0$ on B_ρ and $w = 0$ on ∂B_ρ. The solution is radial, and as such it belongs to $C^2([0,\rho])$ as well as to $C^2(\overline{B_\rho})$. It satisfies the bounds*

$$-\frac{Y_f}{\mathrm{Le}} \min\left(1, \frac{R}{|x|}\right) \leq w \leq 0.$$

Remark 6. The estimate (12) is independent of the parameters α and β. It provides us with a uniform estimate on the decay rate of w towards zero as $r \to \infty$.

Proof. We first establish the existence of w. The function $\overline{w} \equiv 0$ is a supersolution of (11a) with zero Dirichlet boundary data, because substituting $w = \overline{w} \equiv 0$ into (11a), we end up with $\alpha\beta((\theta_h^R)^4 - \theta_f^4) > 0$. On the other hand, the function $\underline{w} = -\frac{Y_f}{\mathrm{Le}} \min(1, \frac{R}{r})$ is a subsolution: it is negative in $r = \rho$, and substituting $w = \underline{w}$, we obtain $3\alpha^2\underline{w} - \Delta\underline{w}$. The first term is negative, the latter too, but in the sense of the distributions. More precisely, $-\Delta\underline{w}$ is a negative "Dirac" measure supported on $r = R$. It is straightforward to mollify $\underline{w}$ into a family of smooth subsolutions $\underline{w}^\varepsilon$ with $\underline{w}^\varepsilon \to \underline{w}$ uniformly as $\varepsilon \to 0$, and $\underline{w}^\varepsilon \equiv \underline{w}$ outside the interval $(R-\varepsilon, R+\varepsilon)$. By standard arguments, e.g., [26], it follows that there is a solution of (11a) with $w = 0$ in $r = \rho$ which lies between $\underline{w}$ and $\overline{w}$. This solution is obtained using an iteration argument starting from either the sub- or the supersolution, both of which are radial.

As a consequence, the constructed solution is also radial. The regularity of w, i.e., $w \in C^2(B_\rho)$, follows directly from ODE arguments. In fact the bounded solutions w of (11a) with $w = 0$ on ∂B_ρ are in $C^2(\overline{B_\rho})$; see again [26].

If w_1 and w_2 are two such solutions, then we set

$$f(x, w) = \alpha\beta\left((w + \theta_h^R(x))^4 - \theta_f^4\right)$$

and

$$c(x) = \int_0^1 \frac{\partial f}{\partial w}(x, tw_1(x) + (1-t)w_2(x))dt.$$

The function $v = w_1 - w_2$ is a solution of

$$\begin{cases} -\Delta v + (3\alpha^2 + c(x))v = 0 & \text{in } B_\rho, \\ v = 0 & \text{on } \partial B_\rho, \end{cases}$$

where $c \in C(\overline{B}_\rho)$. By the maximum principle (see [26]), $v \equiv 0$ if c is nonnegative. Thus we have uniqueness in the class of functions w which satisfy $w(x) + \theta_h^R(x) \geq 0$, i.e., the functions w for which the corresponding temperature profile θ is positive, and it is natural to restrict to this class. This completes the proof. ∎

Remark 7. Writing (11a) as an ODE, i.e.,

$$w'' = -\frac{2}{r}w' + 3\alpha^2 w + \alpha\beta\big(\big(w + \theta_h^R(r)\big)^4 - \theta_f^4\big),$$

with initial conditions $w(0) = w_0$ and $w'(0) = 0$, this initial value problem is well-posed and behaves nicely in terms of continuous dependence on parameters. In particular, w, w', w'', and w''' are uniformly bounded on bounded intervals (for bounded ranges of α^2 and $\alpha\beta$). Alternatively, to examine regularity, one could proceed from the PDE (11a) directly using bootstrap arguments and Hölder estimates for elliptic equations; see, e.g., [26].

Remark 8. We emphasize that w is defined for $0 \leq r \leq \rho$ and that ρ as well as R are parameters with $0 < R < \rho$. Thus we write $w = w_\rho^R$.

A.2. Solutions on the whole space. In this section we take the limit $\rho \to \infty$ to prove existence of a solution w of problem (11).

LEMMA 10. *For R fixed, there exists a solution w of problem* (11) *which satisfies the bound* (12). *The solution belongs to $C^2(\mathbb{R})$ and is unique in the class of radial and nonradial functions.*

Proof. Take a sequence $\rho_n \to \infty$ as $n \to \infty$, and set $w_n = w_{\rho_n}^R$, so w_n is a solution of

$$\begin{cases} -\Delta w_n + 3\alpha^2 w_n = -\alpha\beta\big((w_n + \theta_h^R)^4 - \theta_f^4\big) & \text{in } B_{\rho_n}, \\ w_n = 0 & \text{on } \partial B_{\rho_n}, \end{cases}$$

as constructed in section A.1. We extend w_n to the whole of $\mathbb{R}^3$ by setting $w_n \equiv 0$ for $r \geq \rho_n$. Clearly estimate (12) continues to hold for w_n.

Now fix some $\rho = \bar{\rho}$ and consider the solutions w_n with $\rho_n > \bar{\rho}$, and in particular their restrictions to $B_{\bar{\rho}}$. It follows directly from Remark 7 that w_n and its first and second order derivatives are bounded and equicontinuous. Note that the nonlinear term in (11) is Lipschitz continuous if w is. Thus, we may extract a subsequence along which w_n converges in $C^2(\overline{B_{\bar{\rho}}})$. Choosing $\bar{\rho} = 1, 2, 3, \ldots$, a standard diagonal argument now produces a subsequence along which w_n converges in $C^2(\overline{B_{\bar{\rho}}})$ for every

$\bar{\rho} > 0$. It follows that the limit w exists on the whole space, and that it satisfies (11a) as well as the bound (12). Clearly w corresponds to a temperature profile $\theta = \theta_h^R + w > 0$ on $\mathbb{R}^3$.

Now suppose we have two such profiles. Reasoning as in the uniqueness proof in Lemma 9, we find that $v = w_1 - w_2$ is bounded and satisfies

$$-\Delta v + (3\alpha^2 + c(x))v = 0 \quad \text{in}\,\mathbb{R}^3.$$

When $v \to 0$ as $|x| \to \infty$ (uniformly) the maximum principle implies that $v \equiv 0$, provided that the coefficient $3\alpha^2 + c(x)$ of v is nonnegative. Thus we have uniqueness in the class of solutions w which have $w(x) \to 0$ as $|x| \to \infty$ uniformly. □

A.3. Proof of Theorem 1. In the previous section we proved Theorem 2 and showed that, omitting the reaction rate from the problem formulation, there exists for every $R > 0$ a unique solution triple (θ, Y, u) with $\theta > 0$. It remains to solve (7) with $\theta_R(R)$ given by Theorem 2.

Remark 9. In view of the estimate (12) the flame temperature $\theta(R)$ is bounded between θ_f and $\theta_f + \frac{Y_f}{\mathrm{Le}}$. As F is a continuous positive function, let us define the positive numbers

$$m = \min_{\theta \in [\theta_f,\, \theta_f + Y_f/\mathrm{Le}]} F(\theta) \quad \text{and} \quad M = \max_{\theta \in [\theta_f,\, \theta_f + Y_f/\mathrm{Le}]} F(\theta).$$

Then any solution of the full flame ball problem must satisfy

$$\frac{Y_f}{\mathrm{Le}}\,\frac{1}{M} \le R \le \frac{Y_f}{\mathrm{Le}}\,\frac{1}{m}. \tag{19}$$

Equation (7) has a left-hand side which goes from $+\infty$ to 0 as R goes from 0 to ∞. Its right-hand side is bounded between m and M. Thus the existence of the solution in Theorem 1 is immediate once we know that Remark 1 (about continuity of $\theta_R(R)$) is true. More precisely, we have the following claim.

LEMMA 11. *If $R_n \to R > 0$, then the corresponding functions θ_{R_n} converge uniformly to θ_R on $[0, \infty)$.*

Proof. Clearly this will follow from the same statement for w_R, where w_R is the solution of (11) obtained in Lemma 3. In view of the bound (12), uniform convergence on bounded subsets implies uniform convergence on $[0, \infty)$. By exactly the same arguments as in the proof of Lemma 3 in section A.2, it follows that along a subsequence of $n \to \infty$, w_{R_n} (as well as its first and second order derivatives) converges uniformly on any bounded interval to a solution of (11) satisfying (12). Since this solution is unique, it follows that $w_{R_n} \to w_R$ along this subsequence. In fact, every sequence of $n \to \infty$ has a subsequence for which this is the case. But then there cannot be a sequence of n for which $||w_{R_n} - w_R||_\infty$ is bounded away from zero. This completes the proof of Lemma 11 and thereby of Theorem 1. □

Remark 10. Instead of using this sequence argument, one could also invoke an implicit function argument to conclude that $R \to \theta_R(R)$ (or $R \to w_R(R)$) is smooth. Furthermore, assuming the derivatives of the left- and right-hand sides of (7) to be different at solutions, it follows immediately that the number of solutions is odd. This is the statement that in general situations the number of solutions is odd.

A.4. Uniform estimates. As we have seen, solutions of the flame ball problem are given by $\theta = \theta_h^R + w_R$, where R is such that (7) holds, and where w_R is a C^2-function (of course, θ_h^R is not). Moreover, w_R satisfies (11). In this section we show that w is monotone in r.

LEMMA 12. *The solution $w = w_R$ of (11) has $w' \geq 0$.*

Proof. w solves

$$-w'' - \frac{2}{r}w' = g(r,w) = -3\alpha^2 w - \alpha\beta\left((w + \theta_h^R(r))^4 - \theta_f^4\right), \tag{20}$$

where g satisfies

$$\frac{\partial g}{\partial w} < 0 \quad \text{and} \quad \frac{\partial g}{\partial r} \geq 0,$$

the latter being discontinuous in $r = R$, of course, but with limits existing from both sides. Moreover, $w'(0) = 0$ by symmetry.

If w' is negative somewhere, then there must be points r_1 and r_2 such that $w'(r_1) = w'(r_2) = 0$, while $w' < 0$ on (r_1, r_2). This follows from $w'(0) = 0$ and $0 > w(r) \to 0$ as $r \to \infty$.

Clearly then $g(r_1, w_1(r_1)) = -w''(r_1) \geq 0$ and $g(r_2, w_1(r_2)) = -w''(r_2) \leq 0$, contradicting

$$\frac{d}{dr}g(r,w(r)) = \frac{\partial g}{\partial r} + \frac{\partial g}{\partial w}\frac{\partial w}{\partial r} > 0 \quad \text{on } (r_1, r_2). \qquad \square$$

LEMMA 13. *There exists a constant C depending on α^2 and $\alpha\beta$ such that*

$$\int_0^\infty w'(r)^2 dr < C.$$

Proof. Multiplying (20) by w and integrating from r_1 to r_2 $(0 < r_1 < r_2 < \infty)$, we obtain

$$-\int_{r_1}^{r_2} w''w dr = \int_{r_1}^{r_2} \frac{2}{r}w'w dr + \int_{r_1}^{r_2} g(r,w)w dr,$$

so that

$$\int_{r_1}^{r_2} |w'|^2 dr + w'(r_1)w(r_1) = w'(r_2)w(r_2) + \int_{r_1}^{r_2} \frac{2}{r}w'w dr + \int_{r_1}^{r_2} g(r,w)\, w dr.$$

Letting $r_1 \to 0$ and using $w' \geq 0$, $w < 0$, it follows that, also using (12),

$$\int_0^{r_2} |w'|^2 dr \leq \int_0^{r_2} g(r, w(r))\, w(r) dr \leq C,$$

where C is a constant depending linearly on α^2 and $\alpha\beta$, but not on r_2. This proves the claim. $\square$

Going one step further, we get the following.

LEMMA 14. *w belongs to $H^2(0,\infty)$.*

Proof. Multiplying (20) by $-w''$ and integrating from r_1 to r_2, we find

$$\int_{r_1}^{r_2} |w''|^2 dr + \int_{r_1}^{r_2} \frac{2}{r}w'w'' dr + \int_{r_1}^{r_2} g(r,w)\, w'' dr = 0.$$

Hence, with, e.g., $r_1 = 2$,

$$\begin{aligned}\int_2^{r_2} |w''|^2 dr \leq &\left(\int_2^{r_2} |w'|^2 dr\right)^{\frac{1}{2}} \left(\int_2^{r_2} |w''|^2 dr\right)^{\frac{1}{2}} \\ &+ \left(\int_2^{r_2} |g(r,w(r))|^2 dr\right)^{\frac{1}{2}} \left(\int_2^{r_2} |w''|^2 dr\right)^{\frac{1}{2}}.\end{aligned}$$

In view of (12), (20), and Lemma 13, we conclude that

$$\int_2^\infty |w''|^2 dr \le C,$$

where C depends on α^2 and $\alpha\beta$. The fact that w is C^2 implies that also $\int_0^\infty |w''|^2$ is bounded. Lemma 13 and inequality (12) finish the proof. □

Remark 11. If we consider the problem in $\mathbb{R}^3$, one can easily check that w belongs to $W^{2,p}(\mathbb{R}^3)$ if $p > 3$.

REFERENCES

[1] Y. B. Zeldovich, G. I. Barenblatt, V. B. Librovich, and G. M. Makhviladze, *The Mathematical Theory of Combustion and Explosions* (translated from the Russian by D. H. McNeill), Consultants Bureau (Plenum), New York, 1985.

[2] P. D. Ronney, *Near-limit flame structures at low-Lewis number*, Combust. Flame, 82 (1990), pp. 1–14.

[3] P. D. Ronney, M. S. Wu, H. G. Pearlman, and K. J. Weiland, *Experimental study of flame balls in space: Preliminary results from sts-*83, AIAA J., 36 (1998), pp. 1361–1368.

[4] *Structure of Flame Balls at Low Lewis-number (SOFBALL) Home Page,* online at http://exploration.grc.nasa.gov/combustion/sofball/sofball_index.htm.

[5] G. Joulin and B. Deshaies, *On radiation-affected flame propagation in gaseous mixtures seeded with inert particules*, Combust. Sci. Tech., 47 (1986), pp. 299–315.

[6] G. Joulin and M. Eudier, *Radiation-dominated propagation and extinction of slow, particle-laden gaseous flames*, in Proceedings of the 22nd International Symposium on Combustion, The Combustion Institute, Pittsburgh, PA, 1988, pp. 1579–1585.

[7] J. D. Buckmaster and T. L. Jackson, *The effects of radiation on the thermal-diffusive stability boundaries of premixed flames*, Combust. Sci. Tech., 103 (1994), pp. 299–313.

[8] J. D. Buckmaster, G. Joulin, and P. D. Ronney, *The structure and stability of non adiabatic flame balls*, Combust. Flame, 79 (1990), pp. 381–392.

[9] C. Lederman, J.-M. Roquejoffre, and N. Wolansky, *Mathematical justification of a non-linear integrodifferential equation for the propagation of spherical flames*, Ann. Mat. Pura Appl., 183 (2004), pp. 173–239.

[10] A. A. Shah, R. W. Thachter, and J. W. Dold, *Stability of a spherical flame ball in a porous medium*, Combust. Theory Modelling, 4 (2000), pp. 511–534.

[11] G. I. Sivashinsky, *On flame propagation under condition of stoichiometry*, SIAM J. Appl. Math., 39 (1980), pp. 67–82.

[12] C. M. Brauner and A. Lunardi, *Instabilities in a two-dimensional combustion model with free boundary*, Arch. Ration. Mech. Anal., 154 (2000), pp. 157–182.

[13] J. D. Buckmaster, G. Joulin, and P. D. Ronney, *The structure and stability of non adiabatic flame balls:* II. *Effects of far field losses*, Combust. Flame, 84 (1991), pp. 411–422.

[14] D. Mihalas and B. Mihalas, *Foundation of Radiation Hydrodynamics*, Oxford University Press, Oxford, UK, 1984.

[15] G.-C. Pomraning, *The Equation of Radiation Hydrodynamics*, Pergamon Press, Elmsford, NY, 1973.

[16] B. Dubroca and J. L. Feugeas, *Etude théorique et numérique d'une hiérarchie de modèles aux moments pour le transfert radiatif*, C. R. Acad. Sci. Paris Sér. I Math., 329 (1999), pp. 915–920.

[17] M.-N. Ozisik, *Radiative Transfer*, Wiley, New York, 1973.

[18] M.-F. Modest, *Radiative Heat Transfer*, Series in Mechanical Engineering, McGraw–Hill, New York, 1993.

[19] R. Siegel and J. R. Howell, *Thermal Radiation Heat Transfer*, McGraw–Hill, New York, 1972.

[20] C. M. Brauner, J. Hulshof, and J.-F. Ripoll, *Existence of travelling wave solutions in a combustion-radiation model*, Discrete Contin. Dynam. Systems, 1 (2001), pp. 193–208.

[21] O. Baconneau, J. B. van den Berg, C. M. Brauner, and J. Hulshof, *Multiplicity and stability of travelling wave solutions in a free boundary combustion-radiation problem*, European J. Appl. Math., 15 (2004), pp. 79–102.

[22] E. J. Doedel, A. R. Champneys, T. F. Fairgrieve, Y. A. Kuznetsov, B. Sandstede, and X. Wang, *Auto*97, *Continuation and bifurcation software for ordinary differential*

equations (*with homcont*), 1997; available by anonymous ftp from ftp.cs.concordia.ca/pub/doedel/auto.

[23] F. A. Williams, *Combustion Theory*, Addison–Wesley, Reading, MA, 1994.

[24] J. D. Buckmaster and G.S.S. Ludford, *Theory of Laminar Flames*, Cambridge University Press, Cambridge, UK, 1982.

[25] V. Guyonne and L. Lorenzi, *Instability in a flame ball problem*, Discrete Contin. Dyn. Syst. Ser. B, 2006 to appear.

[26] L. C. Evans, *Partial Differential Equations,* Grad. Stud. Math. 19, American Mathematical Society, Providence, RI, 1998.

SIAM J. APPL. MATH.
Vol. 67, No. 1, pp. 138–163

SPATIAL SPREAD OF RABIES REVISITED: INFLUENCE OF AGE-DEPENDENT DIFFUSION ON NONLINEAR DYNAMICS*

CHUNHUA OU† AND JIANHONG WU‡

Abstract. We consider the spatio-temporal patterns of disease spread involving structured populations. We start with a general model framework in population biology and spatial ecology where the individual's spatial movement behaviors depend on its maturation status, and we show how delayed reaction diffusion equations with nonlocal interactions arise naturally. We then consider the impact of this delayed nonlocal interaction on the disease spread by revisiting the spatial spread of rabies in continental Europe during the period between 1945 and 1985. We show how the distinction of territorial patterns between juvenile and adult foxes, the main carriers of the rabies under consideration, yields a class of partial differential equations involving delayed and nonlocal terms that are implicitly defined by a hyperbolic-parabolic equation, and we show how incorporating this distinction into the model leads to a formula describing the relation of the minimal wave speed and the maturation time of foxes. We show how the homotopy argument developed by Chow, Lin, and Mallet-Paret can be applied to obtain the existence of a heteroclinic orbit between a disease-free equilibrium and an endemic state for the spatially averaged system of delay differential equations, and we illustrate how the technique developed by Faria, Huang, and Wu can be used to establish the existence of a family of traveling wavefronts in the neighborhood of the heteroclinic orbit for the corresponding spatial model.

Key words. time delay, nonlocal, reaction, reaction diffusion, traveling waves, fronts, stability, structured model, disease modeling, minimal wave speed

AMS subject classifications. 34C25, 34K15, 34K18, 35K55

DOI. 10.1137/060651318

1. Introduction.

Spatial movement and reaction time lag are certainly two intrinsic features in biological systems; their interaction seems to be one of the many factors for possible complicated spatio-temporal patterns in a single species population without an external time-dependent forcing term. Modeling this interaction is nevertheless a highly nontrivial task, and recent progress indicates diffusive (partial or lattice) systems with nonlocal and delayed reaction nonlinearities arise very naturally. Such systems were investigated in the earlier work of Yamada [34], Pozio [24, 25], Redlinger [26, 27], and the modeling and analysis effort in the ground-breaking work by Britton [3], Gourley and Britton [9], Smith and Thieme [28] marked the beginning of the systematic study of a new class of nonlinear dynamical systems directly motivated by consideration of biological realities [10, 11].

This new class of nonlinear dynamical systems can be derived from the classical structured population model involving maturation-dependent spatial diffusion rates and nonlinear birth and natural maturation processes. More specifically, if we use $u(t, x)$ to denote the total number of matured individuals in a single species population

*Received by the editors January 31, 2006; accepted for publication (in revised form) August 10, 2006; published electronically November 14, 2006. The work was partially supported by the Canada Research Chairs Program, by NCE Center of Mathematics for Information Technology and Complex Systems, and by the Natural Sciences and Engineering Research Council of Canada.
http://www.siam.org/journals/siap/67-1/65131.html
†Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, Newfoundland, A1C 5S7, Canada (ou@math.mun.ca).
‡Laboratory for Industrial and Applied Mathematics, Department of Mathematics and Statistics, York University, Toronto, Ontario, M3J 1P3, Canada (wujh@mathstat.yorku.ca).

and if we assume the maturation time is a fixed constant τ, then we have

$$\frac{\partial}{\partial t}u(t,x) = D\frac{\partial^2}{\partial x^2}u(t,x) - du(t,x) + j(t,\tau,x), \tag{1.1}$$

where D and d are the diffusion and death rates of the adult population (that are assumed to be age-independent), and $j(t,\tau,x)$ is the maturation rate that is given by the rate where an individual was born exactly time $t-\tau$ ago in all possible spatial locations but moved to the current position x upon maturation. This maturation rate is thus regulated by the birth process and the dynamics of the individual during the maturation phase. In the work of So, Wu, and Zou [29], this is derived from the structured population model

$$\Big(\frac{\partial}{\partial t} + \frac{\partial}{\partial a}\Big)j(t,a,x) = D_I\frac{\partial^2}{\partial x^2}j(t,a,x) - d_I j(t,a,x) \tag{1.2}$$

for the density $j(t,a,x)$ of the immature individual with $a \in (0,\tau]$ as the variable for age, subject to some (spatial) boundary conditions (if the space is bounded) and the following (age) boundary condition:

$$j(t,0,x) = b(u(t,x)), \tag{1.3}$$

where b is the birth rate function that is assumed to be dependent on the matured population, and D_I and d_I are the diffusion and death rates of the immature individual (these rates are allowed to depend on the age a in [29]). The maturation rate $j(t,\tau,x)$ can be obtained by solving the linear hyperbolic-parabolic equation (1.2) subject to the boundary condition (1.3). In the case of unbounded one-dimensional space, we have

$$j(t,\tau,x) = e^{-d_I\tau}\int_R b(u(t-\tau,y))f(x-y)dy. \tag{1.4}$$

In other words, the maturation rate at time t and spatial location x is the sum of the birth rate at time $t-\tau$ at the spatial location y, times the probability $f(x-y)$ of the individual moved from y to the current position x, and then times the survival rate $e^{-d_I\tau}$ during the entire maturation phase. Incorporating (1.4) into (1.1), we then obtain a closed system of reaction diffusion equations with nonlocal delayed nonlinearity as follows:

$$\frac{\partial}{\partial t}u(t,x) = D\frac{\partial^2}{\partial x^2}u(t,x) - du(t,x) + e^{-d_I\tau}\int_R b(u(t-\tau,y))f(x-y)dy, \tag{1.5}$$

where

$$f(x) = \frac{1}{\sqrt{4\pi D_I\tau}}e^{\frac{-x^2}{4D_I\tau}}.$$

For the existence of positive solutions to (1.5) with various initial and boundary conditions, we refer to [19, 33]. Recently, there has been some rapid development towards a qualitative theory for the asymptotic behaviors of solutions to the above equation with various types of assumptions on the birth functions. Notably, in comparison with the ordinary reaction diffusion analogue, we will have more prototypes than the so-called monostable and bistable cases. See [11].

The analytic form above for f was derived in [29]. It is possible to obtain such an analytic form here since the dynamical process during the maturation phase is governed by a linear hyperbolic-parabolic equation with time-independent constant

coefficients. Such a possibility disappears in an ecological system consisting of multiple species with age- or stage-dependent diffusion rates when these species interact during their maturation phases. This is also the case for the spread of a disease even if its main carrier involves only a single species, since the model describing the infection process must involve the transfers of individuals from one compartment to another, and some of these transfers such as the force of infection from the susceptible compartment to the infective compartment are nonlinear.

We will illustrate the above difficulty and usefulness of modeling the spread of diseases involving stage-dependent spatial diffusion by considering the spatial spread of rabies in continental Europe during the period 1945–1985. Our focus is on the front of the epizootic wave of rabies, starting on the edge of the German/Polish border and moved westward at an average speed of about 30–60 km a year. This traveling wavefront has been investigated quite successfully (see [15, 22]), where the minimal wave speed was calculated from basic epidemiological and ecological parameters, and compared well with field observation data. It was also noted that juvenile foxes leave their home territory in the autumn traveling distances that typically may be 10 times a territory size in search of a new territory. If a fox happened to have contracted rabies around the time of such long-distance movement, it could certainly increase the spreading of the disease into uninfected areas. This observation has not been considered in the existing models. It turns out that incorporating the differential spatial movement behaviors of adult and juvenile foxes into a deterministic model yields a much more complicated system of reaction diffusion equations with delayed nonlinear nonlocal interactions.

More precisely, the celebrated work [15, 22] used a system of a reaction diffusion equation for the infective, coupled with an ODE for the susceptible foxes—the main carrier of the disease—under the assumption that the infective compartment consists of both rabid foxes and those in the incubation stage, and that susceptible foxes are territorial and thus their spatial movement can be ignored. It was already pointed out, in both papers mentioned above and their later extensions and further detailed studies, that the spatial movement behaviors of susceptible juvenile foxes are different since they prefer to leave their home territories in search of new territories of their own. How to describe this stage-dependent diffusion pattern of susceptible foxes and how stage-dependent diffusion affects the spatial spread of rabies are the main focus of the current paper.

It turns out, as will be shown in section 2, that such a stage-dependent diffusion of susceptible foxes and the random movement of rabid foxes due to the loss of the sense of direction and territorial behaviors yield a coupled system of reaction diffusion equations with nonlocal delayed nonlinearity for the juvenile susceptible foxes $M(t,x)$ and total rabies foxes $J(t,x)$. Unlike system (1.5) for a single species population with simple dynamics during the maturation phase, the coupled system for (M,J) involves the density of the juvenile foxes $S(t,a,y)$ for all $y \in R$ and the maturation rate $S(t,\tau,x)$ (again, τ is assumed to be a constant maturation time of the foxes) and the force of infection that is proportional to the product of $J(t,x)\int_0^\tau S(t,a,x)da$. This density of the juvenile foxes cannot be solved explicitly in terms of $M(s,\cdot)$ with $s \le t$ although it is given implicitly by solving a hyperbolic-parabolic equation with a nonlinear term.

Some of the key issues related to the spatial spread can nevertheless be addressed, despite the aforementioned difficulty in obtaining an explicit analytic formula of $S(t,a,x)$ in terms of the historical values of M at all spatial locations. As shall be shown in section 3, the linear stability of two spatially homogeneous equilibria

can be fully investigated and the minimal wave speed can be calculated. One of the results we obtain from this calculation is that the minimal wave speed is a function of the average maturation time. More precisely, knowing the carrying capacities for the adult and juvenile foxes, the minimal wave speed is a decreasing function of the maturation period. This results coincide in principle with the speculation in [15, 22], and give a more precise qualitative description of the influence of maturation time on the propagation of the disease in space.

Establishing the existence of traveling waves turns out to be a very difficult task due to the loss of monotonicity of the nonlocal delayed nonlinearity. In section 4, we utilize a general result of Faria, Huang, and Wu [8] that claims the existence of traveling waves in the neighborhood of a heteroclinic orbit between the two equilibria of a corresponding ordinary delay differential system obtained from the delayed reaction diffusion system for (J, M) through a spatial average, and we obtain the existence of this heteroclinic orbit by an approach based on a combination of perturbation analysis [23], the Fredholm theory, and some fixed point theorems [5, 13]. This will be developed in detail in section 4, along with some numerical simulations to show how the maturation time affects the calculation of the minimal wave speed, and how the diffusion of the juvenile foxes impacts the amplitudes and frequencies of the oscillatory long tails of the traveling wavefronts.

2. Derivation of the model. Here we use a deterministic approach to describe the spatial spread of rabies. Following [15, 22], we divide the fox population into two groups: the infective and the susceptible. The former consists of both rabid foxes and those in the incubation stage. The basic facts and assumptions of our model are as follows:

(H1) The rabies virus is contained in the saliva of the rabid fox and is normally transmitted by bite. Therefore, contact between a rabid and a susceptible fox is necessary for the transmission of the disease.

(H2) Rabies is invariably fatal in foxes.

(H3) Adult susceptible foxes are territorial and seem to divide the countryside into nonoverlapping home ranges which are marked out by scent. They do occasionally travel considerable distances but always return to their home territory. However, for young susceptible juvenile foxes, their behaviors are different, because they prefer to leave their home territories in search of new territories of their own.

(H4) The rabies virus enters the central nervous system and induces behavioral changes of foxes. If the spinal cord is involved, it often takes the form of paralysis. However, if the virus enters the limbic system, the foxes become aggressive, lose their sense of direction and territorial behavior, and wander about in a more or less random way.

Modeling the distinction of diffusion patterns of young and adult susceptible foxes, already observed in [15, 22], is the main focus of this paper. Because of this distinction, we shall incorporate age structure into our model and consider the fox population with two age classes: the immature and the mature. Let $I(t, a, x)$ and $S(t, a, x)$ denote the population density at time t, age $a \geq 0$, and spatial location $x \in R = (-\infty, \infty)$ for the infective and the susceptible foxes, respectively, and let τ be the maturation time which is assumed to be a constant. Then the integral

$$J(t, x) = \int_0^\infty I(t, a, x)da \tag{2.1}$$

is the total population of the infective foxes and

$$M(t,x)=\int_{\tau}^{\infty} S(t,a,x)da \tag{2.2}$$

is the total population of the adult susceptible foxes. Using Fick's diffusive law and the mass active incidence, we have

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial a}\right) I(t,a,x)=D_I\frac{\partial^2}{\partial x^2}I(t,a,x)+\beta S(t,a,x)J(t,x)-d_I I(t,a,x), \tag{2.3}$$

where D_I is the diffusive coefficient, d_I is the death rate for the infective foxes, and β is the transmission rate. Using $I(t,\infty,x)=0$ and $I(t,0,x)=0$, we obtain from (2.1) and (2.3) that

$$\begin{aligned}\frac{\partial J(t,x)}{\partial t}&=\int_0^{\infty}\frac{\partial I(t,a,x)}{\partial t}da\\ &=D_I\frac{\partial^2 J(t,x)}{\partial x^2}+\beta M(t,x)J(t,x)-d_I J(t,x)\\ &\quad+\beta J(t,x)\int_0^{\tau} S(t,a,x)da.\end{aligned} \tag{2.4}$$

For $S(t,a,x)$ with $a\geq\tau$, we have the structured population model (see [20] or [32])

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial a}\right) S(t,a,x)=-\beta S(t,a,x)J-d_S S(t,a,x), \tag{2.5}$$

where the constant d_S is the death rate for the susceptible foxes. Using $S(t,\infty,x)=0$, we get from (2.2) and (2.5) that

$$\frac{\partial M(t,x)}{\partial t}=-\beta M(t,x)J(t,x)-d_S M(t,x)+S(t,\tau,x). \tag{2.6}$$

To obtain a closed system for (J,M), we need to formulate $S(t,a,x)$ with $0\leq a\leq\tau$ in terms of (J,M). This is achieved by using the following structured hyperbolic-parabolic equation with the initial condition given by the birth process:

$$\begin{cases}\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial a}\right) S(t,a,x)=D_Y\frac{\partial^2}{\partial x^2}S(t,a,x)-\beta S(t,a,x)J(t,x)-d_Y S(t,a,x),\\ S(t,0,x)=b(M(t,x)),\end{cases} \tag{2.7}$$

where D_Y is the diffusive coefficient for the immature susceptible foxes and $b(\cdot)$ is the birth function of the susceptible foxes.

Combining (2.4) and (2.6) together gives

$$\begin{cases}\frac{\partial J(t,x)}{\partial t}=D_I\frac{\partial^2 J(t,x)}{\partial x^2}+\beta M(t,x)J(t,x)-d_I J(t,x)+\beta J(t,x)\int_0^{\tau}S(t,a,x)da,\\ \frac{\partial M(t,x)}{\partial t}=-\beta M(t,x)J(t,x)-d_S M(t,x)+S(t,\tau,x),\end{cases} \tag{2.8}$$

where $S(t,a,x)$, $0\leq a\leq\tau$, is determined by solving the hyperbolic-parabolic system (2.7).

REMARK 2.1. *$S(t,a,x)$ in (2.7) depends on t, a, $M(t,y)$, and $J(s,y)$ for all $0 \le s \le t$ and $y \in R$, but an explicit formula for $S(t,a,x)$ cannot be found. We shall write $F(t,a,x) = F(a,M,J)(t,a,x)$ to indicate this functional relation. It is easy to show that*

$$F(a,M,J_1)(t,a,x) \ge F(a,M,J_2)(t,a,x) \quad \text{if } 0 \le J_1(s,y) \le J_2(s,y) \tag{2.9}$$

for $0 \le s \le t$ and $y \in R$, and for $t \ge a$,

$$F(a,M,J_1)(t,a,x) = b(M(t-a,x))e^{-\int_{t-a}^{t}(d_Y+\beta J(u,x))du} \quad \text{when } D_Y = 0. \tag{2.10}$$

REMARK 2.2. *When $\tau = 0$, system (2.8) reduces to*

$$\begin{cases} \frac{\partial J(t,x)}{\partial t} = D_I \frac{\partial^2 J(t,x)}{\partial x^2} + \beta J(t,x)S(t,x) - d_I J(t,x), \\ \frac{\partial M(t,x)}{\partial t} = -\beta J(t,x)M(t,x) - d_S M(t,x) + b(M(t,x)). \end{cases} \tag{2.11}$$

This model was studied in [7] *and* [15] *by assuming that the birth function obeys the well-known logistic growth, that is, the (gross) birth function $b(M) := d_S M + b_0 M(1 - M/S_0)$, where S_0 is the carrying capacity of the susceptible fox population and b_0 is the net birth rate for the susceptible foxes when the population density is close to zero. After rescaling by*

$$u(t,x) = J(t,x)/S_0,\ v(t,x) = M(t,x)/S_0,\ x^* = (\beta S_0/D_I)^{1/2}x,\ t^* = \beta S_0 t,\ r = \frac{d_I}{\beta S_0}$$

and dropping the asterisk, we can transform (2.11) into

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u(v - r), \\ \frac{\partial v}{\partial t} = -uv + kv(1 - v), \end{cases} \tag{2.12}$$

where $k = b_0/\beta S_0$. In [7], *it was proved that if $r = \frac{d_I}{\beta S_0} < 1$, then the infective and the susceptible foxes coexist, and there exists a family of traveling wavefronts $(J = S_0 u(x+ct), M = S_0 v(x+ct))$ for (2.11) which connect $(0, S_0)$ to $(kS_0(1-r), S_0 r)$ with the wave speed c satisfying*

$$c \ge c_{\min} = 2\sqrt{\beta S_0 D_I}\sqrt{1 - \frac{d_I}{\beta S_0}} = 2\sqrt{D_I(\beta S_0 - d_I)}.$$

In addition, it was shown that there is a constant $k^* > 0$ so that (a) *if $k = b_0/\beta S_0 > k^*$, then the wavefront (u,v) approaches $(k(1-r)S_0, rS_0)$ monotonically;* (b) *if $k = b_0/\beta S_0 \le k^*$, then the wavefront (u,v) approaches $(k(1-r)S_0, rS_0)$ with oscillatory damping.*

In [14] *and* [15], *instead of the logistic growth, a static population of the susceptible is assumed in the sense that deaths are equally balanced by births. This yields the simple model*

$$\begin{cases} \frac{\partial J}{\partial t} = D_I \frac{\partial^2 J}{\partial x^2} + \beta JM - d_I J, \\ \frac{\partial M}{\partial t} = -\beta JM, \end{cases} \tag{2.13}$$

where M and J are the total numbers of susceptible and infective foxes, respectively. It was shown that with initial (maximum) susceptible population S_0, if $r = \frac{d_I}{\beta S_0} > 1$

(that is, the mortality rate of the infective foxes is greater than the rate of recruitment of new infective), the infection dies out quickly. If $r = \frac{d_I}{\beta S_0} < 1$*, there is a family of traveling wavefronts to the system* (2.13) *with the minimal speed*

$$c_{\min} = 2[D_I(\beta S_0 - d_I)]^{1/2}.$$

See [1, 6] for related work.

REMARK 2.3. *When* $D_Y = 0$*, system* (2.7)–(2.8) *reduces to the following model:*

$$\begin{cases} \frac{dJ}{dt} = D_I \frac{\partial^2 J}{\partial x^2} + \beta MJ - d_I J + \beta J \int_0^\tau b(M(t-a)) e^{\int_0^a -(d_Y + \beta J(t-s))ds} da, \\ \frac{dM}{dt} = -\beta MJ - d_S M + b(M(t-\tau)) e^{\int_{t-\tau}^t -(d_Y + \beta J(s))ds}. \end{cases} \tag{2.14}$$

This is a delayed reaction diffusion system with distributed delay but without spatial averaging. We will numerically compare the behavior of solutions to (2.14) *with that of* (2.7)–(2.8) *in section* 4.3.

In the remaining part of this paper, we consider the dynamics of (2.7)–(2.8) using the birth function

$$b(M) = b_0 M e^{-\bar{a}M},$$

where $\bar{a} > 0$ is a positive parameter, and $b_0 = b'(0)$ is the birth rate when the population size is small. This birth function exhibits the logistic growth nature of the fox population in the absence of the disease. Such a function has been used in the well-studied Nicholson blowfly model [12] and is common in models of fish. The specific form of such a function is not so important for the method developed below, though the specific form facilitates and simplifies our qualitative analysis since, as will be shown, constant equilibria can be explicitly described.

3. Structure of equilibria. In this section, we describe the structure of equilibria of biological interest. At an equilibrium, (J, M) takes on a constant value, namely,

$$J \equiv J_0, \quad M \equiv M_0,$$

for constants J_0 and M_0. Then from (2.7) we have

$$\begin{cases} (\frac{\partial}{\partial t} + \frac{\partial}{\partial a})S = D_Y \frac{\partial^2}{\partial x^2} S - d_Y S - \beta S J_0, \\ S(t, 0, x) = b(M_0). \end{cases} \tag{3.1}$$

To solve (3.1), we define $V^s(t, x) = S(t, t-s, x)$ and obtain, for $t \geq s$, that

$$\begin{aligned} \frac{\partial}{\partial t} V^s(t, x) &= \frac{\partial S}{\partial t}(t, a, x)|_{a=t-s} + \frac{\partial S(t, a, x)}{\partial a}|_{a=t-s} \\ &= D_Y \frac{\partial^2}{\partial x^2} V^s(t, x) - d_Y V^s(t, x) - \beta J_0 V^s(t, x). \end{aligned} \tag{3.2}$$

Note that (3.2) is a linear reaction diffusion equation with constant coefficients. The associated initial condition is

$$V^s(s, x) = b(M_0), \quad x \in R. \tag{3.3}$$

To ensure uniqueness of solutions, we also impose biologically realistic boundary conditions as follows:

$$|V^s(t,\pm\infty)| < \infty. \tag{3.4}$$

The solution of (3.2)–(3.4) is given by

$$V^s(t,x) = b(M_0)e^{-(d_Y+\beta J_0)(t-s)}. \tag{3.5}$$

That is,

$$S(t,t-s,x) = b(M_0)e^{-(d_Y+\beta J_0)(t-s)} \tag{3.6}$$

and

$$S(t,a,x) =: F(a,M_0,J_0) =: b(M_0)e^{-(d_Y+\beta J_0)a},\ 0 \le a \le \tau,$$

from which, with (2.8), it follows that equilibrium (M_0, J_0) is given by the following algebraic equations:

$$\begin{cases} \beta M_0 J_0 - d_I J_0 + \beta J_0 \frac{b(M_0)}{(d_Y+\beta J_0)}(1 - e^{-(d_Y+\beta J_0)\tau}) = 0, \\ -\beta M_0 J_0 - d_S M_0 + b(M_0)e^{-(d_Y+\beta J_0)\tau} = 0. \end{cases} \tag{3.7}$$

We now solve (3.7) for equilibria.

When $J_0 = 0$, the second equation in (3.7) gives

$$-d_S M_0 + b(M_0)e^{-d_Y\tau} = 0. \tag{3.8}$$

Thus M_0 can take on two different values: $M_0 = 0$ or $M_0 = M^\tau_{\max} = \frac{1}{\bar{a}}\ln(b_0/d_S e^{d_Y\tau}))$. Biological consideration requires that

$$\frac{b_0}{d_S e^{d_Y\tau}} > 1$$

or, equivalently,

$$\tau < \tau_{\max} = \frac{1}{d_Y}\ln\frac{b_0}{d_S}, \tag{3.9}$$

so that $M^\tau_{\max} > 0$.

When $J_0 \neq 0$, obviously from the second equation of (3.7) we can simplify the relation between J_0 and M_0 to yield

$$M_0 = \frac{1}{\bar{a}}\left(\ln\frac{b_0}{\beta J_0 + d_S} - (d_Y + \beta J_0)\tau\right).$$

Viewing M_0 as a function of J_0, that is, $M_0 = h_0(J_0)$ with h_0 being given by

$$h_0(J_0) = \frac{1}{\bar{a}}\left(\ln\frac{b_0}{\beta J_0 + d_S} - (d_Y + \beta J_0)\tau\right), \tag{3.10}$$

we find that $h_0(J_0)$ is decreasing for $J_0 \ge 0$ with

$$h_0(0) = M^\tau_{\max} > 0 \quad \text{and} \quad h_0(+\infty) < 0.$$

From the first equation of (3.7) we have

$$\frac{\beta b(M_0)}{d_I - \beta M_0} = \frac{d_Y + \beta J_0}{1 - e^{-(d_Y+\beta J_0)\tau}}. \tag{3.11}$$

The monotonic increasing property of the function on the right-hand side of (3.11) is obvious for $J_0 \in [0, \infty)$. We now check the monotonicity of the function on the left-hand side. Using the definition of $b(\cdot)$ and defining $f(x) := \beta b(x)/(d_I - \beta x)$, we have

$$\begin{aligned} f'(x) &= \beta \frac{b'(x)(d_I - \beta x) + \beta b(x)}{(d_I - \beta x)^2} \\ &= \beta \frac{b_0 e^{-\bar{a}x}(\bar{a}\beta x^2 - \bar{a} d_I x + d_I)}{(d_I - \beta x)^2}. \end{aligned}$$

It is easy to know that the function $f(x) = \beta b(x)/(d_I - \beta x)$ is increasing with respect to x provided that

$$\bar{a} d_I < 4\beta.$$

Therefore, under the above condition, a careful examination of the left-hand side of (3.11) shows that (3.11) gives a unique function $M_0 = h_1(J_0)$ ($M_0 < d_I/\beta$, $J_0 \geq 0$) which is increasing for $J_0 \in (0, \infty)$ and satisfies

$$h_1(\infty) = \frac{d_I}{\beta}.$$

It is easy to see that the intersection point of the two curves $M_0 = h_0(J_0)$ and $\mathrm{M}_0 = h_1(J_0)$ corresponds to the third equilibrium (J_*^τ, M_*^τ) of our system. As to the existence and positivity of this particular point, we have the following.

THEOREM 3.1. *Assume $\tau < \tau_{\max}$ and $ad_I < 4\beta$. Then system* (2.7)–(2.8) *has a unique positive equilibrium (J_*^τ, M_*^τ) if and only if*

$$C_0(\tau) := \frac{d_I}{\beta M_{\max}^\tau} - \frac{b(M_{\max}^\tau)(1 - e^{-d_Y\tau})}{M_{\max}^\tau d_Y} < 1, \tag{3.12}$$

where

$$M_{\max}^\tau = \frac{1}{\bar{a}} \ln \frac{b_0}{d_S e^{d_Y\tau}}.$$

Proof. The condition $\tau < \tau_{\max}$ implies that $M_{\max}^\tau$ is positive and the condition $ad_I < 4\beta$ guarantees that h_1 is increasing. Note that $h_0(0) = M_{\max}^\tau$. By the monotonicity properties of functions h_0 and h_1 and the fact that $h_0(\infty) < 0$ and $h_1(\infty) = \frac{\beta}{d_I} > 0$, it follows that the functions h_0 and h_1 have a positive intersection point if and only if $h_1(0) < h_0(0) = M_{\max}^\tau$. Now we show that $h_1(0) < h_0(0) = M_{\max}^\tau$ if and only if $C_0(\tau) < 1$. We consider two cases:

(i) $M_{\max}^\tau \geq \beta/d_I$;
(ii) $M_{\max}^\tau < \beta/d_I$.

In the first case, the proof is obvious and will be omitted here. For the second case, the inequality $C_0(\tau) < 1$ is actually equivalent to

$$\frac{d_I}{\beta M_{\max}^\tau} - \frac{b(M_{\max}^\tau)(1 - e^{-d_Y\tau})}{M_{\max}^\tau d_Y} < 1,$$

or to

$$\frac{d_Y}{(1-e^{-d_Y\tau})} < \frac{\beta b(M^\tau_{\max})}{d_I - \beta M^\tau_{\max}}. \tag{3.13}$$

From (3.11) we note that $h_1(0)$ $(< \frac{d_I}{\beta})$ is determined by

$$\frac{\beta b(h_1(0))}{d_I - \beta h_1(0)} = \frac{d_Y}{1-e^{-d_Y\tau}}. \tag{3.14}$$

This means by (3.13) that $C_0(\tau) < 1$ is equivalent to

$$\frac{\beta b(h_1(0))}{d_I - \beta h_1(0)} = \frac{d_Y}{1-e^{-d_Y\tau}} < \frac{\beta b(M^\tau_{\max})}{d_I - \beta M^\tau_{\max}}. \tag{3.15}$$

Since the function $f(y) = \beta b(y)/(d_I - \beta y)$ is strictly increasing for $y \in (-\infty, d_I/\beta)$, from (3.15) we have the desired result that $h_1(0) < M^\tau_{\max} \Longleftrightarrow C_0(\tau) < 1$, and the proof is complete.

REMARK 3.2. *Inequality* (3.12) *can be rewritten as*

$$\frac{d_I}{\beta} < M^\tau_{\max} + b(M^\tau_{\max})\frac{(1-e^{-d_Y\tau})}{d_Y}. \tag{3.16}$$

The right side of (3.16) *is the sum of the population of the mature and the immature foxes when they reach equilibria in the disease-free case. This sum is the carrying capacity of the environment. The left-hand side d_I/β is the critical minimum threshold fox density; see* [14]. *Theorem* 3.1 *means that when the carrying capacity of the environment is greater than the critical threshold-value d_I/β, the rabid foxes and the susceptible foxes can coexist and a positive equilibrium exists.*

4. Traveling wave solutions. In this section, we consider the behavior of solutions to system (2.7)–(2.8) in unbounded domain $(-\infty, \infty)$ under the conditions in Theorems 3.1. In section 4.1, we use the standard stability analysis to investigate possible patterns of traveling waves. An explicit formula for the minimal wave speed is given and this wave solution is confirmed by numerical simulations in section 4.3. In section 4.2, we prove that traveling wavefronts with large wave speeds indeed exist by using perturbation analysis developed in [8].

4.1. Local analysis of the traveling wavefronts. Standard stability analysis is employed here to discuss the existence of traveling wavefronts. As usual, we linearize the wave equation of (2.7)–(2.8) near their equilibria and find the associated eigenvalues and eigenvectors. Sketching this information in the system's phase plane yields a useful suggestion about a possible heteroclinic connection between these equilibria. We show the details as follows.

First of all, we linearize (2.7)–(2.8) around its equilibrium (J_0, M_0). Recall that when $J \equiv J_0$, $M \equiv M_0$, we have $S(t, t-s, x) = F(t-s, M_0, J_0)$. Assume that

$$J = J_0 + \Delta J, \quad M = M_0 + \Delta M, \quad S(t, t-s, x) = F(t-s, M_0, J_0) + \Delta S.$$

We first obtain the following linearized system for ΔS:

$$\begin{cases} \frac{\partial \Delta S}{\partial t} = D_Y \frac{\partial^2}{\partial x^2}\Delta S - d_Y \Delta S - \beta J_0 \Delta S - \beta F(t-s, M_0, J_0)\Delta J, \\ \Delta S|_{t=s} = b'(M_0)\Delta M. \end{cases} \tag{4.1}$$

We then use Fourier transforms to solve this equation. Let

$$\Delta\tilde{S} = \int_{-\infty}^{\infty} \Delta S\, e^{i\omega y} dy,$$

and let f be the Fourier transform of the term $-\beta F(t-s, M_0, J_0)\Delta J$, that is,

$$f = -\beta F(t-s, M_0, J_0) \int_{-\infty}^{\infty} \Delta J\, e^{i\omega y} dy.$$

Then, after taking Fourier transforms to both sides of (4.1), we arrive at a new linear equation for $\Delta\tilde{S}$ that can be solved easily to yield

$$\begin{aligned}
\Delta\tilde{S} &= e^{-(D_Y\omega^2+d_Y+\beta J_0)(t-s)} \int_{-\infty}^{\infty} b'(M_0)\Delta M(s,y)e^{i\omega y}dy \\
&\quad + \int_s^t f\, e^{-\int_{u-s}^{t-s}(D_Y\omega^2+d_Y+\beta J_0)dv} du \\
&= e^{-(D_Y\omega^2+d_Y+\beta J_0)(t-s)} \int_{-\infty}^{\infty} b'(M_0)\Delta M(s,y)e^{i\omega y}dy \\
&\quad - \beta \int_s^t F(u-s, M_0, J_0) \int_{-\infty}^{\infty} \Delta J(u,y)\, e^{i\omega y} dy\, e^{-(D_Y\omega^2+d_Y+\beta J_0)(t-u)} du.
\end{aligned}$$

We now take inverse Fourier transforms to obtain

$$\begin{aligned}
\Delta S(t,s,x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} b'(M_0)\Delta M(s,y)e^{i\omega y}dy\, e^{-(D_Y\omega^2+d_Y+\beta J_0)(t-s)} e^{-i\omega x} d\omega \\
&\quad - \frac{\beta}{2\pi} \int_{-\infty}^{\infty}\int_s^t F(u-s, M_0, J_0) \\
&\quad \times \int_{-\infty}^{\infty} \Delta J(u,y)\, e^{-(D_Y\omega^2+d_Y+\beta J_0)(t-u)+i\omega(y-x)} dy\, du d\omega \\
&= \frac{b'(M_0)}{\sqrt{4\pi D_Y(t-s)}} e^{-(d_Y+\beta J_0)(t-s)} \int_{-\infty}^{\infty} \Delta M(s,y) e^{-(x-y)^2/(4D_Y(t-s))} dy \\
&\quad - \frac{\beta}{2\pi} \int_{-\infty}^{\infty} dy \int_s^t F(u-s, M_0, J_0)\Delta J(u,y) e^{-(d_Y+\beta J_0)(t-u)} \\
&\quad \times \int_{-\infty}^{\infty} e^{-\omega^2 D_Y(t-u)+i\omega(y-x)} du d\omega \\
&= \frac{b'(M_0)}{\sqrt{4\pi D_Y(t-s)}} e^{-(d_Y+\beta J_0)(t-s)} \int_{-\infty}^{\infty} \Delta M(t-a,y) e^{-(x-y)^2/(4D_Y(t-s))} dy \\
&\quad - \beta \int_{-\infty}^{\infty} dy \int_0^{t-s} F(t-v-s, M_0, J_0)\Delta J(t-v,y) \\
&\quad \times e^{-(d_Y+\beta J_0)v} \frac{e^{-(x-y)^2/(4D_Y v)}}{\sqrt{4\pi D_Y v}} dv
\end{aligned}$$

and

$$\Delta S(t,a,x) = \frac{b'(M_0)}{\sqrt{4\pi D_Y a}} e^{-(d_Y+\beta J_0)a} \int_{-\infty}^{\infty} \Delta M(t-a,y) e^{-(x-y)^2/(4D_Y a)} dy$$
$$- \beta \int_{-\infty}^{\infty} dy \int_0^a F(a-v, M_0, J_0) \Delta J(t-v,y)$$
$$\times e^{-(d_Y+\beta J_0)v} \frac{e^{-(x-y)^2/(4D_Y v)}}{\sqrt{4\pi D_Y v}} dv.$$

Thus we obtain from (2.4) and (2.6) the following linearized system:

$$\text{(4.2)} \quad \begin{cases} \frac{\partial \Delta J}{\partial t} = D_I \frac{\partial^2 \Delta J}{\partial x^2} + \beta M_0 \Delta J + \beta J_0 \Delta M - d_I \Delta J + \beta \Delta J \int_0^\tau F(a, M_0, J_0) da \\ \qquad + \beta J_0 \int_0^\tau \Delta S(t,a,x) da, \\ \frac{\partial \Delta M}{\partial t} = -\beta M_0 \Delta J - \beta J_0 \Delta M - d_S \Delta M + \Delta S(t,\tau,x). \end{cases}$$

Near the equilibrium $(J,M) = (0, M^\tau_{\max})$, it gives

$$\text{(4.3)} \quad \begin{cases} \frac{\partial \Delta J}{\partial t} = D_I \frac{\partial^2 \Delta J}{\partial x^2} + \beta M^\tau_{\max} \Delta J - d_I \Delta J + \beta b(M^\tau_{\max}) \frac{1-e^{d_Y \tau}}{d_Y} \Delta J, \\ \frac{\partial \Delta M}{\partial t} = -\beta M^\tau_{\max} \Delta J - d_S \Delta M + \Delta S(t,\tau,x), \end{cases}$$

where

$$\Delta S(t,\tau,x) = \frac{b'(M^\tau_{\max})}{\sqrt{4\pi D_Y \tau}} e^{-d_Y \tau} \int_{-\infty}^{\infty} \Delta M(t-\tau,y) e^{-(x-y)^2/(4D_Y \tau)} dy$$
$$- \beta \int_{-\infty}^{\infty} dy \int_0^\tau F(\tau - v, M_0, J_0) \Delta J(t-v,y) e^{-d_Y v} \frac{e^{-(x-y)^2/(4D_Y v)}}{\sqrt{4\pi D_Y v}} dv.$$

Looking for a traveling wavefront $\Delta J = f_1(x+ct)$, $\Delta M = g(x+ct)$, we have from (4.3) that

$$\text{(4.4)} \quad \begin{cases} c f_1' = D_I f_1'' + f_1(\beta M^\tau_{\max} - d_I + \frac{\beta b(M^\tau_{\max})}{d_Y}(1 - e^{-d_Y \tau})), \\ c g' = -\beta M^\tau_{\max} f_1 - d_S g + \frac{b'(M^\tau_{\max})}{\sqrt{4\pi D_Y \tau}} e^{-d_Y \tau} \int_{-\infty}^{\infty} g(y - c\tau) e^{-(\xi - y)^2/(4D_Y \tau)} dy \\ \qquad - \beta \int_{-\infty}^{\infty} dy \int_0^\tau F(\tau - v, M_0, J_0) f_1(y - cv) e^{-d_Y v} \sqrt{\frac{1}{4\pi D_Y v}} e^{-(\xi - y)^2/(4D_Y v)} dv. \end{cases}$$

This is a linear system of functional differential equations with mixed arguments. The corresponding eigenvalues are determined by either

$$\text{(4.5)} \qquad \lambda^2 - \frac{c}{D_I}\lambda + \frac{k_1}{D_I} = 0$$

or

$$-d_S + e^{-d_Y \tau} b'(M^\tau_{\max}) e^{\alpha \lambda^2 - \lambda c \tau} = c\, \lambda,$$

where

$$k_1 = \beta M^\tau_{\max} - d_I + \frac{\beta b(M^\tau_{\max})}{d_Y}(1 - e^{-\tau d_Y}).$$

Solving (4.5) yields

$$\lambda_{1,2} = \frac{c \pm \sqrt{c^2 - 4k_1 D_I}}{2D_I}.$$

The corresponding eigenvectors to the following system, which is equivalent to (4.4) by letting $f_2 = f_1'$,

$$\begin{cases} f_1' = f_2, \\ D_I f_2' = cf_2 - f_1(\beta M^\tau_{\max} - d_I + \frac{\beta b(M^\tau_{\max})}{d_y}(1 - e^{-d_y\tau})), \\ cg' = -\beta M^\tau_{\max} f_1 - d_S g + \frac{b'(M^\tau_{\max})}{\sqrt{4\pi D_Y \tau}} e^{-d_Y\tau} \int_{-\infty}^{\infty} g(y - c\tau) e^{-(\xi-y)^2/(4D_Y\tau)} dy \\ \quad - \beta \int_{-\infty}^{\infty} dy \int_0^\tau S_0(\tau - v) f_1(y - cv) e^{-d_Y v} \sqrt{\frac{1}{4\pi D_Y v}} e^{-(\xi-y)^2/(4D_Y v)} dv \end{cases}$$

are

$$\vec{v}_1 = \begin{pmatrix} 1 \\ \lambda_1 \\ 0 \end{pmatrix}, \ \vec{v}_2 = \begin{pmatrix} 1 \\ \lambda_2 \\ 0 \end{pmatrix}.$$

When

$$0 < c < 2\sqrt{k_1 D_I},$$

the eigenvalues $\lambda_{1,2}$ are complex and the eigensolutions are oscillatory and can be negative. This is not biologically meaningful. Therefore, a natural condition for the existence of traveling wavefronts starting from $(0, M^\tau_{\max})$ is

$$\begin{aligned} c \geq c_{\min}(\tau) &:= 2\sqrt{\beta M^\tau_{\max} D_I} \sqrt{1 - \frac{d_I}{\beta M^\tau_{\max}} + \frac{b(M^\tau_{\max})}{M^\tau_{\max} d_Y}(1 - e^{-d_Y\tau})} \\ &= 2\sqrt{\beta M^\tau_{\max} D_I} \sqrt{1 - C_0(\tau)}. \end{aligned} \tag{4.6}$$

We should mention that the minimal speed can also be expressed as

$$c_{\min}(\tau) = 2\sqrt{\beta D_I} \sqrt{M^\tau_{\max} + b(M^\tau_{\max}) \frac{1 - e^{-d_Y\tau}}{d_Y} - \frac{d_I}{\beta}},$$

from which we find the speed $c_{\min}$ depends not only on the diffusive coefficient D_I and the transmission rate β, but also on the difference between the carrying capacity $M^\tau_{\max} + b(M^\tau_{\max}) \frac{1-e^{-d_Y\tau}}{d_Y}$ and the critical threshold value d_I/β.

We now argue that it is impossible for a positive trajectory to go from $(0, M^\tau_{\max})$ to $(0, 0)$. To see this, linearizing around $(0, 0)$, we obtain

$$\begin{cases} \frac{\partial \Delta J}{\partial t} = D_I \frac{\partial^2 \Delta J}{\partial x^2} - d_I \Delta J, \\ \frac{\partial \Delta M}{\partial t} = -d_S \Delta M + \frac{b'(0)}{\sqrt{4\pi D_Y \tau}} e^{-d_Y\tau} \int_{-\infty}^{\infty} \Delta M(t - \tau, y) e^{-(x-y)^2/(4D_Y\tau)} dy. \end{cases}$$

This gives, by substituting $\Delta J = f_1(x + ct)$, $\Delta M = g(x + ct)$, the following:

$$\begin{cases} cf_1' = D_I f_1'' - d_I f_1, \\ cg' = -d_S g + \frac{b'(0)}{\sqrt{4\pi D_Y \tau}} e^{-d_Y\tau} \int_{-\infty}^{\infty} g(y - c\tau) e^{-(\xi-y)^2/(4D_Y\tau)} dy. \end{cases} \tag{4.7}$$

Thus at $(0, 0)$, the eigenvalues satisfy

$$\left[\lambda\left(\lambda - \frac{c}{D_I}\right) - \frac{d_I}{D_I}\right]\left[\frac{1}{c}\left(-d_S + e^{-d_Y\tau} b'(0) e^{\alpha\lambda^2 - \lambda c\tau}\right) - \lambda\right] = 0. \tag{4.8}$$

The second factor corresponds to the second equation of (4.7) that is in fact decoupled from the first equation of (4.7).

By (3.9) it is easy to see that every eigenvalue to equation

$$\frac{1}{c}\left(-d_S + e^{-d_Y\tau}b'(0)e^{\alpha\lambda^2-\lambda c\tau}\right) - \lambda = 0$$

cannot be negative and real, and hence there is no positive solution g such that $\lim_{t\to\infty} g(t) = 0$. This means that there's no positive orbit of (2.7)–(2.8) starting from $(0, M^\tau_{\max})$ and approaching $(0,0)$.

So the solution starting from $(0, M^\tau_{\max})$ could arrive at (J^τ_*, M^τ_*) under the condition (4.6). The asymptotic behavior of traveling wavefronts approaching (J^τ_*, M^τ_*) depends on eigenvalues of system (4.2) near the equilibrium (J^τ_*, M^τ_*). If all the eigenvalues with negative real parts are complex, then the traveling wave will tend to (J^τ_*, M^τ_*) with oscillatory damping. Otherwise it will approach (J^τ_*, M^τ_*) monotonically. We will see numerical evidence for oscillatory damping of wave patterns in later sections.

4.2. A rigorous proof of traveling wavefronts with large wave speeds. In this section, the existence of traveling wavefronts is rigorously established for system (2.7)–(2.8). To present our result, we first show the existence of a heteroclinic connection for a nondiffusive delayed system and then show that this is perturbed to a traveling wavefront with large wave speed for (2.7)–(2.8).

4.2.1. Heteroclinic connection for a nondiffusion delay system. We now study the heteroclinic connection of the delayed system

$$\begin{cases} \frac{dJ}{dt} = \beta MJ - d_I J + \beta J\int_0^\tau b(M(t-a))e^{\int_0^a -(d_Y+\beta J(t-s))ds}da, \\ \frac{dM}{dt} = -\beta MJ - d_S M + b(M(t-\tau))e^{\int_{t-\tau}^t -(d_Y+\beta J(s))ds}, \end{cases} \tag{4.9}$$

which is a reduced version of (2.7)–(2.8) when $D_I = D_Y = 0$. It is easy to see that (4.9) has three equilibria: $E_1 := (0,0)$, $E_2 := (0, M^\tau_{\max})$, and $E_3 := (J^\tau_*, M^\tau_*)$.

For initial continuous data $(J, M) = (j_0(s), m_0(s)) \geq 0$ for $s \in [-\tau, 0]$ with $(j_0(0), m_0(0)) > 0$, we claim that

$$(J(t), M(t)) > 0$$

for all $t > 0$. Indeed, dividing the first equation in (4.9) by J and integrating it from 0 to t, we have

$$J(t) = J(0)\exp\left(\beta M - d_I + \beta\int_0^\tau b(M(t-a)e^{\int_0^a -(d_Y+\beta J(t-s))ds}\right) > 0.$$

We then use the variation-of-constants formula in consecutive interval $[0,\tau], [\tau, 2\tau], \ldots$ to obtain

$$M(t) > 0$$

for $t \geq 0$.

When $\tau = 0$, the above system reduces to the ODE system

$$\begin{cases} \frac{dJ}{dt} = \beta MJ - d_I J, \\ \frac{dM}{dt} = -\beta MJ - d_S M + b(M). \end{cases} \tag{4.10}$$

Obviously, the three equilibria reduces to $E_1 = (0,0)$, $E_2 = (0, M^0_{\max})$, and $E_3 = (J^0_*, M^0_*) = (\frac{1}{\beta}(-d_S + b_0 e^{-\bar{a}d_I/\beta}), \frac{d_I}{\beta})$, and we have $J^0_* > 0$ if and only if $C_0(0) = \frac{d_I}{\beta M^0_{\max}} < 1$.

THEOREM 4.1. *When $\tau = 0$ and $C_0(0) = \frac{d_I}{\beta M^0_{\max}} < 1$, system (4.9) has a heteroclinic orbit $(J_0(t), M_0(t))$ connecting E_2 and E_3.*

Proof. First, we prove that the third equilibrium E_3 is a global attractor in the sense that it attracts every positive solution of (4.9) when $\tau = 0$. To see this, define a Lyapunov function as

$$V = \left[M - M^0_* - M^0_* \log \frac{M}{M^0_*}\right] + \left[J - J^0_* - J^0_* \log \frac{J}{J^0_*}\right].$$

Differentiating the function V along the solution (4.10) yields

$$\frac{dV}{dt} = b_0(e^{-\bar{a}M} - e^{-\bar{a}d_I/\beta})\left(M - \frac{d_I}{\beta}\right) < 0$$

provided that $M \neq d_I/\beta$. This means that the equilibrium E_3 is a global attractor by LaSalle's well-known invariance principle. Linearizing (4.10) around E_2 gives

$$\begin{pmatrix} \beta M^0_{\max}(1 - C_0(0)) - \lambda & 0 \\ -\beta M^0_{\max} & -d_S + b'(M^0_{\max}) - \lambda \end{pmatrix}$$

and the following characteristic equation:

$$(\beta M^0_{\max} - d_I - \lambda)(-d_S + b'(M^0_{\max}) - \lambda) = 0. \tag{4.11}$$

For $\lambda_1 = \beta M^0_{\max}(1 - C_0(0)) > 0$, we find an eigenvector $\vec{v}_1$ which points into the first quadrant of the $J - M$ plane. Therefore, the solution starting from the local unstable manifold of E_2 along the $\vec{v}_1$ direction will permanently stay in the first quadrant and tends to $(J^0_*, M^0_*)^T$ as $t \to \infty$ due to the global attractivity of E_3. This completes the proof.

When $\tau \neq 0$, deriving the global stability of the equilibrium E_3 is nontrivial. Even for the local stability, providing an explicit criterion is not easy. To demonstrate this, we linearize (4.9) around E_3 to obtain

$$\begin{aligned} \frac{dJ}{dt} = &\left(\beta M^\tau_* - d_I + \beta b(M^\tau_*)\frac{1 - e^{-(d_Y + \beta J^\tau_*)\tau}}{(d_Y + \beta J^\tau_*)}\right) J \\ &- \beta^2 J^\tau_* b(M^\tau_*) \int_0^\tau e^{-(d_Y + \beta J^\tau_*)a} \int_0^a J(t-s)dsda + \beta J^\tau_* M \\ &+ \beta J^\tau_* b'(M^\tau_*) \int_0^\tau M(t-a) e^{-(d_Y + \beta J^\tau_*)a} da \end{aligned} \tag{4.12}$$

and

$$\begin{aligned} \frac{dM}{dt} = &-\beta M^\tau_* J - \beta b(M^\tau_*) e^{-(d_Y + \beta J^\tau_*)\tau} \int_0^\tau J(t-s)ds \\ &+ \beta J^\tau_* M - d_S M + b'(M^\tau_*) e^{-(d_Y + \beta J^\tau_*)\tau} M(t-\tau). \end{aligned} \tag{4.13}$$

In order to understand the linear system (4.12) and (4.13), we first consider the special case when $\tau = 0$. In this case the characteristic equation is given by

$$\lambda^2 + (\beta J^0_* + d_S - b'(M^0_*))\lambda + \beta^2 M^0_* J^0_* = 0.$$

Since (4.10) yields $\beta J_*^0 + d_S = b(M_*^0)/M_*^0$, the above equation becomes

$$\lambda^2 + \bar{a}b_0 e^{-\bar{a}M_*^0}\lambda + \beta^2 M_*^0 J_*^0 = 0.$$

Hence, the eigenvalues are given by

$$\lambda_{1,2} = \frac{-\bar{a}b_0 e^{-\bar{a}M_*^0} \pm \sqrt{(\bar{a}b_0)^2 e^{-2\bar{a}M_*^0} - 4\beta^2 M_*^0 J_*^0}}{2}, \tag{4.14}$$

the real parts of which are negative as long as $J_*^0 > 0$ (or, equivalently, $C_0(0) < 1$). Because λ depends continuously on the parameter τ, we conclude that there exists a number $\tau_1 > 0$ so that when $\tau < \tau_1$, all the eigenvalues of the linearization at E_3 have a negative real part.

At the equilibrium E_2 when $\tau \neq 0$, we have the following characteristic equation:

$$(1 - C_0(\tau) - \lambda)\,(-d_S + b'(M_{\max}^\tau)e^{-d_Y\tau}e^{-\lambda\tau} - \beta M_{\max}^\tau \lambda) = 0. \tag{4.15}$$

It can be shown easily that there exists a constant τ_2 so that equilibrium E_2 is hyperbolic for $C(\tau) < 1$ and $\tau \in [0, \tau_2)$, where τ_2 is the first positive number satisfying

$$|b'(M_{\max}^{\tau_2})e^{-dy\tau_2}| > |d_S|, \quad \text{and} \quad \tau_2 = \frac{\pi - \arccos\frac{d_s}{|b'(M_{\max}^{\tau_2})e^{-d_y\tau_2}|}}{\sqrt{(b'(M_{\max}^{\tau_2})e^{-d_y\tau_2})^2 - d_s^2}}. \tag{4.16}$$

We should mention that formula (4.16) can be obtained by the well-known Hopf bifurcation theory, and that if there is no τ_2 satisfying (4.16), then we assume that $\tau_2 = \infty$.

With the above preparation, we are now ready to prove a theorem concerning the heteroclinic connection for (4.9) when $\tau \neq 0$. To present our result, we first introduce some notation.

- For a vector $x \in R^2$, we denote $||x|| = ||x||_{R^2}$.
- Let $X(R, R^2)$ be the space of continuous and bounded functions from R to R^2 equipped with the standard norm $||\phi|| = \sup\{||\phi(t)||,\, t \in R\}$.
- Let $X^1 = X^1(R, R^2) = \{\phi \in X : \phi' \in X\}$.
- Let $X_0 = \{\phi \in X : \lim_{t\to\pm\infty}\phi = 0\}$ and $X_0^1 = \{\phi \in X_0 : \phi' \in X_0\}$.

Under the conditions in Theorem 3.1, we have the following result.

THEOREM 4.2. *Assume that $C_0(\tau) < 1$. Then there exists a positive constant δ so that for $0 \le \tau \le \delta$, equation (4.9) has a heteroclinic orbit $(J(t), M(t))$ which connects E_2 and E_3.*

Proof. We first introduce the transformation

$$U = \frac{J(t)}{J_*^\tau}, \quad V = \frac{M_{\max}^\tau - M}{M_{\max}^\tau - M_*^\tau}$$

to get rid of the τ-dependence of E_2 and E_3. Substituting this into (4.9), we have the following system for U and V:

$$\begin{cases} \frac{dU}{dt} = \beta(M_{\max}^\tau - V(M_{\max}^\tau - M_*^\tau))U - d_I U \\ \qquad + \beta U \int_0^\tau \bar{b}(V(t-a)e^{\int_0^a -(d_Y + \beta J_*^\tau U(t-s))ds}da, \\ \frac{dV}{dt} = \frac{\beta J_*^\tau (M_{\max}^\tau - V(M_{\max}^\tau - M_*^\tau))U}{M_{\max}^\tau - M_*^\tau} + \frac{d_S(M_{\max}^\tau - V(M_{\max}^\tau - M_*^\tau))}{M_{\max}^\tau - M_*^\tau} \\ \qquad - \frac{\bar{b}(V(t-\tau)}{M_{\max}^\tau - M_*^\tau} e^{\int_{t-\tau}^t -(d_Y + \beta J_*^\tau U(s))ds}, \end{cases} \tag{4.17}$$

where $\bar{b}(V(t-\tau)) = b(M^\tau_{\max} - V(M^\tau_{\max} - M^\tau_*))$. Equation (4.17) has two equilibria $E_2 := (0,0)$ and $E_3 := (1,1)$. In particular when $\tau = 0$, we know from Theorem 4.1 that there exists a heteroclinic solution $(U_0(t), V_0(t))$ that connects two points E_2 and E_3 and satisfies

(4.18)
$$\begin{cases} \frac{dU_0}{dt} = F_1(u,v)|_{u=U_0,v=V_0} = \beta(M^0_{\max} - V_0(M^0_{\max} - M^0_*))U_0 - d_I U_0, \\ \frac{dV_0}{dt} = F_2(u,v)|_{u=U_0,v=V_0} = \frac{\beta J^0_*(M^0_{\max} - V_0(M^0_{\max} - M^0_*))V_0}{M^0_{\max} - M^0_*} + \frac{d_S(M^0_{\max} - V(M^0_{\max} - M^0_*))}{M^0_{\max} - M^0_*} \\ \qquad\qquad - \frac{\bar{b}(V_0(t))}{M^0_{\max} - M^0_*}. \end{cases}$$

Note that the relation between $M^\tau_{\max}$ and $M^0_{\max}$, M^τ_* and M^0_*, and J^τ_* and J^0_* can be described as

$$M^\tau_{\max} = M^0_{\max} + O(\tau), \quad M^\tau_* = M^0_* + O(\tau), \quad J^\tau_* = J^0_* + O(\tau). \tag{4.19}$$

We now show that there exists a constant δ such that (4.17) has a heteroclinic orbit $(U(t), V(t))$ connecting two points E_2 and E_3 provided $\tau < \delta$.

First of all, we let $W_1 = U - U_0$ and $W_2 = V - V_0$ and obtain the following equation for the remainder (W_1, W_2):

$$\begin{cases} \frac{dW_1}{dt} = \frac{\partial F_1(U_0,V_0)}{\partial u} W_1 + \frac{\partial F_1(U_0,V_0)}{\partial v} W_2 + \Gamma_1(t,\tau,W_1,W_2), \\ \frac{dW_2}{dt} = \frac{\partial F_2(U_0,V_0)}{\partial u} W_1 + \frac{\partial F_2(U_0,V_0)}{\partial v} W_2 + \Gamma_2(t,\tau,W_1,W_2), \end{cases} \tag{4.20}$$

where

$$\begin{aligned} &\Gamma_1(t,\tau,W_1,W_2) \\ &= \beta(M_{\max} - (V_0+W_2)(M^\tau_{\max} - M^\tau_*))(U_0+W_1) - d_I(U_0+W_1) \\ &\quad + \beta(U_0+W_1)\int_0^\tau \bar{b}(V_0(t-a) + W_1(t-a)) e^{\int_0^a -(d_Y + \beta J^\tau_*(U_0(t-s)+W(t-s)))ds} da \\ &\quad - F_1(U_0,V_0) - \left(\frac{\partial F_1(U_0,V_0)}{\partial u} W_1 + \frac{\partial F_1(U_0,V_0)}{\partial v} W_2\right) \end{aligned} \tag{4.21}$$

and

$$\begin{aligned} \Gamma_2(t,\tau,W_1,W_2) &= \frac{\beta J^\tau_*(M^\tau_{\max} - (V_0+W_2)(M^\tau_{\max} - M^\tau_*))(U_0+W_1)}{M^\tau_{\max} - M^\tau_*} \\ &\quad + \frac{d_S(M^\tau_{\max} - (V_0+W_2)(M^\tau_{\max} - M^\tau_*))}{M^\tau_{\max} - M^\tau_*} \\ &\quad - \frac{\bar{b}(V_0(t-\tau) + W_2(t-\tau))}{M^\tau_{\max} - M^\tau_*} e^{\int_{t-\tau}^t -(d_Y + \beta J^\tau_*(U_0(s)+W_1(s)))ds} \\ &\quad - F_2(U_0,V_0) - \left(\frac{\partial F_2(U_0,V_0)}{\partial u} W_1 + \frac{\partial F_2(U_0,V_0)}{\partial v} W_2\right). \end{aligned} \tag{4.22}$$

Define an operator $T : \Psi \in X^1 \to X$ from the homogeneous part of (4.20) as follows:

$$T\Psi = \Psi' - A(t)\Psi, \ t \in R, \tag{4.23}$$

where

$$A(t) = \begin{pmatrix} \frac{\partial F_1(U_0(t),V_0(t))}{\partial u} & \frac{\partial F_1(U_0(t),V_0(t))}{\partial v} \\ \frac{\partial F_2(U_0(t),V_0(t))}{\partial u} & \frac{\partial F_2(U_0(t),V_0(t))}{\partial v} \end{pmatrix}.$$

We remark that $(U_0(t), V_0(t))$ tends, respectively, to E_2 and E_3 when $t \to -\infty$ and $t \to \infty$. This means that the linear operator T is asymptotically hyperbolic as $t \to \pm\infty$ in the sense that

$$\Psi' - A(-\infty)\Psi = 0 \quad \text{and} \quad \Psi' - A(\infty)\Psi = 0$$

are hyperbolic due to (4.11) and (4.14). Furthermore, we know that every eigenvalue for the linear equation $\Psi' - A(\infty)\Psi = 0$ has a negative real part. Define the formal adjoint equation of $T\Psi = \Psi' - A(t)\Psi = 0$ as

$$\Phi' + A^T(t)\Phi = 0, \quad t \in R. \tag{4.24}$$

We now divide our proof into five steps.

Step 1. We claim that if $\Phi \in X$ is a solution of (4.24) and Φ is C^1-smooth, then $\Phi = 0$. Moreover, we have $R(T) = X$, where $R(T)$ is the range of T.

Indeed, assuming to the contrary that Φ is not zero at some point t_0, then we can solve (4.24) to obtain

$$\Phi(t) = \Phi(t_0)e^{-\int_{t_0}^{t} A^T(t)dt}.$$

Since when $t \to \infty$, $A^T(t)$ tends to $A^T(\infty)$ whose eigenvalues are negative, we deduce that

$$\lim_{t\to\infty} \Phi(t) = \infty,$$

which contradicts the fact that Φ is bounded.

By the classical Fredholm theory, this claim means further that $R(T) = X$ in the sense that for any $\Theta \in X$, there exists $\Psi \in X^1$ so that

$$T\Psi = \Theta.$$

Step 2. Let $\Theta \in X_0$ be given. If Ψ is a bounded solution of $T\Psi = \Theta$, then $\Psi \in X_0^1$. In fact, we need to show only that

$$\lim_{t\to\pm\infty} \Psi(t) = 0.$$

Actually when $t \to \infty$, equation

$$\Psi' - A(t)\Psi = \Theta \tag{4.25}$$

asymptotically tends to

$$\Psi' - A(\infty)\Psi = 0. \tag{4.26}$$

Note that for (4.26), the ω-limit set of every bounded solution is just the critical point $\Psi = 0$. Using the result from [21] or [18], every bounded solution of (4.25) also satisfies

$$\lim_{t\to\infty} \Psi(t) = 0.$$

When inverting the time from $-t$ to t, we can similarly prove that

$$\lim_{t\to-\infty}\Psi(t)=0.$$

Step 3. We rewrite (4.20) as

$$W'(t)+W=W+A(t)W+\Gamma(t), \tag{4.27}$$

where

$$W=(W_1,W_2)^T,\ \ \Gamma(t)=(\Gamma_1(s,\tau,W_1,W_2),\Gamma_2(s,\tau,W_1,W_2))^T.$$

Changing (4.27) into an integral equation gives

$$W(t)=\int_{-\infty}^{t}e^{-(t-s)I}(W(s)+A(s)W(s)+\Gamma(s))ds, \tag{4.28}$$

where I is the 2×2 identity matrix and $W(t)=(W_1(t),W_2(t))^T$.

Define a linear operator $L: X_0\to X_0$ as follows:

$$L(W)(t)=W(t)-\int_{-\infty}^{t}e^{-(t-s)I}(W(s)+A(s)W(s))ds,\ W\in X_0.$$

Obviously $L(W)\in X_0$ if $W\in X_0$. Now we prove that $R(L)=X_0$, that is, for each $Z\in X_0$, we can have a $W\in X_0$ so that

$$W(t)-\int_{-\infty}^{t}e^{-(t-s)I}(W(s)+A(s)W(s))ds=Z(t).$$

To see this, assuming that $\xi=W-Z$, we obtain an equation for ξ as follows:

$$\xi(t)=\int_{-\infty}^{t}e^{-(t-s)I}(\xi(s)+A(s)\xi(s))ds+\int_{-\infty}^{t}e^{-(t-s)I}(Z(s)+A(s)Z(s))ds.$$

Differentiating both sides yields

$$T(\xi)(t)=\xi'(t)-A(t)\xi(t)=Z(t)+A(t)Z(t). \tag{4.29}$$

Using the results that $R(T)=X$ in Step 2, one can obtain that there exists a solution ξ for (4.29) and $\xi\in X_0^1$. Returning to the variable W, we have $W=\xi+Z\in X_0$.

Step 4. Let $N(L)$ be the null space of the operator L. Define $N^{\perp}(L)=X_0/N(L)$. It is clear that $N^{\perp}(L)$ is a Banach space. If we let $S=L|_{N^{\perp}(L)}$ be the restriction of L on $N^{\perp}(L)$, then $S: N^{\perp}(L)\to X_0$ is one-to-one and onto. By the well-known Banach inverse operator theorem, we have that $S^{-1}: X_0\to X_0/N(L)$ is a linear bound operator.

Step 5. When L is restricted on $N^{\perp}(L)$, equation (4.28) can be written as

$$S(W)(t)=\int_{-\infty}^{t}e^{-(t-s)I}\Gamma(s,W,\tau)ds$$

or

$$W(t)=S^{-1}\left(\int_{-\infty}^{t}e^{-(t-s)I}\Gamma(s,W,\tau)ds\right). \tag{4.30}$$

The term $\int_{-\infty}^{t} e^{-(t-s)I}\Gamma(s)ds$ on the right-hand side can be estimated. Actually when τ is small and $W \in X_0^1$, from (4.19), (4.21), and (4.22), we have the following estimations:

$$\left|\int_{-\infty}^{t} e^{-(t-s)}\Gamma_1(s)ds\right| = O(\tau) + O(\tau||W||_{X_0}) + O(||W||_{X_0}^2) \tag{4.31}$$

and

$$\left|\int_{-\infty}^{t} e^{-(t-s)}\Gamma_2(s)ds\right| = O(\tau) + O(\tau||W||_{X_0}) + O(||W||_{X_0}^2) \tag{4.32}$$

as $\tau \to 0$ and $||W|| \to 0$. To derive (4.31) and (4.32), we have made use of the following result:

$$\int_{-\infty}^{t} e^{-(t-s)}\left(W_i(s-\tau)) - W_i(s)\right) ds = O(\tau||W||), \quad i = 1, 2. \tag{4.33}$$

Actually, if $W \in X_0^1$, by exchanging the order integration and by integration by parts, we have

$$\begin{aligned}
&\left|\int_{-\infty}^{t} e^{-(t-s)}\big(W_i(s-\tau) - W_i(s)\big)ds\right| \\
&= \left|\tau\int_{-\infty}^{t} e^{-(t-s)}\int_0^1 W_i'(s-\tau u)duds\right| \\
&= \left|\tau\int_0^1\int_{-\infty}^{t} e^{-(t-s)}W_i'(s-\tau u)dsdu\right| \\
&= \left|\tau\int_0^1\left(W_i(t-\tau u) - \int_{-\infty}^{t} e^{-(t-s)}W_i(s-\tau u)ds\right)du\right| \\
&= O(\tau||W||),\ i = 1, 2,
\end{aligned}$$

leading to (4.33). Using the fact that X_0^1 is dense in X_0, we conclude that (4.31) and (4.32) hold for any $W \in X_0$.

Let $B(\sigma)$ denote the closed ball in X_0 with radius σ and center at the origin. Since the norm $||S^{-1}||$ is independent of τ, it follows from (4.31) and (4.32) that there exist $\sigma > 0$, $\delta > 0$, and $0 < \rho < 1$ such that for all $\tau \in (0, \delta]$ and $\varphi, \psi, W \in B(\sigma) \subset X_0$,

$$\left|\left|S^{-1}\left(\int_{-\infty}^{t} e^{-(t-s)I}\Gamma(s, W, \tau)ds\right)\right|\right| \le \frac{1}{3}\left(||W|| + \sigma\right)$$

and

$$\left|\left|S^{-1}\left(\int_{-\infty}^{t} e^{-(t-s)I}\Gamma(s, \varphi, \tau)ds\right) - S^{-1}\left(\int_{-\infty}^{t} e^{-(t-s)I}\Gamma(s, \psi, \tau)ds\right)\right|\right| \le \rho||\varphi - \psi||.$$

Hence, $S^{-1}\left(\int_{-\infty}^{t} e^{-(t-s)I}\Gamma(s, W, \tau)ds\right)$ is a uniform contractive mapping of $W \in X_0 \cap B(\sigma)$. By using the classical fixed point theorem, it follows that for $\tau \in [0, \delta]$, (4.30) has a unique solution $W \in X_0/N(L)$. Returning to the original variable, we get that $(W_1 + U_0,\ W_2 + V_0)$ is a heteroclinic connection between E_2 and E_3. This completes our proof.

REMARK 4.3. *When $\tau \geq \delta$, we can rescale the time variable $t \to t\tau$ to obtain*

$$\begin{cases} \frac{dJ}{dt} = \tau\beta MJ - \tau d_I J + \tau\beta J \int_0^1 b(M(t-a)e^{\int_0^a -(d_Y+\beta J(t-s))ds} da, \\ \frac{dM}{dt} = -\tau\beta MJ - \tau d_s M + \tau b(M(t-1))e^{\int_{t-1}^t -(d_Y+\beta J(s))ds}. \end{cases} \tag{4.34}$$

At $\tau = \delta$, by Theorem 4.2*, equation* (4.34) *has a heteroclinic connection. We can as well show that if $C_0(\tau) < 1$, there exists a constant δ_1, $\delta < \delta_1 < \min\{\tau_1, \tau_2\}$, such that if $\delta \leq \tau \leq \delta_1$, equation* (4.9) *has a heteroclinic orbit $(J(t), M(t))$ which connects E_2 and E_3. The proof is the same as that of Theorem* 4.2*. The method is referred as to a homotopy approach (see* [5]*); namely, we view τ as a varying parameter and start with* (4.34)*, and extend the result from δ to $\delta_1 \in (\delta, \min\{\tau_1, \tau_2\})$ by replacing the arguments in Step* 1 *to Step* 5 *by those of the parallel theory in linear delay differential equations. It would be interesting to see how far this homotopy argument can be applied to push the upper bound τ.*

4.2.2. Traveling wavefronts with large wave speeds. We now consider the reaction diffusion system (2.7)–(2.8) for which we will use Theorem 1.1 in [8] to give traveling wavefronts in the case when the wave speed c is large. The main idea of this result is simple: if the nondiffusive equation has a heteroclinic connection between E_2 and E_3, then the diffusive system has a family of traveling wavefronts from E_2 to E_3 with large wave speeds.

THEOREM 4.4. *Assume that $\tau \leq \delta$. Then there exists a $c^* > 0$ such that for any $c \geq c^*$, system* (2.7)–(2.8) *has a traveling wavefront $(J(t,x), W(t,x)) = (u(ct+x), v(ct+x))$ which connects E_2 and E_3.*

Proof. First we observe that if there is no diffusion, that is, if $D_I = 0$ and $D_Y = 0$, our equations (2.7)–(2.8) reduce to (4.9). When $\tau \leq \delta$, the equilibria E_2 and E_3 are hyperbolic, and, in particular, all the eigenvalues to E_3 have negative real parts. From Theorem 4.2, we know that when $\tau \leq \delta$, equation (4.9) has a heteroclinic connection. So conditions (H_1), (H_2), and (H_3) in [8, Theorem 1.1] are satisfied. Last, for our kernel function $f(x) = \frac{1}{\sqrt{4\pi}} \exp(-\frac{y^2}{4})$, it is easy to see that

$$\frac{1}{\sqrt{4\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{4}\right) |y| dy < \infty.$$

So all conditions in [8, Theorem 1.1] are satisfied. Hence by [8, Theorem 1.1], we conclude that there exists a $c^* > 0$ so that for any $c > c^*$, system (2.7)–(2.8) has a traveling wavefront $(J(t,x), W(t,x)) = (u(ct+x),\ v(ct+x))$ which connects E_2 and E_3.

4.3. Numerical simulations. In this subsection, we will numerically study the traveling wavefronts of our model (2.7)–(2.8).

We first describe our numerical methods. We give initial data

$$J(s,x) = j_0(s,x),\ M(s,x) = m_0(s,x),\ -\tau \leq s \leq 0,\ x \in [-L, L],$$

and solve (2.7) and (2.8) to obtain $(J(t,x), M(t,x))$ in a sufficiently large interval $[-L, L]$ for $t \geq 0$ and some $L > 0$. As usual, in the process of finding numerical solutions, we take the homogeneous Neumann boundary conditions at the end points $x = \pm L$. Depending on other parameters in our model and the solution patterns, we may adjust the parameter L from 100 to 1000 so as to present a clear view of our graphs. We take a constant h satisfying

$$M_*^\tau < h < M_{\max}^\tau.$$

For any fixed t, we find the first position $x = z(t,h) > -L$ so that

$$M(t, z(t,h)) = h.$$

Choose a sequence $\{t_j\}_{j=1}^{\infty}$, and consider

$$(J(t_j, x + z(t_j,h)), M(t_j, x + z(t_j,h))). \tag{4.35}$$

If the numerical solutions $(J(t_j, x + z(t_j,h)),\ M(t_j, x + z(t_j,h)))$, $j = 1, 2, 3, \ldots$, converge uniformly to a nonconstant function $(J(\cdot), M(\cdot))$ which satisfies the boundary conditions

$$\lim_{\xi \to -\infty} (J(\xi), M(\xi)) = E_2 \ \text{ and } \lim_{\xi \to +\infty} (J(\xi), M(\xi)) = E_3,$$

then the limit $(J(\cdot), M(\cdot))$ is viewed as a traveling wavefront. Theoretically, this process has also been used to prove the existence of traveling wavefronts for certain monotone dynamics; see [4]. The limit

$$\lim_{j \to \infty} \frac{z(t_{j+1},h) - z(t_j,h)}{t_{j+1} - t_j} \tag{4.36}$$

is correspondingly thought of as the asymptotic wave speed of the traveling wavefront.

We now discuss the parameter values from relevant references [2, 15, 16]. First of all, we note that [16, p. 126] suggests 9 to 12 months for the maturation time, and we will therefore restrict our attention to the range of τ to $[0.5, 0.8]$ (year). The diffusion coefficient $D_I = 60$ km^2/year will be used, based on the value in [15].

For red foxes, the average per capita intrinsic death rate is 0.5 year^{-1} [2], so we take $d_S = 0.5$ year^{-1}. Since it is known that the death rate of juvenile foxes is greater than that of adult foxes, we take $d_Y = 0.8$ year^{-1} [16, p. 127].

An infective fox first goes through an incubation period that can vary from 12 to 110 days. A life expectancy of about 35 days gives d_I as approximately 10 year^{-1}. For the transmission coefficient, we derive $\beta = 10$ km^2/year by using formula (5) in [15]. The number of cubs in a litter ranges from 1 to 10, with a mean of 4.7 in Europe [2, 17, 16, 30, 31]. Sex ratios are in general close to unity at birth, and the pregnancy rate is in the region of 90% [17, 16], with a further 10% of vixens failing to produce offspring [2]. In view of this information, the average per capita birth rate b_0 is taken to be 1.9 year^{-1}.

We now calculate the minimal wave speed $c_{\min}$. The carrying capacity S_0 is assumed to be 2 foxes per km^2, as in Figure 4 of [2], and it is the sum of the population of the immature and the adult foxes when they reach the stable equilibria in the disease-free case, that is,

$$S_0 = M^{\tau}_{\max} + b(M^{\tau}_{\max}) \frac{1 - e^{-d_Y \tau}}{d_Y}.$$

We need further information in order to estimate the maturation time, which is related to the parameter $\bar{a}$ in the birth function. By Table 26 in [16], the number of adult foxes per km^2 varies from 0.5 to 1.8, and the number of litters found per km^2 varies from 0.16 to 0.6. Thus we take the mean value of the ratio of the adult foxes to the litter foxes as $1.15 : 0.38$. Using the facts $S_0 = M^{\tau}_{\max} + b(M^{\tau}_{\max}) \frac{1 - e^{-d_Y \tau}}{d_Y} = 2$ and

$$M^{\tau}_{\max} : b(M^{\tau}_{\max}) \frac{1 - e^{-d_Y \tau}}{d_Y} = 1.15 : 0.38,$$

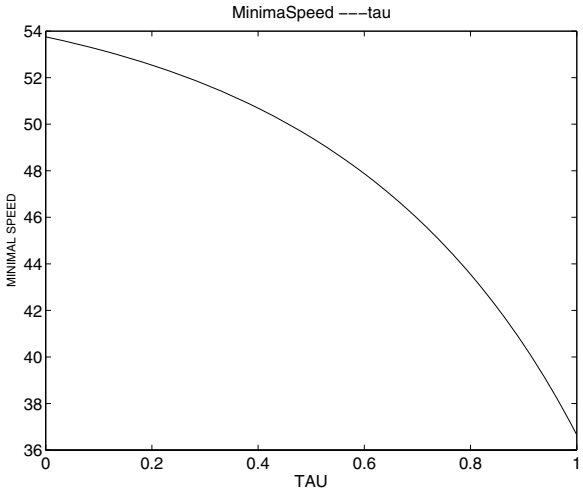


FIG. 1. *Graph of $c_{min}(\tau)$ as a decreasing function of τ.*

we calculate that the carrying capacity is 2 when the parameters $\tau = 0.5305$ and

$$\bar{a} = \frac{1}{S_0}\left(1 + \frac{d_S}{d_Y}(e^{d_Y\tau} - 1)\right)\ln\frac{b_0}{d_S e^{d_Y\tau}} = 0.6057.$$

The result calculated in [15] gives the minimal speed $c_{old} = 48.9898$ km/year. In our calculation, $c_{\min}$ is a decreasing function of τ, with $c_{\min} = 53.757$ km/year if $\tau = 0$; and

$$\begin{aligned} c_{\min} &= 2\sqrt{\beta M^\tau_{\max} D_I}\sqrt{1 - \frac{d_I}{\beta M^\tau_{\max}} + \frac{b(M^\tau_{\max})}{M^\tau_{\max} d_Y}(1 - e^{-d_Y\tau})} \\ &= 2\sqrt{\beta D_I}\sqrt{M^\tau_{\max} + b(M^\tau_{\max})\frac{(1 - e^{-d_Y\tau})}{d_Y} - \frac{d_I}{\beta}} \\ &= 43.549 \text{ km/year} \end{aligned}$$

if $\tau = 0.8$. The graph of $c_{\min}$ as a function of τ is given in Figure 1.

To describe numerically the solution patterns, we first scale the variable x by $\sqrt{D_I}x$ so that the diffusion rate for rabid foxes in our simulations becomes constant 1. The length L of the half interval is taken to be 300. We use the the Neumann boundary condition and the initial values

$$M(t,x) = \begin{cases} M^\tau_{\max}, & -300 \le x \le 150,\ \tau \le t \le 0, \\ 0.6, & 150 < x \le 300,\ \tau \le t \le 0, \end{cases}$$

and

$$J(t,x) = \begin{cases} 0, & -300 \le x \le 150,\ \tau \le t \le 0, \\ 0.05, & 150 < x \le 300,\ \tau \le t \le 0. \end{cases}$$

A finite difference method coupled with iterative techniques is used in our numerical approximation via the software MATLAB, and the numerical result when $D_Y = 0$

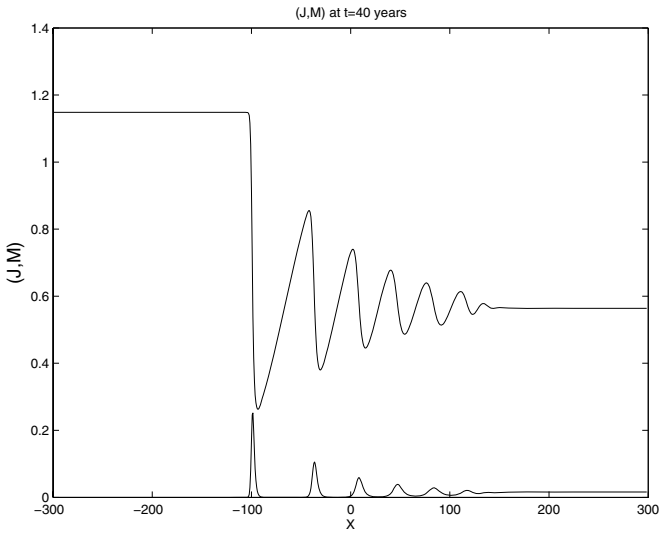


FIG. 2. *Graph of solutions M (upper) and J (lower) at $t = 40$ years, here $\tau = 0.8$, $d_I = 10$, $d_S = 0.5$, $d_Y = 0.8$, $S_0 = 2$, $b_0 = 2$, and $\beta = 10$. There exist long wave tails for both J and M.*

and $\tau = 0.8$ shows that the solution stabilizes to a traveling wavefront with minimal speed 43.549 km/year. The numerical result when $t = 40$ years is shown in Figure 2.

Fixing other parameters, we carry out simulations in the cases when $D_Y = 0.25D_I$ and $D_Y = D_I$. It is found that in both cases, the spreading speeds stabilize to the same minimal wave speed 43.549 km/year and the change of the diffusion rate D_Y has impact only on the amplitudes and frequencies of oscillation for the long tail in the traveling wave, and its impact on the shape of the solution is less apparent if we confine the value D_Y/D_I to the interval $[0, 1]$. This result is what we should expect because the maturation time τ is relatively small that the contribution of D_Y to the pattern of solutions is limited. See Figure 3 for the comparison of M up to $t = 40$ years between the case $D_Y/D_I = 0$ and the case $D_Y/D_I = 1$.

Our simulations agree with the theoretical analysis in the above sections that the minimal wave speed of rabies depends on the maturation time τ, while the amplitude and frequencies of oscillations of the long tail are influenced also by the diffusion rate of juvenile foxes.

We conclude with a remark about the limitation of this work. We assumed two age classes and homogeneity within each age class. Namely, many parameters in the model such as death and diffusion rates and force of infection are all assumed to be constants that depend on the age class but are independent of the precise age. This is certainly only an approximation to the biological reality, and parameter values should be thought of as some sort of averages during the whole juvenile or adult period. For example, newborn susceptible juveniles would not be moving at all and the search for new territories by juveniles must happen only during a particular phase of childhood. In [16], it was noted that breeding season varies from region to region but usually begins early in the year, then in the autumn following birth the pups of the litter will disperse to their own territories. Ideally, we should use age-dependent coefficients and parameters, and hence the model would become an age-structured reaction diffusion equation that cannot be reduced to a system of reaction diffusion equations with delayed nonlocal nonlinearities.

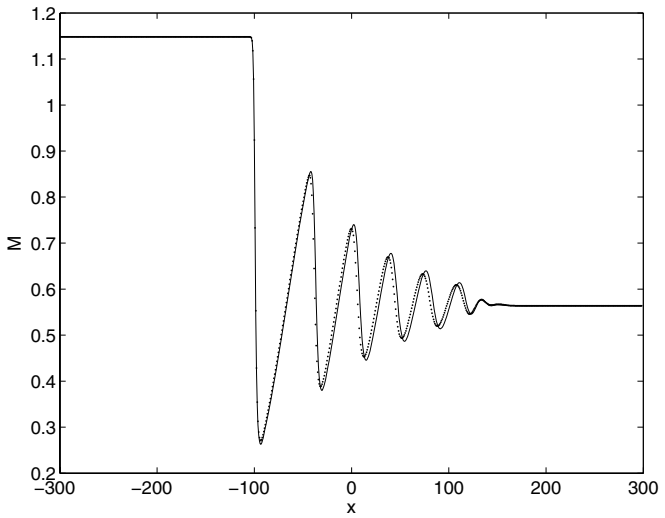


FIG. 3. *Graph of solutions M for $D_Y/D_I = 0$ (solid line) and $D_Y/D_I = 1$ (dashed line) where $t = 40$ years.*

REFERENCES

[1] L. J. S. Allen, D. A. Flores, R. K. Ratnayake, and J. R. Herbold, *Discrete-time deterministic and stochastic models for the spread of rabies*, Appl. Math. Comput., 132 (2002), pp. 271–292.

[2] R. M. Anderson, H. C. Jackson, R. M. May, and A. M. Smith, *Population dynamics of fox rabies in Europe*, Nature, 289 (1981), pp. 765–771.

[3] N. F. Britton, *Spatial structures and periodic travelling waves in an integro-differential reaction-diffusion population model*, SIAM J. Appl. Math., 50 (1990), pp. 1663–1688.

[4] X. Chen, *Existence, uniqueness, and asymptotic stability of traveling waves in nonlocal evolution equations*, Adv. Differential Equations, 2 (1997), pp. 125–160.

[5] S.-N. Chow, X.-B. Lin, and J. Mallet-Paret, *Transition layers for singularly perturbed delay differential equations with monotone nonlinearities*, J. Dynam. Differential Equations, 1 (1989), pp. 3–43.

[6] P. Daszak, A. A. Cunningham, and A. D. Hyatt, *Wildlife ecology—Emerging infectious diseases of wildlife—Threats to biodiversity and human health*, Science, 287 (2000), pp. 443–449.

[7] S. R. Dunbar, *Travelling wave solutions of diffusive Lotka-Volterra equations*, J. Math. Biol., 17 (1983), pp. 11–32.

[8] T. Faria, W. Huang, and J. Wu, *Traveling waves for delayed reaction-diffusion equations with non-local response*, Proc. Roy. Soc. London Ser. A, 462 (2006), pp. 229–261.

[9] S. A. Gourley and N. F. Britton, *A predator prey reaction diffusion system with nonlocal effects*, J. Math. Biol., 34 (1996), pp. 297–333.

[10] S. A. Gourley, J. W.-H. So, and J. Wu, *Non-locality of reaction-diffusion equations induced by delay: Biological modeling and nonlinear dynamics*, J. Math. Sci. (N.Y.), 124 (2004), pp. 5119–5153.

[11] S. A. Gourley and J. Wu, *Delayed reaction diffusion equations: Theory and applications to biological invasion and disease spread*, in Nonlinear Dynamics and Evolution Equations, H. Brunner, X. Zhao, and X. Zou, eds., AMS, Providence, RI, 2006, pp. 137–200.

[12] W. S. C. Gurney, S. P. Blythe, and R. M. Nisbet, *Nicholson's blowflies revisited*, Nature, 287 (1980), pp. 17–21.

[13] J. K. Hale and X.-B. Lin, *Heteroclinic orbits for retarded functional-differential equations*, J. Differential Equations, 65 (1986), pp. 175–202.

[14] A. Kallen, *Thresholds and travelling waves in an epidemic model for rabies*, Nonlinear Anal., 8 (1984), pp. 851–856.

[15] A. Kallen, P. Arcuri, and J. D. Murray, *A simple model for the spatial spread and control of rabies*, J. Theoret. Biol., 116 (1985), pp. 377–393.

[16] H. G. Lloyd, *The Red Fox*, B. T. Batsford Ltd., London, 1980.
[17] D. W. MacDonald, *Rabies and Wildlife. A Biologist's Perspective*, Oxford University Press, Oxford, UK, 1980.
[18] L. Markus, *Asymptotically autonomous differential systems*, in Contributions to the Theory of Nonlinear Oscillations, Vol. 3, Ann. of Math. Stud. 36, Princeton University Press, Princeton, NJ, 1956, pp. 17–29.
[19] R. H. Martin, Jr., and H. L. Smith, *Abstract functional-differential equations and reaction-diffusion systems*, Trans. Amer. Math. Soc., 321 (1990), pp. 1–44.
[20] J. A. J. Metz and O. Diekmann, eds., *The Dynamics of Physiologically Structured Populations*, papers from the colloquium held in Amsterdam, 1983, Lecture Notes in Biomath. 68, Springer-Verlag, Berlin, 1986.
[21] K. Mischaikow, H. Smith, and H. R. Thieme, *Asymptotically autonomous semiflows: Chain recurrence and Lyapunov functions*, Trans. Amer. Math. Soc., 347 (1995), pp. 1669–1685.
[22] J. D. Murray, E. A. Stanley, and D. L. Brown, *On the spatial spread of rabies among foxes*, Proc. Roy. Soc. London Ser. B, 229 (1986), pp. 111–150.
[23] R. E. O'Malley, Jr., *Singular Perturbation Methods for Ordinary Differential Equations*, Appl. Math. Sci. 89, Springer-Verlag, New York, 1991.
[24] M. A. Pozio, *Behaviour of solutions of some abstract functional differential equations and application to predator-prey dynamics*, Nonlinear Anal., 4 (1980), pp. 917–938.
[25] M. A. Pozio, *Some conditions for global asymptotic stability of equilibria of integro-differential equations*, J. Math. Anal. Appl., 95 (1983), pp. 501–527.
[26] R. Redlinger, *Existence theorems for semilinear parabolic systems with functionals*, Nonlinear Anal., 8 (1984), pp. 667–682.
[27] R. Redlinger, *On Volterra's population equation with diffusion*, SIAM J. Math. Anal., 16 (1985), pp. 135–142.
[28] H. L. Smith and H. R. Thieme, *Strongly order preserving semiflows generated by functional-differential equations*, J. Differential Equations, 93 (1991), pp. 332–363.
[29] J. W.-H. So, J. Wu, and X. Zou, *A reaction diffusion model for a single species with age structure,* I. *Travelling wave fronts on unbounded domains*, Proc. Roy. Soc. London Ser. A, 457 (2001), pp. 1841–1853.
[30] A. Wandeler, G. Wachendorfer, U. Forster, H. Krekel, W. Schale, J. Muller, and F. Steck, *Rabies in wild carnivores in central Europe.* 1. *Epidemiological studies*, Zbl. Veter. Med. B, 21 (1974), pp. 735–756.
[31] A. Wandeler, G. Muller, G. Wachendorfer, W. Schale, U. Forster, and F. Steck, *Rabies in wild carnivores in central Europe.* 3. *Ecology and biology of the fox in relation to control operations*, Zbl. Veter. Med. B, 21 (1974), pp. 765–773.
[32] G. F. Webb, *Theory of Nonlinear Age-Dependent Population Dynamics*, Monogr. Textbooks Pure Appl. Math. 89, Marcel Dekker, New York, 1985.
[33] J. Wu, *Theory and Applications of Partial Functional-Differential Equations*, Appl. Math. Sci. 119, Springer-Verlag, New York, 1996.
[34] Y. Yamada, *Asymptotic stability for some systems of semilinear Volterra diffusion equations*, J. Differential Equations, 52 (1984), pp. 295–326.

SIAM J. APPL. MATH.
Vol. 67, No. 1, pp. 164–193

MODELING VISCOELASTIC BEHAVIOR OF ARTERIAL WALLS AND THEIR INTERACTION WITH PULSATILE BLOOD FLOW*

SUNČICA ČANIĆ†, JOSIP TAMBAČA‡, GIOVANNA GUIDOBONI†, ANDRO MIKELIĆ§, CRAIG J. HARTLEY¶, AND DOREEN ROSENSTRAUCH‖

Abstract. Fluid-structure interaction describing wave propagation in arteries driven by the pulsatile blood flow is a complex problem. Whenever possible, simplified models are called for. One-dimensional models are typically used in arterial sections that can be approximated by the cylindrical geometry allowing axially symmetric flows. Although a good first approximation to the underlying problem, the one-dimensional model suffers from several drawbacks: the model is not closed (an ad hoc velocity profile needs to be prescribed to obtain a closed system) and the model equations are quasi-linear hyperbolic (oversimplifying the viscous fluid dissipation), typically producing shock wave solutions not observed in healthy humans. In this manuscript we derived a simple, *closed* reduced model that accounts for the viscous fluid dissipation to the leading order. The resulting fluid-structure interaction system is of hyperbolic-parabolic type. Arterial walls were modeled by a novel, linearly *viscoelastic* cylindrical Koiter shell model and the flow of blood by the incompressible, viscous Navier–Stokes equations. Kelvin–Voigt-type viscoelasticity was used to capture the hysteresis behavior observed in the measurements of the arterial stress-strain response. Using the a priori estimates obtained from an energy inequality, together with the asymptotic analysis and ideas from homogenization theory for porous media flows, we derived an effective model which is an ϵ^2-approximation to the three-dimensional axially symmetric problem, where ϵ is the aspect ratio of the cylindrical arterial section. Our model shows two interesting features of the underlying problem: bending rigidity, often times neglected in the arterial wall models, plays a nonnegligible role in the ϵ^2-approximation of the original problem, and the viscous fluid dissipation imparts long-term viscoelastic memory effects on the motion of the arterial walls. This does not, to the leading order, influence the hysteresis behavior of arterial walls. The resulting model, although two-dimensional, is in the form that allows the use of one-dimensional finite element method techniques producing fast numerical solutions. We devised a version of the Douglas–Rachford time-splitting algorithm to solve the underlying hyperbolic-parabolic problem. The results of the numerical simulations were compared with the experimental flow measurements performed at the Texas Heart Institute, and with the data corresponding to the hysteresis of the human femoral artery and the canine abdominal aorta. Excellent agreement was observed.

Key words. blood flow, viscoelastic arteries, fluid-structure interaction, effective equations

AMS subject classifications. 35Q30, 74K15, 76D27

DOI. 10.1137/060651562

*Received by the editors February 5, 2006; accepted for publication (in revised form) July 31, 2006; published electronically November 16, 2006.

http://www.siam.org/journals/siap/67-1/65156.html

†Department of Mathematics, University of Houston, 4800 Calhoun Rd., Houston, TX 77204-3476 (canic@math.uh.edu, gio@math.uh.edu). The first author's research was supported by the NSF under grants DMS0245513 and DMS-0337355, and by the NSF and NIH under grant DMS-0443826.

‡Department of Mathematics, University of Zagreb, Bijenička 30, 10000 Zagreb, Croatia (tambaca@math.hr). This author's research was supported by the NSF and NIH under grant DMS-0443826.

§Institue Camille Jordan, UFR Mathématiques, Site de Gerland, Université Claude Bernard Lyon 1, Bat. A, 50 avenue Tony Garnier, 69367 Lyon Cedex 07, France (mikelic@univ-lyon1.fr). This author's research was supported by the NSF and NIH under grant DMS-0443826.

¶Department of Medicine, Section of Cardiovascular Sciences, Baylor College of Medicine, Houston, TX 77030 (chartley@bcm.edu). This author's research was supported by the NSF and NIH under grant DMS-0443826, and by the NIH under grant HL22512.

‖Texas Heart Institute at St. Luke's Episcopal Hospital, Houston, TX 77030, and the University of Texas Health Science Center at Houston, Houston, TX 77030 (doreen.rosenstrauch@uth.tmc.edu). This author's research was supported by the NSF and NIH under grant DMS-0443826, and by the Roderick Duncan McDonald Foundation at St. Luke's Episcopal Hospital.

1. Introduction. The study of flow of a viscous incompressible fluid through a compliant tube is of interest to many applications. A major application is blood flow through human arteries. Understanding wave propagation in arterial walls, local hemodynamics, and temporal wall shear stress gradient is important in understanding the mechanisms leading to various complications in the cardiovascular function. Many clinical treatments can be studied in detail only if a reliable model describing the response of arterial walls to the pulsatile blood flow is considered.

It has been well accepted that in medium-to-large arteries blood can be modeled as a viscous, incompressible Newtonian fluid. Although blood is a suspension of red blood cells, white blood cells, and platelets in plasma, its non-Newtonian nature due to the particular rheology is relevant in small arteries (arterioles) and capillaries where the diameter of the arteries becomes comparable to the size of the cells. In medium-to-large arteries, such as the coronary arteries (medium) and the abdominal aorta (large), the Navier–Stokes equations for an incompressible viscous fluid are considered to be a good model for blood flow.

Devising an accurate model for the mechanical behavior of arterial walls is more complicated. Arterial walls are anisotropic and heterogeneous, composed of layers with different biomechanical characteristics [21, 22, 29, 44]. A variety of different models has been suggested in the literature to model the mechanical behavior of arteries [1, 2, 3, 21, 22, 23, 29, 27, 33, 44, 51]. They range from the detailed description of each of the layers to the average description of the total mechanical response of the vessel wall assuming homogeneous, linearly elastic behavior.

To study the *coupling* between the motion of the vessel wall and pulsatile blood flow, a detailed description of the vessel wall biomechanical properties may lead to a mathematical and numerical problem whose complexity is beyond today's computational capabilities. The nonlinearity of the underlying fluid-structure interaction is so severe that even simplified description of the vessel wall mechanics assuming homogeneous, linearly elastic behavior leads to the complicated numerical algorithms with challenging stability and convergence properties. To devise a mathematical model that will lead to a problem which is amenable to numerical methods producing computational solutions in a reasonable time-frame, various simplifications need to be introduced. They can be based on the simplifying model *assumptions* capturing only the most important physics of the problem and/or on the simplifications utilizing *special problem features* such as, for example, special geometry, symmetry, and periodicity.

A common set of simplifying assumptions that captures only the most important physics in the description of the mechanical properties of arterial walls includes homogeneity of the material with "small" displacements and "small" deformation gradients leading to the hypothesis of linear elasticity. A common set of special problem features that leads to simplifying models includes "small" vessel wall thickness allowing a reduction from three-dimensional models to two-dimensional shell models, and cylindrical geometry of a section of an artery where no branching is present allowing the use of cylindrical shell models. Neglecting bending rigidity of arteries, studied in [18, 21], reduces the shell model to a membrane model. Further simplifications include axial symmetry of the loading exerted by the blood flow to the vessel walls in the approximately straight cylindrical sections, leading to axially symmetric models with a potential of further reduction to one-dimensional models. One-dimensional models, although a good first approximation to the underlying problem, suffer from several drawbacks: they are not closed (an ad hoc velocity profile needs to be prescribed to

obtain a closed system), and the model equations are quasi-linear hyperbolic, typically producing shock wave solutions, not observed in healthy humans [5]. In particular, the wall shear stress calculated using one-dimensional models is a consequence of the form of the prescribed velocity profile.

Two-dimensional and three-dimensional models of the fluid-structure interaction between the incompressible viscous fluid flow and the motion of a linearly elastic cylindrical membrane are rather complex. Often times additional ad hoc terms of viscoelastic nature are added to the vessel wall model to provide stability and convergence of the underlying numerical algorithm [40, 44], or to provide enough regularity in the proof of the existence of a solution [10, 16, 24, 49], thereby showing well-posedness of the underlying problem. To this day there is no analytical result proving well-posedness of the fluid-structure interaction problem without assuming that the structure model includes the higher-order derivative terms capturing some kind of viscoelastic behavior [10, 16, 24, 49], or with the terms describing bending (flexion) rigidity in elastic shells or plates [10, 15]. In fact, current literature on well-posedness of the fluid-structure interaction between a viscous incompressible Newtonian fluid and a viscoelastic structure includes many additional simplifying assumptions such as the smallness of the data [49], periodic boundary conditions [24, 49], or flow in a closed cavity [10, 15, 16], not appropriate for the blood-flow application. Thus, the well-posedness of the fluid-structure interaction problem describing blood flow in compliant (elastic or viscoelastic) arteries remains an open problem. However, even in those simplifying problems when the data is infinitesimally small the higher-order regularizing terms in the structure model play a crucial role in providing the stabilizing mechanism. Thus, ignoring the terms that account for bending rigidity of the vessel walls and/or viscous dissipation might mean oversimplifying the physics, giving rise to a problem which might not have a solution.

Keeping this in mind we turn to the theory of elastic/viscoelastic shells to model the mechanical properties of arterial walls. Thus, we will be assuming that the vessel walls are homogeneous, that the thickness of the wall is small in comparison to the vessel radius, and that the state of stress is approximately plane, allowing us to consider shell theory. See section 2. The equations of shell theory have been derived by many authors; see [19] and the references therein. Due to variations in approach and rigor the variety of equations occurring in the literature is overwhelming. Among all the equations of shell theory the Koiter shell equations appear to be the simplest consistent first approximation in the general theory of thin elastic shells [32, 31]. In addition, they have been mathematically justified using asymptotic methods to be consistent with three-dimensional elasticity [12, 13]. Ciarlet and Lods showed in [12] that the Koiter shell model has the same asymptotic behavior as the three-dimensional membrane model, the bending model, and the generalized membrane model in the respective regimes in which each of them holds. Motivated by these remarkable properties of the Koiter shell model, in this manuscript we derived the Koiter shell equations for the cylindrical geometry and extended the linearly elastic Koiter model to include the viscous effects observed in the measurements of the mechanical properties of vessel walls [1, 2, 3]. We utilized the Kelvin–Voigt viscoelastic model, which has been shown in [1, 2, 3] to approximate well the experimentally measured viscoelastic properties of the canine aorta and of the human femoral and carotid arteries. In [43] a version of the Kelvin–Voigt model was used to model the vessel walls as a linearly viscoelastic membrane. In the Kelvin–Voigt model the total stress is linearly proportional to the strain and the time-derivative of strain. More

precisely, for a three-dimensional isotropic and homogeneous body, the Kelvin–Voigt model relates the total stress tensor, whose components we denote by t_{kl}, to the infinitesimal strains e_{kl} and the time-derivative of the strains $\partial_t e_{kl}$ through the following relationship [20]:

$$t_{kl} = (\lambda_e + \lambda_v \partial_t) I_e \delta_{kl} + 2(\mu_e + \mu_v \partial_t) e_{kl}, \quad k, l = 1, 2, 3, \tag{1.1}$$

where λ_e and μ_e are the Lamé constants of elasticity, λ_v and μ_v are their corresponding viscoelastic counterparts, δ_{kl} is the Kronecker delta, and $I_e := \sum_{i=1}^{3} e_{ii}$. In section 8 we show that the fluid-structure interaction algorithm based on the viscoelastic Koiter shell equations coupled with the Navier–Stokes equations for a viscous incompressible fluid captures the experimentally measured viscoelastic properties of arterial walls in the human femoral artery and in the canine aorta. This is, in a nutshell, the main result of this manuscript; using the a priori estimates based on an energy inequality, coupled with the asymptotic analysis and homogenization theory, we derived an effective, closed fluid-structure interaction model and a fast numerical solver whose solutions capture the viscoelastic properties of major arteries. We show that our effective model approximates the original three-dimensional axially symmetric problem to the ϵ^2-accuracy, where ϵ is the aspect ratio of the cylindrical domain (vessel). Our reduced, effective model reveals several interesting features of the coupled fluid-structure interaction problem:

(1) Our model explicitly shows how the leading-order viscous fluid dissipation imparts long-term viscoelastic memory effects on the motion of the vessel wall. This is studied in section 5; see (5.11). We show that this does not influence, to the leading order, the viscoelastic hysteresis loop observed in the stress-strain (or the pressure-diameter) measurements of the arterial viscoelastic properties.

(2) Our model shows that bending rigidity of vessel walls plays a nonnegligible role in the asymptotic behavior of the underlying fluid-structure interaction problem. See the equation for p^0 in (4.17). We found that for the parameters describing blood flow through medium-to-large arteries the leading-order terms in the coupling of the stresses at the vessel wall include not only the membrane terms but also a correction accounting for the bending rigidity of the wall, often times neglected in the description of the mechanical properties of vessel walls.

We developed a fast numerical solver based on the one-dimensional finite element approach and compared the computational solution with the experimental measurements. First, the reduced *elastic* model was tested experimentally using a mock circulatory flow loop with latex tubing, assembled at the Research Laboratory at the Texas Heart Institute. Then the *viscoelastic* model was compared to the hysteresis measurements of the viscoelastic properties of the human femoral artery and the canine aorta. In both cases, excellent agreement between the experiment and the numerical solution was obtained.

2. The viscoelastic cylindrical Koiter shell model. In this section we focus on the derivation of the viscoelastic cylindrical Koiter shell model. We begin with the linearly elastic Koiter shell model as it was derived in [31, 32] and specialize it to the cylindrical shell geometry. Following standard texts in conventional plate and shell theories (see, for example, [20, 41, 45, 50, 52]), we then derive the stress-strain relationship for the Koiter shell model and extend it to include the Kelvin–Voigt viscoelasticity, which has been experimentally observed to approximate well the viscoelastic mechanical properties of arterial walls [1, 2, 3]. We summarize the main steps next.

2.1. The linearly elastic Koiter shell model. Consider a clamped cylindrical shell with the reference radius of the middle surface equal to $r = R$, with the shell thickness h and the cylinder length L, $z \in (0, L)$. The basic assumptions under which the Koiter shell model holds are [31, 32] that

- the shell is thin ($h/R \ll 1$);
- the strains are small everywhere, although large deflections are admitted, and the strain energy per unit volume of the undeformed body is represented by the quadratic function of the strain components for an isotropic solid (Hooke's law);
- the state of stress is approximately plane.

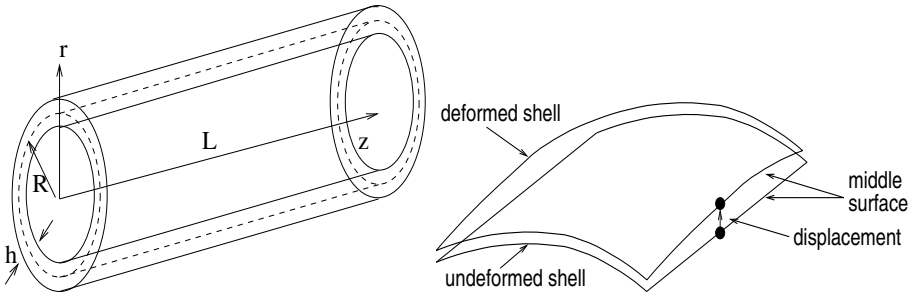


FIG. 2.1. *Left: Cylindrical shell (reference configuration) with middle surface radius R and shell thickness h. Right: Deformed shell.*

The weak formulation, describing the variation of the strain energy density function, depends on the change of metric and the change of curvature tensors of the surface. The change of metric tensor captures the stretching of the surface and the change of curvature tensor captures the bending effects. The weak formulation of the Koiter shell describes variation of the energy that is due to stretching and bending of the shell.

Denote by $\boldsymbol{\xi}(z) = (\xi_z(z), \xi_r(z))$ the displacement of the middle surface at z (see Figure 2.1), where $\xi_z(z)$ and $\xi_r(z)$ denote the longitudinal and the radial component of the displacement, respectively. Here the axial symmetry of the problem has already been taken into account assuming that the displacement in the θ-direction is zero, and that nothing in the problem depends on θ. The change of metric and the change of curvature tensors for a cylindrical shell are given, respectively, by [11]

$$\boldsymbol{\gamma}(\boldsymbol{\xi}) = \begin{bmatrix} \xi_z' & 0 \\ 0 & R\xi_r \end{bmatrix}, \quad \boldsymbol{\varrho}(\boldsymbol{\xi}) = \begin{bmatrix} -\xi_r'' & 0 \\ 0 & \xi_r \end{bmatrix}.$$

Here $'$ denotes the derivative with respect to the longitudinal variable z. Introduce the following function space:

$$\begin{aligned} V_c &= H_0^1(0,L) \times H_0^2(0,L) \\ &= \big\{(\xi_z, \xi_r) \in H^1(0,L) \times H^2(0,L) : \xi_z(0) = \xi_z(L) = \xi_r(0) = \xi_r(L) = 0, \\ &\qquad \xi_r'(0) = \xi_r'(L) = 0\big\}. \end{aligned}$$

Then the weak formulation of the linearly elastic cylindrical Koiter shell is given by the following: find $\boldsymbol{\eta} = (\eta_z, \eta_r) \in V_c$ such that

$$\frac{h}{2}\int_0^L \mathcal{A}\boldsymbol{\gamma}(\boldsymbol{\eta}) \cdot \boldsymbol{\gamma}(\boldsymbol{\xi}) R dz + \frac{h^3}{24}\int_0^L \mathcal{A}\boldsymbol{\varrho}(\boldsymbol{\eta}) \cdot \boldsymbol{\varrho}(\boldsymbol{\xi}) R dz = \int_0^L \boldsymbol{f} \cdot \boldsymbol{\xi} R dz, \quad \boldsymbol{\xi} \in V_c, \tag{2.1}$$

where $\cdot$ denotes the scalar product

$$A \cdot B := \mathrm{Tr}\left(AB^T\right), \quad A, B \in M_2(\mathbb{R}) \cong \mathbb{R}^4. \tag{2.2}$$

Here $\boldsymbol{f}$ is the surface density of the force applied to the shell, and $\mathcal{A}$ is the elasticity tensor given by [11]

$$\mathcal{A}\mathbf{E} = \frac{4\lambda\mu}{\lambda+2\mu}(\mathbf{A}^c \cdot \mathbf{E})\mathbf{A}^c + 4\mu\mathbf{A}^c\mathbf{E}\mathbf{A}^c, \quad \mathbf{E} \in \mathrm{Sym}\,(\mathbb{R}^2), \quad \text{with}$$
$$\mathbf{A}_c = \begin{bmatrix} 1 & 0 \\ 0 & R^2 \end{bmatrix}, \quad \mathbf{A}^c = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{R^2} \end{bmatrix},$$

where λ and μ are the Lamé constants. Written in terms of the displacement, the weak formulation reads

$$\begin{aligned}
&\frac{h}{2}\int_0^L \left(\frac{4\mu\lambda}{\lambda+2\mu}\left(\eta_z' + \frac{1}{R}\eta_r\right)\cdot\left(\xi_z' + \frac{1}{R}\xi_r\right) + 4\mu\left(\eta_z'\xi_z' + \frac{1}{R^2}\eta_r\xi_r\right)\right)dz \\
&+ \frac{h^3}{24}\int_0^L \left(\frac{4\mu\lambda}{\lambda+2\mu}\left(-\eta_r'' + \frac{1}{R^2}\eta_r\right)\cdot\left(-\xi_r'' + \frac{1}{R^2}\xi_r\right) + 4\mu\left(\eta_r''\xi_r'' + \frac{1}{R^4}\eta_r\xi_r\right)\right)dz \\
&= \int_0^L (f_z\xi_z + f_r\xi_r)dz \quad \forall(\xi_z, \xi_r) \in V_c.
\end{aligned}$$

Using the following relationships between the Lamé constants and Young's modulus of elasticity E and the Poisson ratio σ

$$\frac{2\mu\lambda}{\lambda+2\mu} + 2\mu = 4\mu\frac{\lambda+\mu}{\lambda+2\mu} = \frac{E}{1-\sigma^2}, \quad \frac{2\mu\lambda}{\lambda+2\mu} = 4\mu\frac{\lambda+\mu}{\lambda+2\mu}\frac{1}{2}\frac{\lambda}{\lambda+\mu} = \frac{E}{1-\sigma^2}\sigma,$$

the elasticity tensor $\mathcal{A}$ reads

$$\mathcal{A}\mathbf{E} = \frac{2E\sigma}{1-\sigma^2}(\mathbf{A}^c \cdot \mathbf{E})\mathbf{A}^c + \frac{2E}{1+\sigma}\mathbf{A}^c\mathbf{E}\mathbf{A}^c, \quad \mathbf{E} \in \mathrm{Sym}\,(\mathbb{R}^2).$$

From here we get the weak formulation (2.1) as

$$\begin{aligned}
&h\int_0^L \left(\frac{E\sigma}{1-\sigma^2}\left(\eta_z' + \frac{1}{R}\eta_r\right)\left(\xi_z' + \frac{1}{R}\xi_r\right) + \frac{E}{1+\sigma}\left(\eta_z'\xi_z' + \frac{1}{R^2}\eta_r\xi_r\right)\right)dz \\
&+ \frac{h^3}{12}\int_0^L \left(\frac{E\sigma}{1-\sigma^2}\left(-\eta_r'' + \frac{1}{R^2}\eta_r\right)\left(-\xi_r'' + \frac{1}{R^2}\xi_r\right) + \frac{E}{1+\sigma}\left(\eta_r''\xi_r'' + \frac{1}{R^4}\eta_r\xi_r\right)\right)dz \\
&= \int_0^L (f_z\xi_z + f_r\xi_r)dz, \quad (\xi_z, \xi_r) \in V_c.
\end{aligned} \tag{2.3}$$

The terms multiplying $h/2$ account for the stored energy density due to stretching (membrane effects) and the terms multiplying $h^3/12$ account for the stored energy density due to bending (flexural shell effects). Integration by parts gives rise to the static equilibrium equations. Written in differential form they read

$$\boxed{\begin{aligned}
-\frac{hE}{1-\sigma^2}\left(\eta_z'' + \sigma\frac{1}{R}\eta_r'\right) &= f_z, \\
\frac{hE}{R(1-\sigma^2)}\left(\sigma\eta_z' + \frac{\eta_r}{R}\right) + \frac{h^3E}{12(1-\sigma^2)}\left(\eta_r'''' - 2\sigma\frac{1}{R^2}\eta_r'' + \frac{1}{R^4}\eta_r\right) &= f_r.
\end{aligned}} \tag{2.4}$$

The Linearly Elastic Cylindrical Koiter Shell Model

We employ these equations to study the response of arteries to pulsatile blood flow. For this purpose, we assume that the in vivo arteries are *prestretched* under internal pressure load, that the arterial walls are *longitudinally tethered*, and that the *longitudinal displacement is negligible* [38, 42].

The assumption that the longitudinal displacement is negligible has been justified in [38]. More precisely, in [38] we considered the equations of *three-dimensional* linear elasticity to model the vessel wall, coupled with the Navier–Stokes equations for a viscous, incompressible fluid to model the flow of blood in cylindrical geometry. In addition, we assumed that the "thickness" h of the structure (the radial dimension of the three-dimensional elastic body) is less than or comparable to the radius of the domain occupied by the fluid, i.e., $h/R \leq 1$ (this includes the scenario $h/R \ll 1$ considered in this manuscript). Starting from the assumption that both the radial and longitudinal displacement of the three-dimensional structure are nonzero, we showed that the effective model obtained by considering small aspect ratio $\epsilon = R/L$ embodies negligible longitudinal displacement of the structure.

Taking this into account we emply here the equations of a linearly elastic cylindrical Koiter shell model with negligible longitudinal displacement:

$$\boxed{\left(\frac{hE}{R(1-\sigma^2)} + p_{\text{ref}}\right)\frac{\eta_r}{R} + \frac{h^3E}{12(1-\sigma^2)}\left(\eta_r'''' - 2\sigma\frac{1}{R^2}\eta_r'' + \frac{1}{R^4}\eta_r\right) = f_r.} \tag{2.5}$$

This is obtained from the weak formulation (2.3), assuming $\eta_z = 0$, and the test space

$$V_c^0 := V_c \cap \{\xi_z = 0\}.$$

In order to include the fact that the reference configuration is prestressed at reference pressure p_{ref}, and that the arterial walls are viscoelastic, we study the stress-strain relationship corresponding to the Koiter shell model and modify it to include these two effects. This is presented next.

2.2. The linearly viscoelastic Koiter shell model. The stress-strain relationship is given by the "stress resultant," which relates the internal force with the change of metric tensor, and the "stress couples," which describe the bending moments in terms of the change of curvature tensor [20]. As noted by Koiter in his original paper [31], the stress resultant and the stress couples can be obtained from (2.1) as gradients of the stored energy function, given by the integrand on the left-hand side of (2.1), with respect to the middle surface strains and changes of curvature. Following this approach one obtains

- *stress resultant (or the internal force) for the elastic Koiter shell*

$$N := \frac{h}{2}\mathcal{A}\boldsymbol{\gamma}(\boldsymbol{\eta}) = \frac{h}{2}\begin{bmatrix} \frac{2E\sigma}{1-\sigma^2}\frac{\eta_r}{R} & 0 \\ 0 & \frac{2E}{1-\sigma^2}\frac{\eta_r}{R^3} \end{bmatrix}, \tag{2.6}$$

- *stress couples (bending moment) for the elastic Koiter shell*

$$M := \frac{h^3}{24}\mathcal{A}\boldsymbol{\varrho}(\boldsymbol{\eta}) = \frac{h^3}{24}\begin{bmatrix} -\frac{2E}{1-\sigma^2}\eta_r'' + \frac{2E\sigma}{1-\sigma^2}\frac{\eta_r}{R^2} & 0 \\ 0 & \frac{2E}{1-\sigma^2}\frac{u_r}{R^4} - \frac{2E\sigma}{1-\sigma^2}\frac{1}{R^2}\eta_r'' \end{bmatrix}. \tag{2.7}$$

At this point we also introduce the effects of prestress by defining the stress resultant N_{ref} that relates the reference pressure p_{ref} with the circumferential strain [17, 34, 35]

$$\frac{h}{2}N_{\text{ref}} = hR\mathbf{A}^c\begin{bmatrix} 0 & 0 \\ 0 & p_{ref}\frac{R}{h}\eta_r \end{bmatrix}\mathbf{A}^c \tag{2.8}$$

so that the total stress resultant, including the effects of prestress, reads

- *stress resultant for the prestressed elastic Koiter shell*

$$N = \frac{h}{2}\mathcal{A}\boldsymbol{\gamma}(\boldsymbol{\eta}) + \frac{h}{2}N_{\text{ref}}. \tag{2.9}$$

We focus now on introducing the viscous effects to the linearly elastic, prestressed cylindrical Koiter shell model. For this purpose assume that the displacement is not only a function of position z but also a function of time: $\boldsymbol{\eta} = \boldsymbol{\eta}(z,t)$ and that the velocity of the displacement is linearly proportional to the stress as described in (1.1). Employing the Kelvin–Voigt model (1.1) to describe this viscoelastic behavior one writes the constitutive relations in which the stress is linearly proportional to the strain plus the time-derivative of strain [20]. For the linearly viscoelastic Koiter shell model we define

- *stress resultant for the viscoelastic prestressed Koiter shell*

$$N := \frac{h}{2}\mathcal{A}\boldsymbol{\gamma}(\boldsymbol{\eta}) + \frac{h}{2}\mathcal{B}\boldsymbol{\gamma}(\dot{\boldsymbol{\eta}}) + \frac{h}{2}N_{\text{ref}}, \tag{2.10}$$

- *stress couples for the viscoelastic Koiter shell*

$$M := \frac{h^3}{24}\mathcal{A}\boldsymbol{\varrho}(\boldsymbol{\eta}) + \frac{h^3}{24}\mathcal{B}\boldsymbol{\varrho}(\dot{\boldsymbol{\eta}}), \tag{2.11}$$

where $\mathcal{B}$ is given by

$$\mathcal{B}\mathbf{E} = \frac{4\lambda_v\mu_v}{\lambda_v + 2\mu_v}(\mathbf{A}^c \cdot \mathbf{E})\mathbf{A}^c + 4\mu_v\mathbf{A}^c\mathbf{E}\mathbf{A}^c, \quad \mathbf{E} \in \text{Sym}\,(\mathbb{R}^2),$$

with μ_v and λ_v corresponding to the viscous counterpart of the Lamé constants μ and λ. With these constitutive relations we now define the weak formulation of the linearly viscoelastic prestressed Koiter shell model by the following: for each $t > 0$ find $\eta(t) \in V_c$ such that $\forall \boldsymbol{\xi}(t) \in V_c$

$$\begin{aligned}\frac{h}{2}\int_0^L (N_{\text{ref}} + \mathcal{A}\boldsymbol{\gamma}(\boldsymbol{\eta}) + \mathcal{B}\boldsymbol{\gamma}(\dot{\boldsymbol{\eta}})) \cdot \boldsymbol{\gamma}(\boldsymbol{\xi})Rdz + \frac{h^3}{24}\int_0^L (\mathcal{A}\boldsymbol{\varrho}(\boldsymbol{\eta}) + \mathcal{B}\boldsymbol{\varrho}(\dot{\boldsymbol{\eta}})) \cdot \boldsymbol{\varrho}(\boldsymbol{\xi})Rdz \\ + \rho_w h\int_0^L \frac{\partial^2\boldsymbol{\eta}}{\partial t^2} \cdot \boldsymbol{\xi} = \int_0^L \boldsymbol{f} \cdot \boldsymbol{\xi}Rdz,\end{aligned} \tag{2.12}$$

where $\dot{\boldsymbol{\eta}}$ denotes the time-derivative. Written in terms of the displacement, after employing the notation

$$C_v := \frac{2\lambda_v\mu_v}{\lambda_v + 2\mu_v} + 2\mu_v, \quad D_v := \frac{2\lambda_v\mu_v}{\lambda_v + 2\mu_v}, \tag{2.13}$$

the weak formulation of the linearly viscoelastic prestressed Koiter shell model reads

$$\begin{aligned}&\int_0^L f_r\xi_r dz = \rho_w h\int_0^L \frac{\partial^2\eta_r}{\partial t^2}\xi_r + h\int_0^L \left(\left(\frac{E}{1-\sigma^2} + p_{\text{ref}}\frac{R}{h}\right)\frac{1}{R}\eta_r + C_v\frac{1}{R}\frac{\partial\eta_r}{\partial t}\right)\frac{\xi_r}{R}dz \\ &+ \frac{h^3}{12}\int_0^L \left(\left(\frac{E\sigma}{1-\sigma^2}\left(-\frac{\partial^2\eta_r}{\partial z^2} + \frac{\eta_r}{R^2}\right) + D_v\left(-\frac{\partial^3\eta_r}{\partial t\partial z^2} + \frac{1}{R^2}\frac{\partial\eta_r}{\partial t}\right)\right)\left(-\frac{\partial^2\xi_r}{\partial z^2} + \frac{\xi_r}{R^2}\right)\right. \\ &\left. + \left(\frac{E}{1+\sigma}\frac{\partial^2\eta_r}{\partial z^2} + (C_v - D_v)\frac{\partial^3\eta_r}{\partial t\partial z^2}\right)\frac{\partial\xi_r}{\partial z^2} + \left(\frac{E}{1+\sigma}\frac{1}{R^2}\eta_r + (C_v - D_v)\frac{1}{R^2}\frac{\partial\eta_r}{\partial t}\right)\frac{\xi_r}{R^2}\right)dz\end{aligned}$$

$\forall \boldsymbol{\xi}(t) \in V_c^0$. Integration by parts gives rise to the equilibrium equation

(2.14)
$$\boxed{f_r = \rho_w h \frac{\partial^2 \eta_r}{\partial t^2} + C_0 \eta_r - C_1 \frac{\partial^2 \eta_r}{\partial z^2} + C_2 \frac{\partial^4 \eta_r}{\partial z^4} + D_0 \frac{\partial \eta_r}{\partial t} - D_1 \frac{\partial^3 \eta_r}{\partial t \partial z^2} + D_2 \frac{\partial^5 \eta_r}{\partial t \partial z^4}},$$

where ρ_w denotes the shell density (see Table 4.1) and

(2.15)
$$C_0 = \frac{h}{R^2} \frac{E}{1-\sigma^2} \left(1 + \frac{h^2}{12R^2}\right) + \frac{p_{\text{ref}}}{R}, \quad C_1 = 2\frac{h^3}{12R^2} \frac{E\sigma}{1-\sigma^2}, \quad C_2 = \frac{h^3}{12} \frac{E}{1-\sigma^2},$$
$$D_0 = \frac{h}{R^2} C_v \left(1 + \frac{h^2}{12R^2}\right), \qquad D_1 = 2\frac{h^3}{12R^2} D_v, \qquad D_2 = \frac{h^3}{12} C_v.$$

We use this equation to model the motion of compliant arterial walls interacting with the time-dependent fluid flow driven by the pulsatile inlet and outlet pressure data. To simplify notation, from this point on in this manuscript we will be using η to denote the radial displacement η_r.

3. Fluid-structure interaction: The three-dimensional model. In medium to large arteries blood can be modeled as an incompressible, Newtonian viscous fluid. We will be assuming that the viscosity of blood is constant, utilizing the data from biomedical literature (see, e.g., [21, 39, 44]), providing the viscosity coefficient $\mu_F = 3500$ kg/ms. The Navier–Stokes equations for a viscous, incompressible fluid have been well accepted as a model for blood flow in medium-to-large arteries. Assuming cylindrical geometry and axially symmetric flow, the fluid velocity $\boldsymbol{v}(r,z,t) = (v_r(r,z,t), v_z(r,z,t))$ and pressure $p(r,z,t)$ satisfy

$$\rho_F \left\{ \frac{\partial v_r}{\partial t} + v_r \frac{\partial v_r}{\partial r} + v_z \frac{\partial v_r}{\partial z} \right\} - \mu_F \left(\frac{\partial^2 v_r}{\partial r^2} + \frac{\partial^2 v_r}{\partial z^2} + \frac{1}{r}\frac{\partial v_r}{\partial r} - \frac{v_r}{r^2} \right) + \frac{\partial p}{\partial r} = 0, \tag{3.1}$$

$$\rho_F \left\{ \frac{\partial v_z}{\partial t} + v_r \frac{\partial v_z}{\partial r} + v_z \frac{\partial v_z}{\partial z} \right\} - \mu_F \left(\frac{\partial^2 v_z}{\partial r^2} + \frac{\partial^2 v_z}{\partial z^2} + \frac{1}{r}\frac{\partial v_z}{\partial r} \right) + \frac{\partial p}{\partial z} = 0, \tag{3.2}$$

$$\frac{\partial v_r}{\partial r} + \frac{\partial v_z}{\partial z} + \frac{v_r}{r} = 0. \tag{3.3}$$

Here ρ_F is the fluid density and μ_F is the fluid dynamic viscosity coefficient, where the subscript F stands for the fluid quantities. The Navier–Stokes equations hold in the cylindrical domain

$$\Omega(t) = \left\{ x \in \mathbb{R}^3; x = (r\cos\vartheta, r\sin\vartheta, z),\ r < R + \eta(z,t),\ 0 < z < L \right\} \tag{3.4}$$

bounded by the viscoelastic lateral boundary

$$\Sigma(t) = \left\{ ((R(z) + \eta(t,z))\cos\theta, (R(z) + \eta(t,z))\sin\theta, z) \in \mathbb{R}^3 : \theta \in (0, 2\pi), z \in (0, L) \right\}.$$

See Figure 3.1. The reference configuration corresponds to that of a straight cylinder with radius R and length L. (The same results can be obtained for a cylinder with a slowly varying radius $R(z)$ under the assumption that $R'(z) < \epsilon$ [47].) The following inlet ($z = 0$) and outlet ($z = L$) boundary data lead to a well-defined problem:

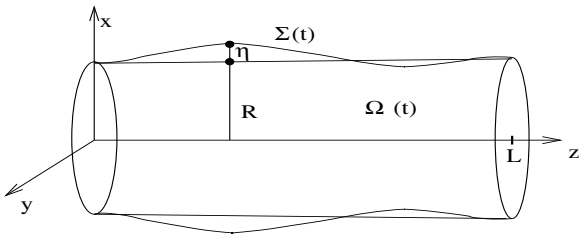


FIG. 3.1. *Deformed domain* $\Omega(t)$.

1. The dynamic pressure is prescribed at both ends:

$$p + \rho(v_z)^2/2 = P_{0,L}(t) + p_{\text{ref}} \text{ at } z = 0, L. \tag{3.5}$$

2. The fluid enters and leaves the tube parallel to the axis of symmetry, with zero displacement:

$$v_r = 0, \quad \eta = 0 \text{ at } z = 0, L. \tag{3.6}$$

3. The tube is clamped so that

$$\frac{\partial \eta}{\partial z} = 0 \text{ at } z = 0, L. \tag{3.7}$$

In the reduced model (see section 4), the zero displacement condition is relaxed. This is typical for reduced models where the boundary layer phenomena near the edges with high stress concentrations are lost [8].

Initially, the fluid and the wall are assumed to be at rest, with zero displacement from the reference configuration:

$$\boldsymbol{v} = 0, \quad \eta = 0, \quad \frac{\partial \eta}{\partial t} = 0. \tag{3.8}$$

These initial and boundary conditions describe well our experimental set up, described in section 7.

The coupling between the fluid flow and vessel wall dynamics is performed via the following kinematic and dynamic lateral boundary conditions [9]:

- The kinematic condition requiring continuity of velocity:

$$v_r(R + \eta(z,t), z, t) = \frac{\partial \eta(z,t)}{\partial t}, \quad v_z(R + \eta(z,t), z, t) = 0. \tag{3.9}$$

- The dynamic condition requiring balance of forces (the contact force of the fluid is counterbalanced by the contact force of the wall):

$$f_r = [(p - p_{\text{ref}})\mathbf{I} - 2\mu_F D(\boldsymbol{v})]\, \boldsymbol{n} \cdot \boldsymbol{e}_r \left(1 + \frac{\eta}{R}\right) \sqrt{1 + (\partial_z \eta)^2}, \tag{3.10}$$

  where f_r is given by the viscoelastic shell model (2.14). The right-hand side of (3.10) describes the contact force of the fluid, where $D(\boldsymbol{v})$ is the symmetrized gradient of velocity, defined in (3.12), $\boldsymbol{n}$ is the vector normal to the deformed boundary $\Sigma(t)$, and $\boldsymbol{e}_r$ is the radial unit vector.

See [9] for more details.

Thus, the complete fluid-structure interaction problem consists of solving the fluid equations (3.1)–(3.3) on the domain $\Omega(t)$ defined by (3.4) with a moving boundary $\Sigma(t)$, satisfying the initial and boundary data given by (3.5)–(3.10) where the contact force of the structure f_r is given by (2.14).

3.1. Weak formulation. To derive a weak formulation of the fluid-structure interaction problem we take the standard approach: multiply the fluid equations by a test function, integrate by parts, and take into account the initial and boundary conditions to obtain the integral form of the problem. For that purpose, introduce the following test spaces.

DEFINITION 3.1 (the test spaces). *Let*

(3.11)

$$V(\Omega(t)) = \{\varphi = \varphi_r \boldsymbol{e}_r + \varphi_z \boldsymbol{e}_z \in H^2(\Omega(t))^2 \mid \varphi_r(r,z) = \partial_z \varphi_r(r,z) = 0 \text{ at } z = 0, L,$$
$$\varphi_z(R + \gamma(z,t), z) = 0, \ \textit{and } \operatorname{div} \varphi = 0 \textit{ in } \Omega(t) \textit{ a.e.}\}.$$

For each $t \in [0,T]$, the test space is the space $H^1(0,T;V(\Omega(t)))$.

To specify the weak solution we introduce the spaces containing the candidates for the radial displacement and the velocity. They are deduced from the a priori solution estimates, presented in section 3.2.

DEFINITION 3.2 (the solution spaces).

- *The space Γ consists of all the functions*

$$\eta \in L^\infty(0,T;H^2(0,L)) \cap C^1([0,T];L^2(0,L)) \cap C([0,T];H^2(0,L))$$

such that $\eta(t,0) = \eta(t,L) = 0$, $\partial_z \eta(t,0) = \partial_z \eta(t,L) = 0$, and $\eta(0,z) = \partial_t \eta(0,z) = 0$.

- *The space V consists of all the functions*

$$v = (v_r, v_z) \in L^2(0,T;H^1(\Omega(t))^2) \cap C([0,T];L^2(\Omega(t))^2)$$

such that $\operatorname{div} \boldsymbol{v} = 0$ in $\Omega(t) \times \mathbb{R}_+$, $v_r = 0$ for $z = 0, L$, and $\boldsymbol{v} = 0$ at $t = 0$.

To define the weak form recall that the symmetrized gradient of velocity $D(\boldsymbol{\varphi})$, defined for an axially symmetric vector valued function $\boldsymbol{\varphi} = \varphi_r \boldsymbol{e}_r + \varphi_z \boldsymbol{e}_z$, is given by

$$D(\boldsymbol{\varphi}) = \begin{pmatrix} \dfrac{\partial \varphi_r}{\partial r} & 0 & \dfrac{1}{2}\Big(\dfrac{\partial \varphi_r}{\partial z} + \dfrac{\partial \varphi_z}{\partial r}\Big) \\ 0 & \dfrac{\varphi_r}{r} & 0 \\ \dfrac{1}{2}\Big(\dfrac{\partial \varphi_r}{\partial z} + \dfrac{\partial \varphi_z}{\partial r}\Big) & 0 & \dfrac{\partial \varphi_z}{\partial z} \end{pmatrix}. \tag{3.12}$$

Define the matrix norm $|\cdot|$ through the scalar product

$$A \cdot B := \operatorname{Tr}\left(AB^T\right), \quad A, B \in \mathbb{R}^9. \tag{3.13}$$

DEFINITION 3.3. *A* weak solution *of problem* (3.1)–(3.10) *is a function $(\eta, \boldsymbol{v}) \in \Gamma \times V$ such that $\forall \varphi \in H^1(0,T;V(\Omega(t)))$ the following integral equation holds:*

$$\begin{aligned} &2\mu_F \int_{\Omega(t)} D(v) \cdot D(\varphi)\ r dr dz + \rho \int_{\Omega(t)} \left\{ \frac{\partial v}{\partial t} + (v(t)\nabla)v \right\} \varphi\ r dr dz \\ &+ R \int_0^L \left\{ C_0 \eta\ \varphi_r|_{R+\eta} + C_1 \frac{\partial \eta}{\partial z}\ \frac{\partial \varphi_r}{\partial z}\bigg|_{R+\eta} + C_2 \frac{\partial^2 \eta}{\partial z^2}\ \frac{\partial^2 \varphi_r}{\partial z^2}\bigg|_{R+\eta} \right. \\ &\left. + D_0 \frac{\partial \eta}{\partial t}\ \varphi_r|_{R+\eta} + D_1 \frac{\partial^2 \eta}{\partial t \partial z}\ \frac{\partial \varphi_r}{\partial z}\bigg|_{R+\eta} + D_2 \frac{\partial^3 \eta}{\partial t \partial z^2}\ \frac{\partial^2 \varphi_r}{\partial z^2}\bigg|_{R+\eta} \right\} dz \\ &+ R\rho_w h \int_0^L \frac{\partial^2 \eta}{\partial t^2} \varphi_r(R + \eta(t,z), z, t)\ dz = -\int_0^R \left\{ P_2(t) - \frac{\rho}{2}(v_z^2)|_{z=L} \right\} \varphi_z|_{z=L} r dr \\ &+ \int_0^R \left\{ P_1(t) - \frac{\rho}{2}(v_z^2)|_{z=0} \right\} \varphi_z|_{z=0} r dr, \end{aligned} \tag{3.14}$$

where $\Omega(t)$ *is given by* (3.4) *and* η *and* v_r *are linked on* $\Sigma(t)$ *through* (3.9).

Notice that the domain as well as the solution and test spaces depend on time. To get a global weak formulation one can use the a priori solution estimates, presented below, and define a global weak solution via a fixed point mapping, defined on a fixed, "fictitious" domain. This approach is used in [9] to define a global weak solution for a related fluid-structure interaction problem using the linearly elastic membrane equations to model the vessel walls. We do not pursue this approach here but continue with the derivation of the energy and a priori estimates.

3.2. The energy and a priori estimates. By replacing the test function with the fluid velocity and using the kinematic lateral boundary condition (3.9) one obtains the following proposition.

PROPOSITION 3.4 (energy equality). *Solution* $(\eta, \boldsymbol{v})$ *of problem* (3.1)–(3.10) *satisfies the following energy equality:*

$$\begin{aligned}
&\frac{\rho}{2}\frac{d}{dt}\int_{\Omega(t)}|\boldsymbol{v}|^2\,dV+\frac{\pi R}{2}\frac{d}{dt}\int_0^L\left\{C_0|\eta|^2+C_1\left|\frac{\partial\eta}{\partial z}\right|^2+C_2\left|\frac{\partial^2\eta}{\partial z^2}\right|^2\right\}dz\\
&+\frac{\pi R}{2}\rho_w h\frac{d}{dt}\int_0^L\left|\frac{\partial\eta}{\partial t}\right|^2dz+\pi R\int_0^L\left\{D_0\left|\frac{\partial\eta}{\partial t}\right|^2+D_1\left|\frac{\partial^2\eta}{\partial t\partial z}\right|^2+D_2\left|\frac{\partial^3\eta}{\partial t\partial z^2}\right|^2\right\}dz\\
&+2\mu_F\|D(\boldsymbol{v})\|^2_{L^2(\Omega(t))}=-\int_0^R P_2(t)v_z(t,r,L)\;rdr+\int_0^R P_1(t)v_z(t,r,0)\;rdr,
\end{aligned}\tag{3.15}$$

with $v_r(t,R+\eta,z)=\frac{\partial\eta}{\partial t}(t,z)$ *and* $v_z(t,R+\eta,z)=0$ *on* $(0,L)\times(0,T)$.

To obtain the a priori estimates and the correct scales for the problem, we introduce the nondimensional time

$$\hat{t}:=\omega t.\tag{3.16}$$

The characteristic frequency ω will be specified later in (3.21). The choice of ω determines the time-scale for the natural oscillations of the structure in terms of the inlet and outlet pressure data. As it will be seen later, the quantity $L\omega$ corresponds to the "sound speed" of the natural oscillations of the structure, and the choice of ω given in (3.21) gives rise to the structure sound speed reported in Fung [21]. From now on we will be working with the nondimensional time $\hat{t}$ but will drop the "hat" notation for simplicity. Whenever physical time t is used, this will be explicitly specified.

Take the rescaled time into account and integrate the energy equality with respect to time to obtain

$$\begin{aligned}
&\frac{\rho\omega}{2}\int_{\Omega(t)}|\boldsymbol{v}|^2\,dV+\frac{\pi R\omega}{2}\int_0^L\left\{C_0|\eta|^2+C_1\left|\frac{\partial\eta}{\partial z}\right|^2+C_2\left|\frac{\partial^2\eta}{\partial z^2}\right|^2\right\}dz\\
&+\frac{\pi R\omega^3}{2}\rho_w h\int_0^L\left|\frac{\partial\eta}{\partial t}\right|^2dz+\pi R\omega^2\int_0^t\int_0^L\left\{D_0\left|\frac{\partial\eta}{\partial t}\right|^2+D_1\left|\frac{\partial^2\eta}{\partial t\partial z}\right|^2+D_2\left|\frac{\partial^3\eta}{\partial t\partial z^2}\right|^2\right\}dzd\tau\\
&+2\mu_F\int_0^t\|D(\boldsymbol{v})\|^2_{L^2(\Omega(\tau))}d\tau=-\int_0^t\int_0^R\left(P_2(\tau)v_z(\tau,r,L)-P_1(\tau)v_z(\tau,r,0)\right)rdr\tau.
\end{aligned}\tag{3.17}$$

By estimating the right-hand side in a manner similar to the estimates in [9] and [6]

one obtains

$$\frac{\rho\omega}{2}\|\boldsymbol{v}\|^2_{L^2(\Omega(t))} + \pi\omega^3\rho_w hR\,\|\partial_t\eta\|^2 + \frac{\pi\omega RC_0}{2}\|\eta\|^2 \leq \frac{16\pi LR\omega}{C_0}\left(\sup_{z,t}|\hat{p}|^2 + \left(\sup_z\int_0^t|\partial_t\hat{p}|d\tau\right)^2\right) + \frac{8T\pi R^2}{\rho\omega L}\int_0^t|A(\tau)|^2d\tau, \tag{3.18}$$

where

$$A(t) = P_L(t) - P_0(t), \quad \hat{p}(t) = \frac{A(t)}{L}z + P_0(t), \tag{3.19}$$

and $T > 0$ denote the physical time such that

$$T \leq \frac{1}{4}\frac{R\sqrt{\rho_w hC_0}}{\|p\|_\infty}. \tag{3.20}$$

For example, for $p_{\text{ref}} = 0$, this inequality reads $T \leq 1/[4(1-\sigma^2)]h\sqrt{E\rho_w}/\|p\|_\infty$.

This is the point were we define the frequency ω. Choose ω so that the contribution of all the terms involving the pressure data have the same weight. Namely, choose ω so that the time-scale of the captured oscillations is determined by the pressure drop $A(t)$, the inlet and outlet maximum pressure, and by the time-average of the steepness of the pressure front $\partial_t\hat{p}$ to obtain

$$\omega = \frac{1}{L}\sqrt{\frac{RC_0}{2\rho}}. \tag{3.21}$$

This choice of ω gives rise to the sound speed of the waves in the "structure" ωL which is exactly the sound speed reported by Fung in [21]. After taking this form of ω into account, and after dividing (3.18) by ω, we obtain the following energy inequality from which the a priori estimates will follow.

PROPOSITION 3.5. *Weak solution* $(\eta, \boldsymbol{v})$ *satisfies*

$$\frac{\rho}{2}\|\boldsymbol{v}\|^2_{L^2(\Omega(t))} + \pi\omega^2\rho_w hR\,\|\partial_t\eta\|^2 + \frac{\pi R}{2}C_0\|\eta\|^2 \leq \frac{16\pi LR}{C_0}\,\mathcal{P}^2, \text{ where}$$

$$\mathcal{P}^2 := \sup_{z,t}|\hat{p}|^2 + \left(\sup_z\int_0^t|\hat{p}_t|d\tau\right)^2 + T\int_0^t|A(\tau)|^2. \tag{3.22}$$

Using this result we obtain the a priori estimates for the L^2-norms of the fluid velocity, the displacement, and the time-derivative of the displacement.

LEMMA 3.6. *Weak solution* $(\eta, \boldsymbol{v})$ *satisfies the following a priori estimates:*

$$\frac{1}{L}\|\eta(t)\|^2_{L^2(0,L)} \leq \frac{32}{C_0^2}\mathcal{P}^2, \quad \frac{1}{L}\,\|\partial_t\eta(t)\|^2_{L^2(0,L)} \leq \frac{16}{\rho_W\omega^2 hC_0}\mathcal{P}^2,$$

$$\frac{1}{LR^2\pi}\|\boldsymbol{v}\|^2_{L^2(\Omega(t))} \leq \frac{32}{\rho_F RC_0}\mathcal{P}^2,$$

$$\int_0^t\left\{\|\partial_r v_r\|^2_{L^2(\Omega(\tau))} + \left\|\frac{v_r}{r}\right\|^2_{L^2(\Omega(\tau))} + \|\partial_z v_z\|^2_{L^2(\Omega(\tau))}\right\}d\tau \leq \frac{4\pi R^2}{\mu_F}\sqrt{\frac{2}{\rho_F RC_0}}\mathcal{P}^2,$$

$$\int_0^t\left\{\|\partial_r v_z\|^2_{L^2(\Omega(\tau))} + \|\partial_z v_r\|^2_{L^2(\Omega(\tau))}\right\}d\tau \leq \frac{4R^2}{\mu_F}\sqrt{\frac{2}{\rho RC_0}}\mathcal{P}^2.$$

Furthermore, we obtain the following estimates for the functions describing the viscoelastic behavior of the structure.

COROLLARY 3.7. *The following estimates hold for the viscoelastic thin shell model:*

$$\frac{\omega}{L}\int_0^t \left\|\frac{\partial\eta}{\partial t}\right\|_{L^2}^2 d\tau \le \frac{32}{C_0D_0}\mathcal{P}^2, \quad \frac{\omega}{L}\int_0^t \left\|\frac{\partial^2\eta}{\partial t\partial z}\right\|_{L^2}^2 d\tau \le \frac{32}{C_0D_1}\mathcal{P}^2,$$
$$\frac{\omega}{L}\int_0^t \left\|\frac{\partial^3\eta}{\partial t\partial^2 z}\right\|_{L^2}^2 d\tau \le \frac{32}{C_0D_0}\mathcal{P}^2,$$

where $\mathcal{P}$ is given by (3.22)*, and ω by* (3.21)*.*

The a priori estimates obtained in this section will be used to derive the reduced model presented below.

4. Fluid-structure interaction: A reduced model. We proceed by deriving a closed, effective, reduced model, approximating the full, original axially symmetric problem to the ϵ^2-accuracy.

We begin by considering (3.1)–(3.3) written in nondimensional form. The scalings for the dependent variables $\boldsymbol{v}$ and η are obtained from the a priori estimates presented in Lemma 3.6

$$\boldsymbol{v} = V\tilde{\boldsymbol{v}}, \text{ where } 2V = \frac{\mathcal{P}}{\sqrt{\rho_F}}\left(\frac{hE}{R(1-\sigma^2)} + p_{\text{ref}}\right)^{-\frac{1}{2}}, \tag{4.1}$$

$$\eta = \Xi\tilde{\eta}, \text{ where } 2\Xi = \mathcal{P}R\left(\frac{hE}{R(1-\sigma^2)} + p_{\text{ref}}\right)^{-1}. \tag{4.2}$$

Consider $p = C_p\tilde{p}$, where C_p will be determined later; see (4.11). The nondimensional independent variables $\tilde{r}$, $\tilde{z}$, and $\tilde{t}$ are introduced via

$$r = R\tilde{r}, \quad z = L\tilde{z}, \quad t = \frac{1}{\omega}\tilde{t}, \text{ where } \omega = \frac{1}{L}\sqrt{\frac{1}{\rho_F}\left(\frac{hE}{R(1-\sigma^2)} + p_{\text{ref}}\right)}. \tag{4.3}$$

At this point we could continue by performing singular perturbation analysis of the rescaled system (3.1)–(3.10), (2.14). As in [9], we would find a two-dimensional reduced free-boundary problem approximating the initial problem to the ϵ^2-accuracy. This problem involves a hydrostatic approximation of the pressure, and it is usually written as an analogue of the shallow water system. Elimination of the radial component of the velocity leads to a nonlocal degenerate term. The resulting equations are too complex to be used in the calculation of the solution, and simplifications involving an ad hoc axial velocity profile are typically considered in the literature. Typically considered v_z-profiles are in the form of a product of an unknown function of z and t and a generalized Poiseuille profile in r (see, e.g., [44]). The resulting variant of the shallow water equations is then closed, but the closure hypothesis could introduce an error of order 1.

In order to find a closure that results from the problem itself and gives rise to an ϵ^2-approximation of the full three-dimensional axially symmetric problem, we are going to use homogenization theory [4]. Homogenization theory is used to find effective equations for nonhomogeneous flows. For porous media problems homogenization theory can be applied when (a) the pore size (characteristic size of the fluid region free of another phase) is smaller than the characteristic length of the macroscopic

problem (here, vessel diameter) or (b) the pore includes a large number of molecules to be considered as continuum [28].

At a first glance using this approach in our setting is pointless. A simple averaging of the equations for the fluid phase over the cross-section of the vessel should provide a good approximation. Unfortunately, as remarked above, this approach leads to a problem that is not closed and might ultimately give rise to the errors of order 1. On the other hand, we know how to obtain closed models related to nonlinear filtration laws in rigid periodic porous media by homogenization [36, 37]. In rigid periodic porous media the expansions are of lower order of precision, but the resulting models are closed. It was shown in [36, 37] that in this case it is possible to link the homogenized equations with the nonlinear algebraic relations between the pressure gradient and the velocity (Forchheimer's filtration law), found in experiments. In a similar way, Robertson and Sequeira [46] obtained a closed model for blood flow in rigid wall tubes by replacing the averaged momentum equation with a variant of Forchheimer's law, and no closure assumption was needed to derive a closed system.

In our case we are concerned with viscoelastic walls. How do we link the flow of blood through viscoelastic arteries with the filtration through porous media? Due to the uniform bound on the maximal value of the radial displacement, obtained in section 3.2, our artery can be placed into a rectangle with the length of order 1 and of small width ϵ. By repeating periodically this geometry in the radial direction, we get a network of parallel, long, and narrow tubes, with no cross-flow from one horizontal tube to another. This is one of the simplest porous media which one can imagine. It is not a rigid but a *deformable* porous medium, just as are the domains in Biot's theories of deformable porous media. All results that are valid for deformable porous media are also valid in our situation. Motivated by the results from [36] and [37], where closed effective porous medium equations were obtained using homogenization techniques, we set up a problem that mimics a similar scenario.

Introduce $y = \frac{1}{\epsilon}\tilde{z}$ and assume periodicity in y of the domain and of the velocity and the pressure. Furthermore, recalling that we have a narrow long tube with $\tilde{r} = \frac{1}{R}r = \frac{1}{\epsilon}\frac{r}{L}$, assume periodicity in the radial direction thereby forming a network of a large number of strictly separated, parallel tubes. Follow the approach first presented in [9]. In [9] a closed, reduced model was derived in the case when the vessel walls were approximated by a linearly elastic membrane equations. In the present manuscript, the introduction of a linearly viscoelastic Koiter shell model introduces minor differences in the derivation of the reduced model. Thus, we present only the main steps in the derivation and omit the details which can be found in [9].

Following standard approach in homogenization theory [28, 4], we look for the unknown functions that explicitly depend on the "slow variables" r and $\tilde{z}$ as well as on the "fast variables" r/ϵ and $\tilde{z}/\epsilon =: y$. In our problem the slow and fast variables are related through $z = L\tilde{z} := L\epsilon y = Ry$, $r = R\tilde{r}$. Thus, we look for the functions

$$\tilde{\boldsymbol{v}} = \tilde{\boldsymbol{v}}(\tilde{t}, r, r/\epsilon, \tilde{z}, \tilde{z}/\epsilon), \quad \tilde{\eta} = \tilde{\eta}(\tilde{t}, r, r/\epsilon, \tilde{z}, \tilde{z}/\epsilon), \quad \text{and } \tilde{p} = \tilde{p}(\tilde{t}, r, r/\epsilon, \tilde{z}, \tilde{z}/\epsilon) \tag{4.4}$$

that are 1-periodic in $y = \tilde{z}/\epsilon$ and r/ϵ and satisfy the Navier–Stokes equations (3.1)–(3.3). Keeping both the fast and the slow variables in the derivation of the equations, namely keeping r, r/ϵ, $\tilde{z}$, and y in the problem, will help us determine the proper scaling for the pressure and lead us to a closed, reduced effective model.

Expand the functions in (4.4) in terms of the small parameter ϵ

$$\boldsymbol{v} = V\left\{\tilde{\boldsymbol{v}}^0 + \epsilon\tilde{\boldsymbol{v}}^1 + \cdots\right\}, \quad \eta = \Xi\left\{\tilde{\eta}^0 + \epsilon\tilde{\eta}^1 + \cdots\right\}, \quad p = C_p\left\{\tilde{p}^0 + \epsilon\tilde{p}^1 + \cdots\right\} \tag{4.5}$$

TABLE 4.1
Table with parameter values.

Parameters	Aorta/iliacs	Latex Tube
Char. radius R(m)	0.006-0.012 [44]	0.011
Char. length L(m)	0.065-0.2 [14]	0.34
Dyn. viscosity $\mu_F(\frac{\text{kg}}{\text{ms}})$	3.5×10^{-3} [44]	3.5×10^{-3}
Young's modulus E(Pa)	$10^5 - 10^6$ [44, 1, 3]	1.0587×10^6
Wall thickness h(m)	$1 - 2 \times 10^{-3}$ [44]	0.0009
Wall density $\rho_W(\text{kg/m}^3)$	1.1×10^3 [44]	1.1×10^3
Fluid density $\rho_F(\text{kg/m}^3)$	1050 [44]	1000
Wall viscosity coef. $hC_v/R(\text{Pa} \cdot \text{s})$	$10^3 - 8 \times 10^3$ [1, 2, 3]	0

and plug this into the Navier–Stokes equations (3.1)–(3.3). We look for a solution to the zeroth-order approximation of the problem plus its ϵ-correction. The zeroth-order approximation corresponds to the leading-order approximation of the flow in the limit in which the wavelength of the disturbance and the length scale of tube variation are large compared with the tube radius.

4.1. The zeroth-order approximation. The leading-order Navier–Stokes equations read

$$Sh_0 \frac{\partial \tilde{v}_z^0}{\partial \tilde{t}} + (\tilde{v}^0 \nabla_{\tilde{r},y}) \tilde{v}_z^0 + \frac{\partial \tilde{p}^0}{\partial \tilde{z}} + \frac{\partial \tilde{p}^1}{\partial y} - \frac{1}{Re_0} \left\{ \frac{1}{\tilde{r}} \frac{\partial}{\partial \tilde{r}} \left(\tilde{r} \frac{\partial \tilde{v}_z^0}{\partial \tilde{r}} \right) + \frac{\partial^2 \tilde{v}_z^0}{\partial y^2} \right\} = 0, \tag{4.6}$$

$$Sh_0 \frac{\partial \tilde{v}_r^0}{\partial \tilde{t}} + (\tilde{v}^0 \nabla_{\tilde{r},y}) \tilde{v}_r^0 + \frac{\partial \tilde{p}^0}{\partial r} + \frac{\partial \tilde{p}^1}{\partial \tilde{r}} - \frac{1}{Re_0} \left\{ \frac{1}{\tilde{r}} \frac{\partial}{\partial \tilde{r}} \left(\tilde{r} \frac{\partial \tilde{v}_r^0}{\partial \tilde{r}} \right) + \frac{\partial^2 \tilde{v}_r^0}{\partial y^2} \right\} = 0, \tag{4.7}$$

$$\nabla_{\tilde{r},y} \tilde{p}^0 = 0, \tag{4.8}$$

$$\frac{\partial}{\partial \tilde{r}} \left(\tilde{r} \tilde{v}_r^0 \right) + \frac{\partial}{\partial y} \left(\tilde{r} \tilde{v}_z^0 \right) = 0, \tag{4.9}$$

$$\text{with } \tilde{v}_r^0, \tilde{v}_z^0, \text{ and } \tilde{p}^1 \text{ 1-periodic in } y \text{ and } \tilde{v}_r^0 = \tilde{v}_z^0 = 0 \text{ at } \tilde{r} = 1 + \frac{\Xi}{R} \tilde{\eta}, \tag{4.10}$$

where $Sh_0 := \frac{\epsilon L \omega^\epsilon}{V}$ and $Re_0 := \frac{\rho_F R V}{\mu_F}$. Here the following scaling for the pressure is used:

$$p = \frac{\rho_F V^2}{\epsilon} \tilde{p}, \quad \text{thus} \quad C_p = \frac{\rho_F V^2}{\epsilon}. \tag{4.11}$$

Notice $Sh_0 = \epsilon Sh$ and $Re_0 = Re/\epsilon$. For the average values from Table 4.1 Sh_0 is of order 1 and Re_0 is around 1000. We remark that (4.8) corresponds to the ϵ^{-1}-term and the others to the ϵ^0-term.

The leading-order behavior for the boundary conditions evaluated at the lateral boundary $\tilde{r} = 1 + \frac{\Xi}{R} \tilde{\eta}^0$ is the following:

- The kinematic boundary condition:

$$\tilde{v}_r^1 = \frac{\partial \tilde{\eta}^0}{\partial \tilde{t}} + \mathcal{O}(\epsilon^2). \tag{4.12}$$

- The dynamic boundary condition:

$$\begin{aligned} \tilde{p}^0 - \tilde{p}_{\text{ref}} = {} & \frac{\epsilon}{\rho_F V^2} \frac{\Xi}{R} \frac{hE}{R(1-\sigma^2)} \left(1 + \frac{h^2}{12R^2} \right) \tilde{\eta}^0 + \tilde{p}_{\text{ref}} \frac{\Xi}{R} \tilde{\eta}^0 \\ & + \frac{\epsilon}{\rho_F V^2} \frac{\Xi}{R} \frac{hC_V \omega}{R} \left(1 + \frac{h^2}{12R^2} \right) \frac{\partial \tilde{\eta}^0}{\partial \tilde{t}} + \mathcal{O}(\epsilon^2). \end{aligned} \tag{4.13}$$

Notice that for the parameter values in Table 4.1, $\omega \approx 100$ and the values of the leading-order coefficients are both of order one: $\frac{\epsilon}{\rho_F V^2}\frac{\Xi}{R}\frac{hE}{R(1-\sigma^2)}(1+\frac{h^2}{12R^2}) = \mathcal{O}(1)$, $\frac{\epsilon}{\rho_F V^2}\frac{\Xi}{R}\frac{hC_V\omega}{R}(1+\frac{h^2}{12R^2}) = \mathcal{O}(1)$. This is the ϵ^2-approximation of the pressure-displacement relationship describing the linearly viscoelastic cylindrical Koiter shell model. The terms multiplying h^3 account for the bending rigidity of the Koiter shell. These terms are not present in the pressure-displacement relationship describing a viscoelastic membrane.

To obtain a closed system of reduced equations notice that system (4.6)–(4.10) admits a unique strong (nonstationary) unidirectional solution independent of y [48] for every given smooth pressure $\tilde{p}^0$:

$$\tilde{v}_r^0 = 0, \quad \tilde{v}_z^0 = \tilde{v}_z^0(\tilde{r}, \tilde{z}, \tilde{t}), \tag{4.14}$$

where $\tilde{v}_z^0$ satisfies

$$\begin{cases} Sh_0 \dfrac{\partial \tilde{v}_z^0}{\partial \tilde{t}} - \dfrac{1}{Re_0}\dfrac{1}{\tilde{r}}\dfrac{\partial}{\partial \tilde{r}}\left(\tilde{r}\dfrac{\partial \tilde{v}_z^0}{\partial \tilde{r}}\right) = -\dfrac{\partial \tilde{p}^0}{\partial \tilde{z}}(\tilde{z}, \tilde{t}), \\ \tilde{v}_z^0(0, \tilde{z}, \tilde{t}) \text{ bounded}, \quad \tilde{v}_z^0(1 + \Xi\tilde{\eta}^0(\tilde{z}, \tilde{t})/R, \tilde{z}, \tilde{t}) = 0, \quad \text{and } \tilde{v}_z^0(\tilde{r}, \tilde{z}, 0) = 0, \end{cases} \tag{4.15}$$

and $\tilde{p}^1$ is a linear function of y, independent of $\tilde{r}$. Since $\tilde{p}^1$ is 1-periodic $\tilde{p}^1$ cannot depend on y. Thus, the derivatives of $\tilde{p}^1$ with respect to $\tilde{r}$ and y are both zero.

To complement (4.15) in the calculation of $\tilde{v}_z^0$ and $\tilde{p}^0$ we use the conservation of mass equation (3.3) averaged with respect to the cross-section. The leading-order terms in (3.3) read

$$\frac{\partial}{\partial \tilde{r}}\left(\tilde{r}\tilde{v}_r^1\right) + \frac{\partial}{\partial \tilde{z}}\left(\tilde{r}\tilde{v}_z^0\right) = 0.$$

Integrated with respect to $\tilde{r}$ from 0 to $1 + \frac{\Xi}{R}\tilde{\eta}^0$ one obtains

$$\frac{\partial\left(1 + \frac{\Xi}{R}\tilde{\eta}^0\right)^2}{\partial \tilde{t}} + \frac{\Xi}{R}\frac{\partial}{\partial \tilde{z}}\int_0^{1+\frac{\Xi}{R}\tilde{\eta}^0} 2\tilde{v}_z^0\tilde{r}\,d\tilde{r} = 0, \tag{4.16}$$

where we have used the kinematic boundary condition (4.12) to couple the flow velocity and lateral boundary motion.

Equations (4.16), (4.15), and (4.13) give rise to a nonlinear free-boundary problem for the zeroth-order approximation of the flow. In dimensional variables, the nonlinear free-boundary problem for $(\boldsymbol{v}^0, \eta^0, p^0) = (v_z^0, 0, \eta^0, p^0)$ reads

$$\begin{aligned} &\frac{\partial (R+\eta^0)^2}{\partial t} + \frac{\partial}{\partial z}\int_0^{R+\eta^0} 2rv_z^0 dr = 0, \\ &\varrho_F \frac{\partial v_z^0}{\partial t} - \mu_F \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_z^0}{\partial r}\right) = -\frac{\partial p^0}{\partial z}, \\ &p^0 - p_{\text{ref}} = \frac{hE}{R^2(1-\sigma^2)}\left(1 + \frac{h^2}{12R^2}\right)\eta^0 + p_{\text{ref}}\frac{\eta^0}{R} + \frac{hC_V}{R^2}\left(1 + \frac{h^2}{12R^2}\right)\frac{\partial \eta^0}{\partial t}, \\ &v_z^0|_{r=0} - \text{ bounded}, \quad v_z^0|_{r=R+\eta^0} = 0, \quad v_z^0|_{t=0} = 0, \\ &\eta^0|_{t=0} = 0, \quad p^0|_{z=0} = P_0, \quad p^0|_{z=L} = P_L. \end{aligned} \tag{4.17}$$

4.2. The first-order correction. The first-order correction to the solution defined by (4.17) is obtained by solving the equations that result from the coefficients at the ϵ^1-terms in the expanded Navier–Stokes equations (3.1)–(3.3)

(4.18)
$$Sh_0\frac{\partial\tilde{v}_z^1}{\partial\tilde{t}}+\tilde{v}_z^0\left\{\frac{\partial\tilde{v}_z^1}{\partial y}+\frac{\partial\tilde{v}_z^0}{\partial\tilde{z}}\right\}+\tilde{v}_r^1\frac{\partial\tilde{v}_z^0}{\partial\tilde{r}}+\frac{\partial\tilde{p}^1}{\partial\tilde{z}}+\frac{\partial\tilde{p}^2}{\partial y}=\frac{1}{Re_0}\left\{\frac{1}{\tilde{r}}\frac{\partial}{\partial\tilde{r}}\left(\tilde{r}\frac{\partial\tilde{v}_z^1}{\partial\tilde{r}}\right)+\frac{\partial^2\tilde{v}_z^1}{\partial y^2}\right\},$$

$$Sh_0\frac{\partial\tilde{v}_r^1}{\partial\tilde{t}}+\tilde{v}_z^0\frac{\partial\tilde{v}_r^1}{\partial y}+\frac{\partial\tilde{p}^2}{\partial\tilde{r}}=\frac{1}{Re_0}\left\{\frac{1}{\tilde{r}}\frac{\partial}{\partial\tilde{r}}\left(\tilde{r}\frac{\partial\tilde{v}_r^1}{\partial\tilde{r}}\right)+\frac{\partial^2\tilde{v}_r^1}{\partial y^2}\right\}, \tag{4.19}$$

$$\frac{\partial}{\partial\tilde{r}}\left(\tilde{r}\tilde{v}_r^1\right)+\frac{\partial}{\partial y}\left(\tilde{r}\tilde{v}_z^1\right)+\tilde{r}\frac{\partial\tilde{v}_z^0}{\partial\tilde{z}}=0, \tag{4.20}$$

$$\tilde{v}_r^1,\tilde{v}_z^1,\tilde{p}^2 \text{ 1-periodic in } y;\ \tilde{v}_r^1=\frac{\partial\tilde{\eta}^0}{\partial\tilde{t}},\quad \tilde{v}_z^0=0 \text{ at } \tilde{r}=1+\frac{\Xi}{R}\tilde{\eta}^0. \tag{4.21}$$

Using the same arguments as in [9] one can show that $\tilde{p}^1=\tilde{p}^2=0$ and we have a closed linear system, known as a nonstationary Oseen system, defined on a fixed domain $(0,L)\times(0,1+\Xi/R\eta^0)$.

To calculate the ϵ-correction to the velocity we look for a solution $\tilde{v}_z^1$ that is independent of the "artificial" fast variable y. In this case the conservation of mass equation (4.20) can be integrated with respect to $\tilde{r}$ to obtain an explicit formula for $\tilde{v}_r^1$ in terms of the already calculated $\tilde{v}_z^0$:

$$\tilde{r}\tilde{v}_r^1(\tilde{r},\tilde{z},\tilde{t})=\left(1+\frac{\Xi\tilde{\eta}^0}{R}\right)\frac{\partial\tilde{\eta}^0}{\partial\tilde{t}}+\int_{\tilde{r}}^{1+\Xi\tilde{\eta}^0/R}\frac{\partial\tilde{v}_z^0}{\partial\tilde{z}}(\xi,\tilde{z},\tilde{t})\ \xi\ d\xi. \tag{4.22}$$

The axial momentum equation (4.18) defines a linear problem for $\tilde{v}_z^1$:

$$Sh_0\frac{\partial\tilde{v}_z^1}{\partial\tilde{t}}-\frac{1}{Re_0}\frac{1}{\tilde{r}}\frac{\partial}{\partial\tilde{r}}\left(\tilde{r}\frac{\partial\tilde{v}_z^1}{\partial\tilde{r}}\right)=-\tilde{v}_r^1\frac{\partial\tilde{v}_z^0}{\partial\tilde{r}}-\frac{\partial}{\partial\tilde{z}}\left(\frac{(\tilde{v}_z^0)^2}{2}\right), \tag{4.23}$$

$$\tilde{v}_z^1(0,\tilde{z},\tilde{t}) \text{ bounded},\quad \tilde{v}_z^1(1+\Xi\tilde{\eta}^0(\tilde{z},\tilde{t})/R,\tilde{z},\tilde{t})=0, \tag{4.24}$$

$$\tilde{v}_z^1(\tilde{r},\tilde{z},0)=0,\quad \tilde{v}_z^1(\tilde{r},0,t)=\tilde{v}_z^1(\tilde{r},L,t)=0. \tag{4.25}$$

Notice that the quadratic transport terms appear in this higher-order approximationi. They are linearized around the zeroth-order approximation of the solution.

Equations (4.22)–(4.25) define the ϵ-correction of the solution. In dimensional form the system reads

$$\begin{aligned}
&v_r^1(r,z,t)=\frac{1}{r}\left(R\frac{\partial\eta^0}{\partial t}+\int_r^R\xi\frac{\partial v_z^0}{\partial z}(\xi,z,t)d\xi\right),\\
&\rho_F\frac{\partial v_z^1}{\partial t}-\mu_F\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_z^1}{\partial r}\right)=-\rho_F\left(v_r^1\frac{\partial v_z^0}{\partial r}+v_z^0\frac{\partial v_z^0}{\partial z}\right),\\
&v_z^1|_{\overline{r}=0}- \text{ bounded},\quad v_z^1|_{r=R}=0,\quad v_z^1|_{t=0}=0.
\end{aligned} \tag{4.26}$$

PROPOSITION 4.1. *The velocity field* $\boldsymbol{v}=(v_z^0+v_z^1,v_r^1)$*, the radial displacement* $\eta=\eta^0$*, and the pressure* $p=p^0$*, defined by* (4.17) *and* (4.26)*, satisfy the original problem* (3.1)–(3.10) *to* $\mathcal{O}(\epsilon^2)$.

The proof is the same as that of Proposition 7.1 in [9].

We end this section by summarizing the main assumptions under which the simplified, effective problem (4.17), (4.26) holds and the parameter values assumed.

Assumptions.

(1) The domain is cylindrical with small aspect ratio $\epsilon = R_{\max}/L$.
(2) The problem is axially symmetric.
(3) Longitudinal displacement is negligible.
(4) Radial displacement is not too large, i.e., $\delta := \Xi/R \leq \epsilon$.
(5) The reference tube radius varies slowly: $R'(z) < \epsilon$.
(6) The Reynolds number Re is small to medium (Re ≈ 1000).
(7) The z-derivatives of the nondimensional quantities are $O(1)$ (not too large).

5. Viscoelasticity of the fluid-structure interaction. We emphasize in this section that the viscoelastic behavior of the coupled fluid-structure interaction problem comes from two distinct effects. One is the viscoelasticity of the structure itself, and the other is the viscoelasticity due to the interaction between the structure (not necessarily viscoelastic) with a viscous fluid. To explicitly capture the leading-order effects that the viscous fluid imparts on the motion of the structure we proceed as follows. First, we simplify the free-boundary problem (4.17) by expanding the underlying problem (4.17), (4.26), with respect to the radial displacement. The free-boundary problem will be approximated by two fixed boundary problems of similar form. Each of the two fixed boundary problems consists of solving a system of two equations (see (5.1), (5.3)) that are of hyperbolic-parabolic type. In each of the two problems, we can "explicitly solve" the parabolic equation for the velocity, plug the velocity into the resulting equation for the structure, and obtain a single equation describing the motion of the structure. The resulting equation incorporates the viscous fluid effects in terms of a convolution integral. If we will assume, for the moment, that the structure is purely elastic, the resulting equation describes the dynamics of an elastic structure under a viscous fluid load; see (5.11). It corresponds to a model of a viscoelastic string with viscous long-term memory effects. Thus, the fluid viscosity influences the dynamics of an elastic structure through a long-term memory effect.

We begin by expanding the free-boundary problem (4.17) and the ϵ-correction (4.26) with respect to the radial displacement whose magnitude is measured, in nondimensional variables, by Ξ/R. Thus, assume that

$$\delta := \frac{\Xi}{R} \leq \epsilon$$

and introduce the following expansions with respect to δ:

$$\tilde{\eta}^0 = \tilde{\eta}^{0,0} + \delta\tilde{\eta}^{0,1} + \cdots, \quad \tilde{p}^0 = \tilde{p}^{0,0} + \delta\tilde{p}^{0,1} + \cdots,$$
$$v_z^0 = v_z^{0,0} + \delta v_z^{0,1} + \cdots, \quad \tilde{v}_z^1 = \tilde{v}_z^{1,0} + \cdots, \quad \tilde{v}_r^1 = \tilde{v}_r^{1,0} + \cdots.$$

The first superscript denotes the expansion with respect to ϵ and the second with respect to δ. Then using the same approach as in [9] one obtains a set of equations approximating the original problem to the ϵ^2-accuracy. The resulting problem, in dimensional variables, consists of finding the functions

$$v_z = v_z^{0,0} + v_z^{0,1} + v^{1,0} + \mathcal{O}(\epsilon^2),\ v_r = v_r^{1,0} + \mathcal{O}(\epsilon^2),\ \eta = \eta^{0,0} + \mathcal{O}(\epsilon^2)\, p = p^{0,0} + \mathcal{O}(\epsilon^2)$$

satisfying the following set of closed, well-defined problems.

The zeroth-order approximation. Find $(\eta^{0,0}, v_z^{0,0})$ such that

$$
\begin{aligned}
&\frac{\partial \eta^{0,0}}{\partial t} + \frac{1}{R}\frac{\partial}{\partial z}\int_0^R r v_z^{0,0} dr = 0,\\
&\varrho_F \frac{\partial v_z^{0,0}}{\partial t} - \mu_F \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_z^{0,0}}{\partial r}\right) = -\frac{\partial p^{0,0}}{\partial z},\\
&v_z^{0,0}|_{r=0} - \text{ bounded}, \quad v_z^{0,0}|_{t=R} = 0, \quad v_z^{0,0}|_{t=0} = 0,\\
&\eta^{0,0}|_{t=0} = 0, \quad p^{0,0}|_{z=0} = P_0, \quad p^{0,0}|_{z=L} = P_L,
\end{aligned}
\tag{5.1}
$$

where

$$
p^{0,0} = \frac{Eh}{(1-\sigma^2)R}\left(1 + \frac{h^2}{12R^2}\right)\frac{\eta^{0,0}}{R} + p_{\text{ref}}\frac{\eta^{0,0}}{R} + \frac{hC_v}{R^2}\left(1 + \frac{h^2}{12R^2}\right)\frac{\partial \eta^{0,0}}{\partial t}. \tag{5.2}
$$

The δ correction. Find $(\eta^{0,1}, v_z^{0,1})$ such that

$$
\begin{aligned}
&\frac{\partial \eta^{0,1}}{\partial t} + \frac{1}{R}\frac{\partial}{\partial z}\int_0^R r v_z^{0,1} dr = -\frac{1}{2R}\frac{\partial}{\partial t}\left(\eta^{0,0}\right)^2,\\
&\varrho_F \frac{\partial v_z^{0,1}}{\partial t} - \mu_F \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_z^{0,1}}{\partial r}\right) = -\frac{\partial p^{0,1}}{\partial z},\\
&v_z^{0,1}|_{r=0} - \text{ bounded}, \quad v_z^{0,1}|_{r=R} = -\eta^{0,0}\frac{\partial v_z^{0,0}}{\partial r}|_{r=R}, \quad v_z^{0,1}|_{t=0} = 0,\\
&\eta^{0,1}|_{t=0} = 0, \quad \eta^{0,1}|_{z=0} = 0, \quad \eta^{0,1}|_{z=L} = 0,
\end{aligned}
\tag{5.3}
$$

where

$$
\begin{aligned}
p^{0,1} = &\left(\frac{Eh}{(1-\sigma^2)R}\left(1 + \frac{h^2}{12R^2}\right) + p_{\text{ref}}\right)\left(\frac{\eta^{0,1}}{R} - \left(\frac{\eta^{0,0}}{R}\right)^2\right)\\
&+ \frac{hC_v}{R^2}\left(1 + \frac{h^2}{12R^2}\right)\left(\frac{\partial \eta^{0,1}}{\partial t} - \frac{\eta^{0,0}}{R}\frac{\partial \eta^{0,1}}{\partial t}\right).
\end{aligned}
\tag{5.4}
$$

The ϵ-correction. Find $(v_r^{1,0}, v_z^{1,0})$ such that

$$
v_r^{1,0}(r,z,t) = \frac{1}{r}\left(R\frac{\partial \eta^{0,0}}{\partial t} + \int_r^R \xi \frac{\partial v_z^{0,0}}{\partial z}(\xi,z,t)d\xi\right), \tag{5.5}
$$

$$
\begin{aligned}
&\rho_F \frac{\partial v_z^{1,0}}{\partial t} - \mu_F \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_z^{1,0}}{\partial r}\right) = -\rho_F\left(v_r^{1,0}\frac{\partial v_z^{0,0}}{\partial r} + v_z^{0,0}\frac{\partial v_z^{0,0}}{\partial z}\right),\\
&v_z^{1,0}|_{\bar{r}=0} - \text{ bounded}, \quad v_z^{1,0}|_{r=R} = 0, \quad v_z^{1,0}|_{t=0} = 0.
\end{aligned}
\tag{5.6}
$$

Systems (5.1) and (5.3) can be solved by considering the auxiliary problem

$$
\begin{cases}
\dfrac{\partial \zeta}{\partial t} - \dfrac{1}{r}\dfrac{\partial}{\partial r}\left(r\dfrac{\partial \zeta}{\partial r}\right) = 0 \quad \text{in } (0,R)\times(0,\infty),\\
\zeta(0,t) \text{ is bounded}, \quad \zeta(R,t) = 0 \text{ and } \zeta(r,0) = 1.
\end{cases}
\tag{5.7}
$$

For example, the solution of the parabolic equation for the velocity $v_z^{0,0}$ can be written as the convolution

$$
v_z^{0,0} = -\frac{1}{\rho_F}\int_0^t \zeta\left(r, \frac{\mu_F(t-\tau)}{\rho_F}\right)\frac{\partial p^{0,0}}{\partial z}(z,\tau)d\tau.
$$

Plugging this expression for the velocity into the first equation one obtains

$$\frac{\partial \eta^{0,0}}{\partial t} - \frac{1}{\rho_F R}\frac{\partial}{\partial z}\int_0^R r \int_0^t \zeta\left(r, \frac{\mu_F(t-\tau)}{\rho_F}\right)\frac{\partial p^{0,0}}{\partial z}(z,\tau)d\tau dr = 0. \tag{5.8}$$

Denote the mean of ζ in the radial direction by

$$\mathcal{K}(t) = 2\int_0^R \zeta(r,t)\, r dr, \tag{5.9}$$

and assume, for the moment, that the Koiter shell is purely elastic so that

$$p^{0,0} = C_0\eta^{0,0}, \text{ where } C_0 = \frac{h}{R^2}\frac{E}{1-\sigma^2}\left(1+\frac{h^2}{12R^2}\right) + \frac{p_{\text{ref}}}{R}.$$

Then (5.8) becomes

$$\frac{\partial \eta^{0,0}}{\partial t} - \frac{C_0}{2\rho_F R}\int_0^t \mathcal{K}\left(\frac{\mu_F(t-\tau)}{\rho_F}\right)\frac{\partial^2 \eta^{0,0}}{\partial z^2}d\tau = 0. \tag{5.10}$$

Differentiate with respect to t to obtain

$$\frac{\partial^2 \eta^{0,0}}{\partial t^2} = \frac{C_0 R}{2\rho_F}\frac{\partial^2 \eta^{0,0}}{\partial z^2} + \mu_F\frac{C_0}{2\rho_F^2 R}\int_0^t \mathcal{K}'\left(\frac{\mu_F(t-\tau)}{\rho_F}\right)\frac{\partial^2 \eta^{0,0}}{\partial z^2}. \tag{5.11}$$

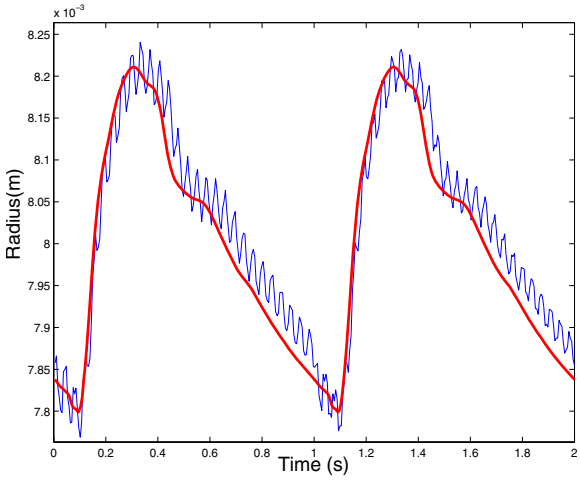


FIG. 5.1. *A comparison between the solutions of* (5.11) *with* $\mu_F = 0$ *(thin solid line) and* $\mu_F = 3.5 \times 10^{-3}$ *(thick solid line). The radius, shown in these graphs, is taken at the midpoint of the tube during two cardiac cycles.*

This is a model describing the motion of a linearly viscoelastic string with the viscous effects described by the convolution integral on the right-hand side of (5.11). The kernel in the convolution corresponds to the derivative of $\mathcal{K}$ which decays in time exponentially fast, with the decay rate equal to the first zero of the Bessel function J_0. This is the only term that incorporates the viscosity of the fluid μ_F. Thus, the fluid impacts the motion of the structure through this long-term memory effect. Numerical simulations presented in Figure 5.1 show the motion of the structure (displacement $\eta^{0,0}$) with $\mu_F = 0$ and with $\mu_F = 3.5 \times 10^{-3}$. The smoothing by the viscous fluid dissipation is obvious.

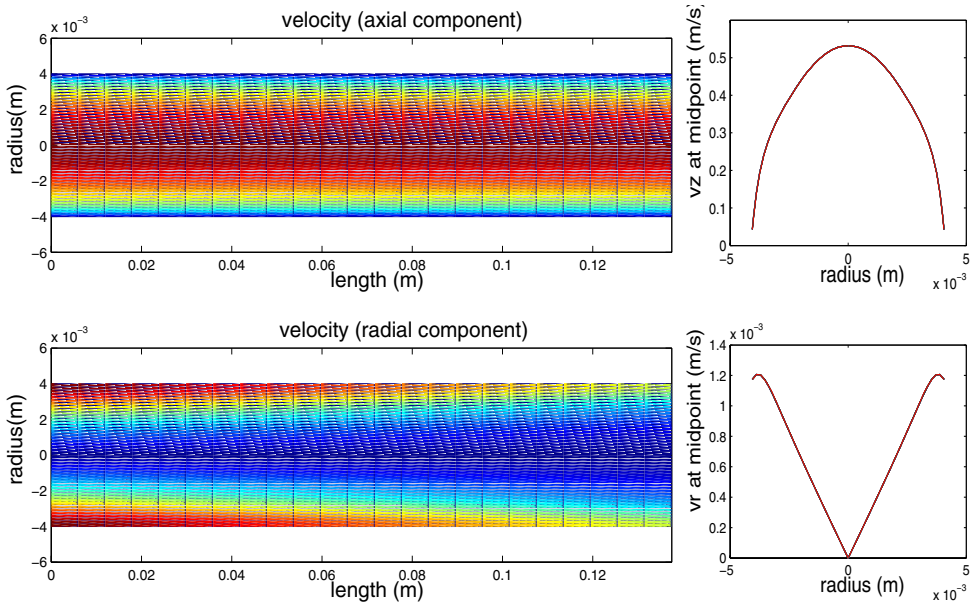


FIG. 6.1. *The axial (top) and radial (bottom) components of the velocity obtained at a first half of the systole (*1/6 *of the cardiac cycle). The magnitude of the axial component of the velocity is between* 0 *and* 0.52 m/s. *The magnitude of the radial component of the velocity is between* 0 *and* 0.0014 m/s. *The pictures on the right show the velocity profiles calculated at the midpoint of the tube.*

6. Numerical algorithm. To solve problems (5.1) and (5.3) numerically it is convenient to rewrite each of the systems of equations as a second-order hyperbolic-parabolic problem. Namely, after differentiating the first equation in (5.1) with respect to time, and plugging the second equation into the first, problem (5.1) can be rewritten as

$$\frac{\partial^2 \eta^{0,0}}{\partial t^2} - \frac{R}{2\rho_F}\frac{\partial^2 p^{0,0}}{\partial z^2} = -\frac{\mu_F}{\rho_F}\frac{\partial}{\partial z}\left(\frac{\partial v_z^{0,0}}{\partial r}\bigg|_{r=R}\right), \tag{6.1}$$

$$\rho_F \frac{\partial v_z^{0,0}}{\partial t} - \mu_F \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_z^{0,0}}{\partial r}\right) = -\frac{\partial p^{0,0}}{\partial z}, \tag{6.2}$$

with the initial and boundary conditions specified in (5.1) and $p^{0,0}$ substituted by (5.2). Similarly, problem (5.3) can be written as

$$\frac{\partial^2 \eta^{0,1}}{\partial t^2} - \frac{R}{2\rho_F}\frac{\partial^2 p^{0,1}}{\partial z^2} = -\frac{\mu_F}{\rho_F}\frac{\partial}{\partial z}\left(\frac{\partial v_z^{0,0}}{\partial r}\bigg|_{r=R}\right) - \frac{1}{2R}\frac{\partial^2}{\partial t^2}\left(\eta^{0,0}\right)^2, \tag{6.3}$$

$$\rho_F \frac{\partial v_z^{0,1}}{\partial t} - \mu_F \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_z^{0,1}}{\partial r}\right) = -\frac{\partial p^{0,1}}{\partial z}, \tag{6.4}$$

with initial and boundary conditions given in (5.3) and $p^{0,1}$ substituted by (5.4). The first equation in both subproblems can be thought off as a one-dimensional wave equation in z and t, and the second as the one-dimensional heat equation in r and t. The systems for the $0,0$ and $0,1$ approximations have the same form. They are solved using a one-dimensional finite element method. Since the mass and stiffness matrices are the same for both problems, up to the boundary conditions, they are

generated only once. Both systems are solved simultaneously using a time-iteration procedure. First, the parabolic equation is solved for $v_z^{0,0}$ at the time step t_{i+1} by explicitly evaluating the right-hand side at the time-step t_i. Then the wave equation is solved for $\eta^{0,0}$ with the evaluation of the right-hand side at the time-step t_{i+1}. Using these results for $v_z^{0,0}$ and $\eta^{0,0}$, computed at t_{i+1}, a correction at t_{i+1} is calculated by repeating the process with the updated values of the right-hand sides. This method is a version of the Douglas–Rachford time-splitting algorithm which is known to be of first-order accuracy.

Calculating approximation $1,0$ is straightforward once the approximations $0,0$ and $0,1$ are obtained. In this algorithm a sequence of one-dimensional problems is solved, so the numerical complexity is that of one-dimensional solvers. However, leading-order two-dimensional effects are captured to the ϵ^2-accuracy. Figure 6.1 presents the axial and radial components of the velocity, showing two-dimensional effects that cannot be captured using one-dimensional models.

7. Experimental validation. A mock circulatory loop was used to validate our simplified, effective mathematical flow model (5.1)–(5.6). The circulatory loop was assembled at the Research Laboratory at the Texas Heart Institute. Figure 7.1 shows the experimental setup and a sketch of the main components of the mock circulatory loop. The main components of the flow loop include the left ventricular assist device (LVAD Heart Mate, Thoratex Corp., Woburn, MA), which is a pulsatile flow pump used in patients with failing hearts to aid the function of the left ventricle, the inlet and outlet LVAD valves, two compliance chambers (wash bottles; 250 ml in volume), a reservoir (Nalgene canister), and pressure transducers (TruEave, Edwards Lifesciences, Irvine, CA) placed at the inlet and outlet of the test segment. Latex tubing (Kent Elastomer Products Inc.) was used to simulate compliant vessels. See Figure 7.1. The straight latex tube segment was attached to the hard plastic connectors placed at the inlet and at the outlet of the segment, keeping the inlet and outlet displacement together with its derivative equal to zero, i.e., $\eta = \partial\eta/\partial z = 0$ at $z = 0, L$, as well as the inlet and outlet velocity approximately such that $v_r = 0$.

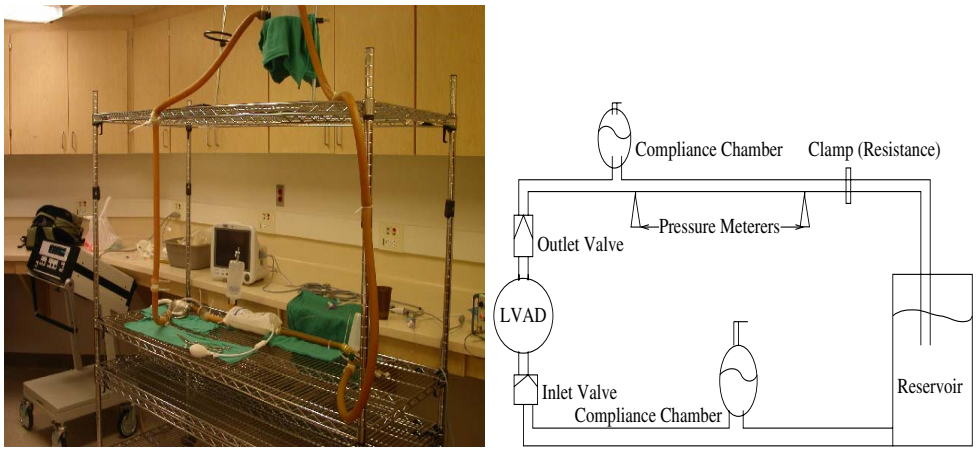


FIG. 7.1. *Flow loop at the Research Laboratory at the Texas Heart Institute (left), a sketch of the flow loop (right).*

One of the goals of this experiment was to recreate the pressure waves and fluid velocity at the middle section of the straight test segment similar to those typical

for the human abdominal aorta. To achieve this goal a clamp located downstream from the test segment was added to mimic downstream resistance by the capillary bed. Figure 7.2(left) shows the measured (filtered) pressure data at the inlet and at the outlet of the test segment. This compares well with the typical inlet and outlet pressure data of the human abdominal aorta, shown in Figure 7.2(right). Ultrasonic imaging and Doppler methods were used to measure the axial velocity of the flow. Nondairy coffee creamer was dispersed in water to enable reflection for ultrasound measurements. A high-frequency (20 MHz) single crystal probe was inserted through a catheter at several locations of the tube. This method has been validated in vivo by measuring the velocity and wall motion in mice to a precision of 0.1 um; see [25, 26].

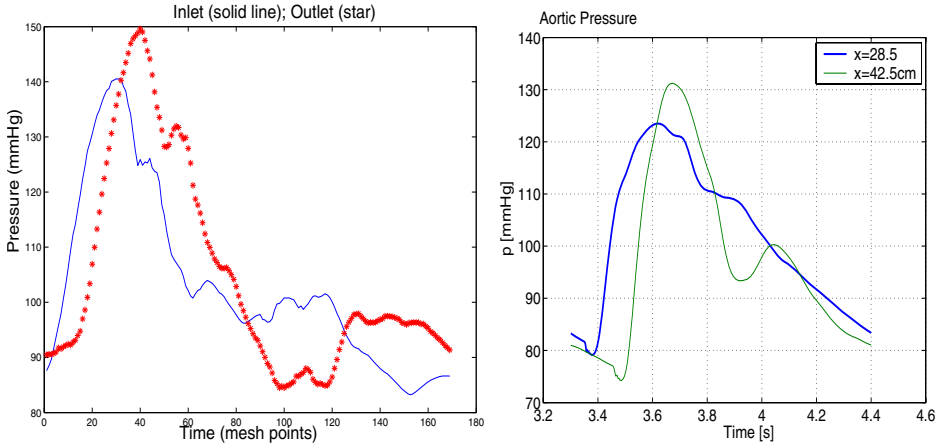


FIG. 7.2. *Inlet and outlet pressure data used in the numerical simulations. Left: Circulatory flow loop data (filtered). Right: aortic data* [14].

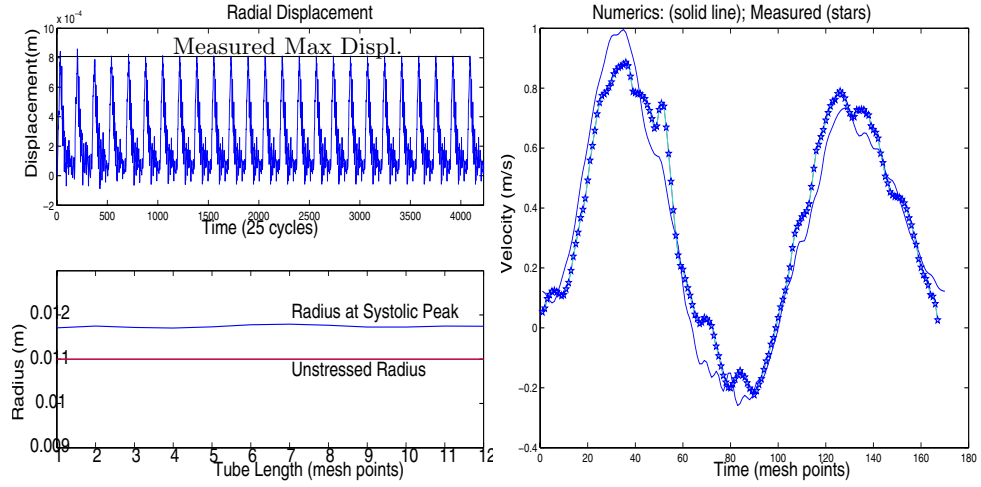


FIG. 7.3. *The figure shows a comparison between the numerical simulations and the experimental measurements. Left: Displacement. Right: axial component of the velocity, evaluated at the midpoint of the tube (midpoint with respect to the length and the diameter of the tube). The solid line corresponds to the numerically calculated solution; the line with asterisks corresponds to the experimentally collected data.*

To determine Young's modulus of the tube wall we measured the tube diameter d at the reference pressure of 84 mmHg ($d = 2.22$ cm) and at the maximal systolic pressure of 148 mmHg ($d = 2.38$ cm), utilizing the linear pressure-displacement relationship described by the equations of p^0 in (4.17) with $C_V = 0$ and the data for the tube wall thickness provided by the manufacturer of the latex tube, Kent Elastomer Products Inc. The value of $E = 1.0587 \times 10^6$ Pa was obtained. Using the numerical solver described in section 6 system of equations (5.1)–(5.6) was solved. The results were compared with the experimental measurements. At the top of Figure 7.3(left) is a comparison between the numerically calculated displacement and the experimentally measured maximal displacement of 0.0008 m. Figure 7.3(right) shows a comparison between the numerically calculated (solid line) and experimentally measured (asterisks) axial velocity. Excellent agreement was obtained indicating that this model captures well the fluid-structure interaction between a linearly elastic structure such as a latex tube, and the flow of a viscous incompressible fluid such as water, in the flow regime corresponding to the abdominal aorta.

8. Hysteresis behavior of viscoelastic arteries. In this section we compare the results of our viscoelastic model with the measurement of the viscoelastic properties of the human and canine arteries presented in [1, 2, 3]. In [1] Armentano et al. studied the viscoelastic aortic properties in dogs. In particular, they measured the magnitude of the viscous modulus corresponding to our coefficient hC_v/R. The values corresponding to dogs aortas, reported in [1], belong to the interval

$$\begin{aligned}\frac{hC_v}{R}|_{(\text{dog aorta})} &\in (3.8 \pm 1.3 \times 10^4, 7.8 \pm 1.1 \times 10^4)\ dyn \cdot \text{s/cm}^2 \\ &= (3.8 \pm 1.3 \times 10^3, 7.8 \pm 1.1 \times 10^3)\ \text{Pa} \cdot \text{s}.\end{aligned}$$

Taking into account the radius of the studied aortas (≈ 0.008 m) and the average wall thickness (≈ 0.0014 m), one obtains

$$C_v|_{(\text{dog aorta})} \in (2.17 \times 10^4, 4.45 \times 10^4)\ \text{Pa} \cdot \text{s}.$$

In [1] the measurements of the viscoelastic properties of the canine aorta were obtained, showing a hysteresis in the stress-strain diagram, where the stress (τ) and strain (e) were defined using

$$\tau = \frac{2p(r_e r_i)^2}{r_e^2 - r_i^2}\frac{1}{R^2}, \quad e = \frac{R+\eta}{R}. \tag{8.1}$$

Here r_e and r_i are the external and internal vessel radii calculated using $r_{e,i} = R \pm 0.5\ h$. The results of the measurements are shown in Figure 8.1(left). We used the data presented in [1] as a guide in the numerical simulation of the dynamics of the canine aorta utilizing the effective viscoelastic model (4.17), (4.26). Unfortunately, [1] does not include the pressure data at the inlet and outlet of the canine aorta. Thus, it was impossible to recreate the simulation that would correspond exactly to the scenario studied in [1]. However, using the data available to us, in particular the viscous modulus C_V, we were able to approximate the scenario studied in [1] and capture the main viscoelastic properties of the canine aorta. The results are shown in Figure 8.1. The top figures show the pressure and the scaled diameter in one cardiac cycle. Both waves exhibit the same morphology, but the diameter shows a time delay with respect to the pressure, which is due to the viscosity of the vessel wall. The

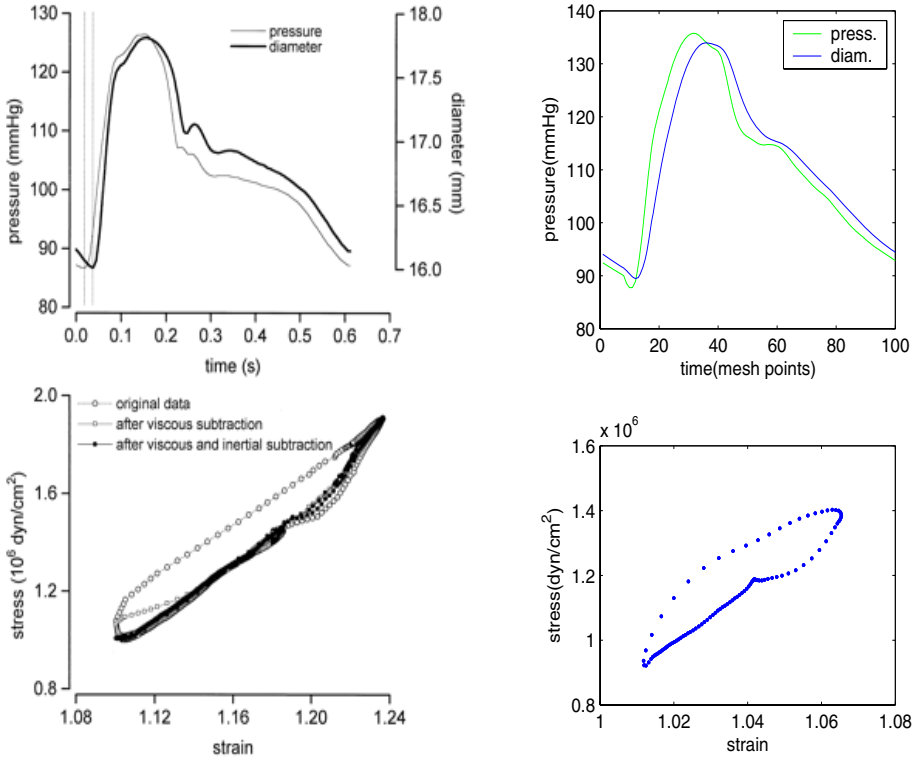


FIG. 8.1. *Left: Measured viscoelastic behavior of the canine aorta reported in* [1] *(top: aortic diameter and pressure wave forms, bottom: stress-strain relationship). Right: Numerical simulation of the reduced one-and-a-half-dimensional model showing viscoelastic behavior of vessel walls (top: aortic diameter and pressure wave forms, bottom: stress-strain relationship).*

bottom figures show the hysteresis behavior in the stress-strain relationship. The upper "half" of the hysteresis corresponds to the loading and the lower "half" to the unloading portion of the cardiac cycle. The hysteresis curves and the time-lag between the pressure and scaled diameter show similar qualitative behavior.

An even better approximation of the hysteresis behavior in the dynamics of major arteries was obtained for the data corresponding to a healthy human femoral artery. One reason for this is that the inlet and outlet pressure data that were used in all of our numerical simulations correspond to the human data. We compared our numerical simulations to the measurements data presented in [2]. In [2] Armentano et al. estimated the magnitude of the coefficient multiplying the term $\partial D/\partial t$, where D is the vessel diameter of a human femoral artery. The value of the coefficient was estimated to be $266 \times$ Pa $\cdot$ s/m. Using the values for the measured femoral artery diameter ($0.00625m$) and the wall thickness (0.001 m), one obtains

$$C_v|_{(\text{human femoral})} \approx 5.2 \times 10^3 \text{ Pa} \cdot \text{s}. \tag{8.2}$$

Thus, the corresponding viscous modulus hC_v/R is

$$\left.\frac{hC_v}{R}\right|_{(\text{human femoral})} \approx 1.6 \times 10^3 \text{ Pa} \cdot \text{s}, \tag{8.3}$$

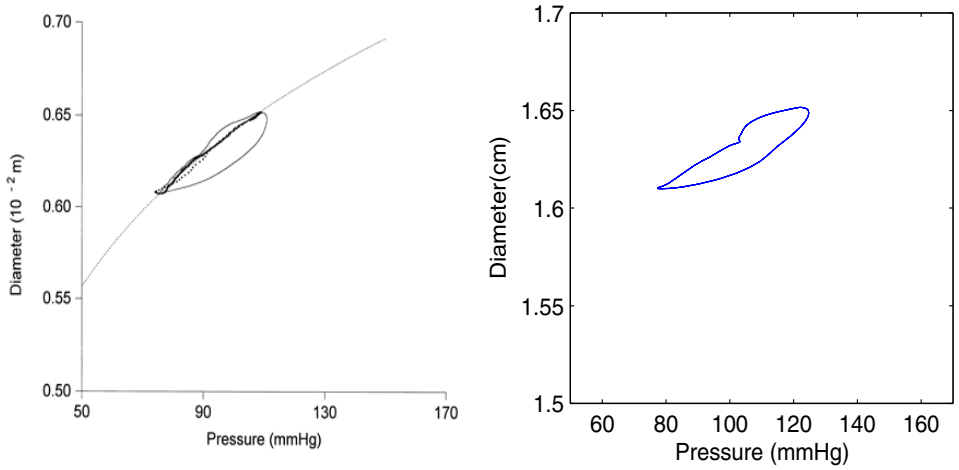


FIG. 8.2. *Left: Measurements of the diameter-pressure hysteresis loop in human femoral artery reported in* [2]. *Right: Numerical simulation of the diameter-pressure hysteresis loop with parameters from Table* 4.1 *($E = 1.3 \times 10^6$ Pa, $h = 0.001$ m, $R = 0.008$ m, $L = 0.13$ m, $hC_v/R = 10^3$ Pa · s).*

which is of the same order of magnitude as the viscous modulus corresponding to the dogs aortas. Figure 8.2 shows a comparison between our numerical simulations and measurements. There, a pressure-diameter relationship is plotted, showing hysteresis behavior. The graph in Figure 8.2(left) corresponds to the measurements of the human femoral artery reported in [2], and the graph in Figure 8.2(right) shows the pressure-diameter relationship in the simulations obtained using the reduced model (5.1), (5.6). Again, similar viscoelastic behavior is detected.

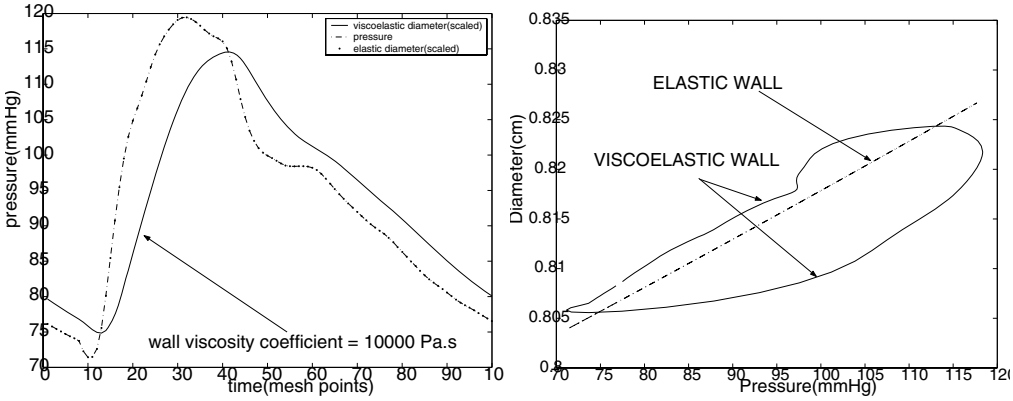


FIG. 9.1. *Elastic vs. viscoelastic wall model. The figure on the left shows the pressure and the scaled diameter (elastic and viscoelastic) over one cardiac cycle (the pressure wave and the elastic diameter coincide). The figure on the right shows the pressure-diameter plot for the viscoelastic model (hysteresis) and the elastic model (straight line).*

9. Elastic vs. viscoelastic model. We conclude this manuscript by presenting a comparison between the results of the fluid-structure interaction models assuming elastic vs. viscoelastic wall model with a relatively large viscoelastic constant $hC_v/R =$

10^4 Pa · s. Figure 9.1(left) shows the pressure and the scaled diameter values for the two models plotted over one cardiac cycle. One can easily detect the time-shift in the diameter of the viscoelastic model compared with the diameter of the elastic wall model which coincides (the scaled diameter) with the pressure wave. Figure 9.1(right) shows the pressure-diameter plot emphasizing the hysteresis in the viscoelastic model superimposed over the straight line pressure-diameter plot corresponding to the elastic model.

10. Conclusions. In this manuscript we derived a simple, effective closed model that describes blood flow through viscoelastic arteries in cylindrical geometry assuming axially symmetric flows. Using homogenization theory and asymptotic analysis, this fluid-structure interaction problem was reduced to a free-boundary problem of hyperbolic-parabolic type in two space dimensions. Although the model is two-dimensional, its simple form allows the use of one-dimensional solvers giving rise to a numerical algorithm of one-dimensional complexity. In contrast with the "classical" one-dimensional models where an ad hoc assumption on the axial velocity profile needs to be used to close the model, the system we obtained in this manuscript is closed, producing the axial as well as radial velocity as a solution of the problem. We showed that the reduced model approximates the original three-dimensional axially symmetric model to the ϵ^2-accuracy, where ϵ is the aspect ratio of the tube approximating straight arterial sections. The main novelty in this manuscript is the derivation of a viscoelastic cylindrical Koiter shell model to describe the behavior of arterial walls. Viscoelasticity of Kelvin–Voigt type was utilized to derive the model which appoximates well the hysteresis behavior observed in the vessel wall measurements. We showed that in this fluid-structure interaction model bending rigidity of arterial walls plays a nonnegligible role in the leading-order approximation of the problem. This effect, together with the viscosity of vessel walls, explicitly derived in this manuscript, provides the regularizing mechanisms for the stability of the solutions.

REFERENCES

[1] R. L. Armentano, J. G. Barra, J. Levenson, A. Simon, and R. H. Pichel, *Arterial wall mechanics in conscious dogs: Assessment of viscous, inertial, and elastic moduli to characterize aortic wall behavior*, Circ. Res., 76 (1995), pp. 468–478.

[2] R. L. Armentano, J. L. Megnien, A. Simon, F. Bellenfant, J. G. Barra, and J. Levenson, *Effects of hypertension on viscoelasticity of carotid and femoral arteries in humans*, Hypertension, 26 (1995), pp. 48–54.

[3] R. D. Bauer, R. Busse, A. Shabert, Y. Summa, and E. Wetterer, *Separate determination of the pulsatile elastic and viscous forces developed in the arterial wall in vivo*, Pflugers Arch., 380 (1979), pp. 221–226.

[4] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North–Holland, Amsterdam, 1978.

[5] S. Čanić and E.-H. Kim, *Mathematical analysis of the quasilinear effects in a hyperbolic model of blood flow through compliant axi–symmetric vessels*, Math. Methods Appl. Sci., 26 (2003), pp. 1161–1186.

[6] S. Čanić, A. Mikelić, and J. Tambača, *A two-dimensional effective model describing fluid-structure interaction in blood flow: Analysis, simulation and experimental validation*, Comptes Rendus Méchanique Acad. Sci. Paris, 333 (2005), pp. 867–883.

[7] S. Čanić and A. Mikelić, *Effective equations describing the flow of a viscous incompressible fluid through a long elastic tube*, Comptes Rendus Méchanique Acad. Sci. Paris, 330 (2002), pp. 661–666.

[8] S. Čanić and A. Mikelić, *Effective equations modeling the flow of a viscous incompressible fluid through a long elastic tube arising in the study of blood flow through small arteries*, SIAM J. Appl. Dyn. Syst., 2 (2003), pp. 431–463.

[9] S. Čanić, D. Lamponi, A. Mikelić, and J. Tambača, *Self-consistent effective equations modeling blood flow in medium-to-large compliant arteries*, Multiscale Model. Simul., 3 (2005), pp. 559–596.
[10] A. Chambolle, B. Desjardins, M. Esteban, and C. Grandmont, *Existence of weak solutions for an unsteady fluid-plate interaction problem*, J Math. Fluid Mech., 7 (2005), pp. 368–404.
[11] P. G. Ciarlet, *Mathematical Elasticity. Vol.* III. *Theory of Shells*, Stud. Math. Appl. 29, North–Holland, Amsterdam, 2000.
[12] P. G. Ciarlet and V. Lods, *Asymptotic analysis of linearly elastic shells.* III. *Justification of Koiter's shell equations*, Arch. Rational Mech. Anal., 136 (1996), pp. 191–200.
[13] P. G. Ciarlet and V. Lods, *Asymptotic analysis of linearly elastic shells. Generalized membrane shells*, J. Elasticity, 43 (1996), pp. 147–188.
[14] C. Chmielevsky, *A Comparison of Two Models Predicting Blood Flow in the Systemic Arteries*, M.S. thesis, North Carolina State University, Raleigh, NC, 2004.
[15] D. Coutand and S. Shkoller, *Motion of an elastic solid inside of an incompressible viscous fluid*, Arch. Ration. Mech. Anal., 176 (2005), pp. 25–102.
[16] D. Coutand and S. Shkoller, *Interaction between quasilinear elasticity and the Navier-Stokes equations*, Arch. Ration. Mech. Anal., 179 (2006), pp. 303–352.
[17] H. Demiray, *Small but finite amplitude waves in a prestressed viscoelastic thin tube filled with an inviscid fluid*, Internat. J. Engrg. Sci., 35 (1997), pp. 353–363.
[18] S. X. Deng, J. Tomioka, J. C. Debes, and Y. C. Fung, *New experiments on shear modulus of elasticity of arteries*, Am. J. Physiol. Heart Circ. Physiol., 266 (1994), pp. 1–10.
[19] P. Destuynder, *A classification of thin shell theories*, Acta Appl. Math., 4 (1985), pp. 15–63.
[20] A. C. Eringen, *Mechanics of Continua*, Wiley, New York, 1967.
[21] Y. C. Fung, *Biomechanics: Circulation*, 2nd ed., Springer, New York, 1993.
[22] Y. C. Fung, *Biomechanics: Mechanical Properties of Living Tissues*, Springer, New York, 1993.
[23] J. B. Grotberg and O. E. Jensen, *Biofluid mechanics in flexible tubes*, in Annual Review of Fluid Mechanics, Annu. Rev. Fluid Mech. 36, Annual Reviews, Palo Alto, CA, 2004, pp. 121–147.
[24] M. Guidorzi, M. Padula, and P. Plotnikov, *Hopf solutions to a fluid-elastic interaction model*, Math. Models Methods Appl. Sci., submitted.
[25] C. J. Hartley, *Ultrasonic blood flow and velocimetry*, in McDonald's Blood Flow in Arteries, Theoretical, Experimental and Clinical Principles, 4th ed., W. W. Nichols and M. F. O'Rourke, Arnold, London, 1998, pp. 154–169.
[26] C. J. Hartley, G. Taffet, A. Reddy, M. Entman, and L. Michael, *Noninvasive cardiovascular phenotyping in mice*, ILAR J., 43 (2002), pp. 147–158.
[27] M. Heil and T. J. Pedley, *Large axisymmetric deformations of a cylindrical shell conveying a viscous flow*, J. Fluids and Struct., 9 (1995), pp. 237–256.
[28] U. Hornung, ed., *Homogenization and Porous Media*, Interdiscip. Appl. Math. 6, Springer, New York, 1997.
[29] J. D. Humphrey, *Mechanics of the arterial wall: Review and directions*, Crit. Rev. Biomed. Eng., 23 (1995), pp. 1–162.
[30] R. E. Klabunde, *Cardiovascular Physiology Concepts*, http://cvphysiology.com/index.html (2005).
[31] W. T. Koiter, *A consistent first approximation in the general theory of thin elastic shells*, in Proceedings of the IUTAM Symposium on the Theory of Thin Elastic Shells (Delft, 1959), North–Holland, Amsterdam, 1960, pp. 12–33.
[32] W. T. Koiter, *On the foundations of the linear theory of thin elastic shells.* I, II, Nederl. Akad. Wetensch. Proc. Ser. B, 73 (1970), pp. 169–182.
[33] G. D. C. Kuiken, *Wave propagation in a thin-walled liquid-filled initially stressed tube*, J. Fluid Mech., 141 (1984), pp. 289–308.
[34] P. Luchini, M. Lupo, and A. Pozzi, *Unsteady Stokes flow in a distensible pipe*, Z. Angew. Math. Mech., 71 (1991), pp. 367–378.
[35] X. Ma, G. C. Lee, and S. G. Lu, *Numerical simulation for the Propagation of nonlinear pulsatile waves in arteries*, ASME J. Biomech. Eng., 114 (1992), pp. 490–496.
[36] E. Marušić-Paloka and A. Mikelić, *The derivation of a nonlinear filtration law including the inertia effects via homogenization*, Nonlinear Anal., 42 (2000), pp. 97–137.
[37] A. Mikelić, *Homogenization theory and applications to filtration through porous media*, in Filtration in Porous Media and Industrial Applications, Lecture Notes in Math. 1734, M. Espedal, A. Fasano, and A. Mikelić, eds., Springer, Berlin, 2000, pp. 127–214.
[38] A. Mikelić, G. Guidoboni, and S. Čanić, *Fluid-Structure Interaction between a Vis-*

cous Incompressible Fluid and a Linearly Elastic Cylinder with Finite Wall Thickness, manuscript.

[39] W. W. Nichols and M. F. O'Rourke, *McDonald's Blood Flow in Arteries: Theoretical, Experimental and Clinical Principles*, 4th ed., Arnold, Oxford University Press, New York, London, Sydney, Auckland, 1997.

[40] F. Nobile, *Numerical Approximation of Fluid-Structure Interaction Problems with Application to Haemodynamics*, Ph.D. thesis, EPFL, Lausanne, Switzerland, 2001.

[41] V. Novozhilov, *Thin Shell Theory*, P. G. Lowe, ed., Groningen, Noordholf, The Netherlands 1964.

[42] M. S. Olufsen, C. S. Peskin, W. Y. Kim, E. M. Pedersen, A. Nadim, and J. Larsen, *Numerical simulation and experimental validation of blood flow in arteries with structured-tree outflow conditions*, Ann. Biomed. Eng., 28 (2000), pp. 1281–1299.

[43] G. Pontrelli, *A mathematical model of flow in a liquid-filled visco-elastic tube*, Med. Biol. Eng. Comput., 40 (2002), pp. 550–556.

[44] A. Quarteroni, M. Tuveri, and A. Veneziani, *Computational vascular fluid dynamics: Problems, models and methods. Survey article*, Comput. Vis. Sci., 2 (2000), pp. 163–197.

[45] H. Reismann, *Elastic Plates: Theory and Applications*, Wiley, New York, 1988.

[46] A. M. Robertson and A Sequeira, *A director theory approach to modeling blood flow in the arterial system: An alternative to classical* 1*D models*, Math. Models Methods Appl. Sci., 15 (2005), pp. 871–906.

[47] J. Tambača, S. Čanić, and A. Mikelić, *Effective model of the fluid flow through elastic tube with variable radius*, in XI. Mathematikertreffen Zagreb-Graz, Grazer Math. Ber. 348, Karl-Franzens-University, Graz, Austria, 2005, pp. 91–112.

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

[49] H. B. da Veiga, *On the existence of strong solution to a coupled fluid structure evolution problem*, J. Math. Fluid Mech., 6 (2004), pp. 21–52.

[50] E. Ventsel and T. Krauthammer, *Thin Plates and Shells*, Marcel Dekker, New York, 2001.

[51] R. P. Vito and S. A. Dixon, *Blood vessels constitutive models* 1995-2002, Ann. Rev. Biomed. Eng., 5 (2003), pp. 413–439.

[52] A. Wempner, *Mechanics of Solids and Shells: Theories and Approximations*, CRC Press, Boca Raton, FL, 2003.

SIAM J. APPL. MATH.
Vol. 67, No. 1, pp. 194–213

INVERSE BOUNDS AND BULK PROPERTIES OF COMPLEX-VALUED TWO-COMPONENT COMPOSITES*

CHRISTIAN ENGSTRÖM†

Abstract. The bulk properties of composites are known to depend strongly on the microstructure. This dependence can be quantified in terms of a representation introduced by D. Bergman, which factorizes the geometry dependence from the contrast. Based on this analytic representation of the effective permittivity, we present a general scheme to estimate the microstructural parameters such as the volume fraction and the anisotropy of two-component composites. The estimates are given as bounds, that is, the largest parameter region which is compatible with the available information. Thus, more information produces better estimates on the microstructural parameters. The method, which uses complex-valued measurements of bulk properties of the composite, is illustrated by numerical examples.

Key words. composite, inverse bounds, inverse homogenization, Stieltjes integral, Padé approximations

AMS subject classifications. 78A48, 41A20, 41A21, 30E05, 30E10

DOI. 10.1137/060649598

1. Introduction.

In many cases of interest when considering the interaction of electromagnetic waves with composites the wavelength is much longer than the characteristic length of the microstructure. The composite then reacts to the slowly varying field in much the same way as a homogeneous material, with some effective material parameters.

The determination of the effective properties of composite materials, with known periodic geometry or from simulations of random materials, constitutes a classical problem in physics. In the case of a two-component mixture, a representation formula that separate the dependence on the phases and the dependence on the microstructure was developed by Bergman [6] and Golden and Papanicolaou [22].

The structural information is associated with a spectral measure, and much effort has been focused on the reconstruction of this measure from a known geometry [24, 19, 29]. When the measure is calculated, a single integral gives the effective property for any value of the phases. One drawback is that a complete knowledge of the geometry rarely is available.

A direct approach to characterize the microstructure is in terms of an infinite set of correlation functions [4, 37]. Except for some special cases, the infinite set of correlation functions is not known, and hence an exact solution is not possible. Using images of cross sections, some correlation functions can be estimated. When the material is finely scaled, the computation of the volume fraction is a large computational problem, and calculations of higher-order correlation functions is in general very demanding.

Instead of using correlation functions, information from measurements of one effective property can be used to improve bounds on a related property. Prager [35] used measurements of the effective magnetic permeability to improve the bounds on

*Received by the editors January 11, 2006; accepted for publication (in revised form) September 18, 2006; published electronically November 22, 2006.
http://www.siam.org/journals/siap/67-1/64959.html
†Department of Electroscience, Lund Institute of Technology, P.O. Box 118, 221 00 Lund, Sweden (christian.engstrom@es.lth.se).

the thermal conductivity. These bounds are called cross-property bounds or coupled bounds. The pioneering work of Prager was followed by the papers of Bergman [5, 6] and Milton [33], among others. The problem of bounding the structural parameters that characterize the microstructure from known values of an effective property is by some authors called inverse homogenization, and the bounds are called inverse bounds.

Inverse bounds for the volume fraction were first derived in [31]. In recent years the representation formula introduced by Bergman [6] has been used to study the inverse problem. Explicit formulas for bounds on the volume fraction can in the case of measurements of lossy materials be found in [14]. If the measurements are on a real-valued effective property, the formulas for the volume fraction in [14] cannot be used. In the case of real-valued measurements the author in [21] provides a schedule to derive inverse bounds and give explicit formulas for bounds on the three lowest moments of the measure, where the first moment corresponds to the volume fraction.

Various inverse algorithms for recovering the structural parameters (the spectral measure) of composites from experimental data have been developed [15, 17, 13]. In [18] the algorithm developed in [17] was successfully used to recover the measure from 4000 reflectance data points.

The numerical algorithms are useful, but one disadvantage with this approach is that we lose the concept of bounds. If we have limited information from measurements (few or inaccurate measurements), the numerical methods cannot recover the measure. Using the numerical approximations of the measure can then result in bounds on an effective property that are not valid.

In this paper inverse bounds using information from measurements of lossy materials are derived. These bounds are used to derive cross-property bounds, which are exemplified by a frequency-dependent permittivity. We use and improve the geometry-independent bounds on the structural parameters that were derived in [21]. In other words, restrictions on the moments of the measure are derived.

The asymptotic behavior of the formulas in this paper is superior to the formulas in [21], but the formulas presented here cannot be used if the effective property is real-valued. The two papers complement each other, and the formulas in the two papers can be combined.

2. Bounds on the effective permittivity. Assume that inside the composite the electric field $\boldsymbol{E}$ and the electric flux density $\boldsymbol{D}$ satisfy the constitutive relation

$$\boldsymbol{D}(\boldsymbol{x}) = \boldsymbol{\epsilon}(\boldsymbol{x})\boldsymbol{E}(\boldsymbol{x}). \tag{2.1}$$

The permittivity matrix $\boldsymbol{\epsilon}$ is the description of the material on the fine scale, where $\boldsymbol{\epsilon}$ and thereby the fields oscillate rapidly. On a much larger scale the averaged fields have no oscillations on the length scale of the microstructure, since they are smoothed out, but they retain slow macroscopic variations.

We seek an effective permittivity matrix $\boldsymbol{\epsilon}^{\text{eff}}$ which relates the average of the electric displacement field $\langle \boldsymbol{D} \rangle$ to the average of the electric field $\langle \boldsymbol{E} \rangle$. The average is over a volume having large size compared with the microstructure.

In general the $\boldsymbol{D}$-field satisfies $\nabla \cdot \boldsymbol{D} = \rho$, where ρ is the charge density. Using for example a two-scale expansion [2, p. 138] of Maxwell's equations, we have $\nabla \times \boldsymbol{E} = 0$.

From the constitutive relation (2.1) it follows that, for a charge-free region, the $\boldsymbol{E}$-field satisfies

$$\nabla \times \boldsymbol{E} = 0, \quad \nabla \cdot (\boldsymbol{\epsilon}\boldsymbol{E}) = 0. \tag{2.2}$$

This system represents, besides dielectrics, several other physical phenomena such as electrical and thermal conductivity, magnetism, diffusion, and flow in porous media.

Let $\langle \boldsymbol{\Psi} \rangle$ denote the average of the vector field $\boldsymbol{\Psi}$ over the unit cell $\mathrm{U} = [0,1[^d$ in d dimensions. If the $\boldsymbol{E}$-field is Lebesgue integrable and the equations (2.2) are satisfied in a weak sense, the homogenization rule

$$\langle \boldsymbol{\epsilon} \boldsymbol{E} \rangle = \boldsymbol{\epsilon}^{\mathrm{eff}} \langle \boldsymbol{E} \rangle \tag{2.3}$$

can be proven [25, p. 15].

The materials in this paper are assumed to be d-dimensional and to consist of two homogeneous, isotropic phases. The two-component material is locally modelled by the scalar relative permittivity

$$\epsilon(\epsilon_1, \epsilon_2) = \epsilon_1 \chi_1(\boldsymbol{x}) + \epsilon_2 \chi_2(\boldsymbol{x}), \tag{2.4}$$

where the components are isotropic with constant permittivity ϵ_1 and ϵ_2. We use complex-valued permittivities and assume that the imaginary parts are greater than or equal to zero.

The volume fraction of phase χ_i is denoted f_i, and the characteristic function χ_i is defined as

$$\chi_i(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \text{ in phase } i, \\ 0 & \text{otherwise} \end{cases}$$

and $f_1 + f_2 = 1$. When the composite is periodic and the characteristic function χ_1 is known, we can calculate $\boldsymbol{\epsilon}^{\mathrm{eff}}$ from (2.2), (2.3) using a standard finite element program, but in many cases the geometry is unknown. Another drawback with this approach is that the problem (2.2), (2.3) depends not only on the microstructure but also on the contrast. If we change the contrast, all calculations need to be repeated.

2.1. Analytic representation of the effective matrix. Due to the homogeneity property $\epsilon^{\mathrm{eff}}(c\epsilon_1, c\epsilon_2) = c\epsilon^{\mathrm{eff}}(\epsilon_1, \epsilon_2)$, the effective permittivity depends on the ratio ϵ_1/ϵ_2. The main property of the solution to the problem in (2.2) and (2.3) is that the function

$$\frac{\boldsymbol{\epsilon}^{\mathrm{eff}}(\epsilon_1, \epsilon_2)}{\epsilon_2} = \boldsymbol{\epsilon}^{\mathrm{eff}} \left(\frac{\epsilon_1}{\epsilon_2}, 1 \right) \tag{2.5}$$

is analytic in $\epsilon_1/\epsilon_2 \in \mathbb{C} \backslash]-\infty, 0]$ and that it maps the upper half-plane to the upper half-plane; i.e., the function $\boldsymbol{\epsilon}^{\mathrm{eff}}/\epsilon_2$ is a Herglotz function [1]. The function $\boldsymbol{\epsilon}^{\mathrm{eff}}/\epsilon_2$ has the Stieltjes-integral representation

$$\boldsymbol{\epsilon}^{\mathrm{eff}}(\epsilon_1, \epsilon_2) = \epsilon_2 \mathbf{I} - \epsilon_2 \mathbf{G}(s), \tag{2.6}$$

where

$$\mathbf{G}(s) = \int_0^1 \frac{\mathrm{d}\mathbf{m}(y)}{s - y}, \quad s = \frac{\epsilon_2}{\epsilon_2 - \epsilon_1}. \tag{2.7}$$

The matrix-valued measure $\mathbf{m}$ on $[0,1]$ is derived from the spectral measure of the operator $\Gamma = P\chi_1$, where $P = \nabla(-\Delta)^{-1}(\nabla\cdot)$. The operator Γ is bounded $||\Gamma|| \leq 1$ and self-adjoint in $L^2(\mathrm{U})^d$ equipped with the scalar product $(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) = \langle \chi_1 \boldsymbol{\Psi}_1 \cdot \boldsymbol{\Psi}_2 \rangle$ [22].

The representation formula (2.6), valid for $s \notin [0,1]$, was derived for the periodic case in [6] and in the general case in [22].

The measure $\mathbf{m}$ is a purely geometric quantity. It depends on the microstructure but not on the value of the two phases. If the microstructure is the same, the single integral (2.7) gives the effective permittivity, independent of the value of the phases. This is particularly useful when the permittivity is frequency- or temperature-dependent.

2.2. Bounds on $\boldsymbol{\epsilon}^{\mathbf{eff}}$ using Padé approximations. If the microstructure is only partly known, we can get bounds on the effective permittivities. When the permittivities of the two materials, together with the volume fraction f_1, are known, the effective permittivity is bounded by the harmonic and arithmetic means. If more structural information is known, we get tighter bounds such as the Hashin–Shtrikman bounds and the Beran bounds.

We focus on the diagonal elements in the effective permittivity matrix and use the power series expansion

$$\boldsymbol{\epsilon}^{\mathrm{eff}} = \epsilon_2 \mathbf{F}(z), \quad \mathbf{F}(z) = \sum_{n=0}^{\infty} \mathbf{c}_n z^n, \tag{2.8}$$

where $z = -1/s = (\epsilon_1 - \epsilon_2)/\epsilon_2$ is the contrast. The series (2.8) is convergent in $|z| < 1$.

The integral (2.7) vanishes in the limit $s \to \infty$, implying $\mathbf{c}_0 = \mathbf{I}$. This is a consequence of (2.8), because $z = 0$ when $\epsilon_1 = \epsilon_2$, which means that we have only one material.

For $|s| > 1$ the function $(s-y)^{-1}$ has a power expansion in y/s. The integral $\mathbf{G}(s)$ then has the power expansion

$$\mathbf{G}(s) = \sum_{n=0}^{\infty} \frac{1}{s^{n+1}} \int_0^1 y^n \, \mathrm{d}\mathbf{m}(y). \tag{2.9}$$

The integral in this expression is, for $n = 0, 1, \ldots$, the (Hausdorff) moments of the measure $\mathbf{m}$. The coefficients $\mathbf{c}_n$ in the power series expansion (2.8) and the measure $\mathbf{m}$ are connected by the moments

$$\mathbf{c}_{n+1} = (-1)^n \int_0^1 y^n \, \mathrm{d}\mathbf{m}(y). \tag{2.10}$$

Since the measure $\mathbf{m}$ is defined on the compact set $[0,1]$ it follows that $\mathbf{m}$ is bounded and uniquely determined by the moments [1]. If all the moments are known, the effective matrix is obtained from the series (2.8). Thus, the local information about $\epsilon_1 = \epsilon_2$ gives the effective permittivity independent of the contrast.

The volume fraction f_1 is given by the total weight [6, 22]

$$\mathbf{c}_1 = \int_0^1 \mathrm{d}\mathbf{m}(y) = f_1 \mathbf{I}. \tag{2.11}$$

Higher-order moments depend on the geometrical structure. Bergman [6] derived the general constraint $\operatorname{Tr} \mathbf{c}_2 = -c_1(1-c_1)$ and that, in the case of a statistically isotropic composite, the second moment is

$$\mathbf{c}_2 = -\int_0^1 y \, \mathrm{d}\mathbf{m}(y) = -\frac{c_1(1-c_1)}{d}\mathbf{I}. \tag{2.12}$$

Higher-order moments can be calculated exactly in a few special cases; see, for instance, [16] or [19].

The power series (2.8) with coefficients given by the moments (2.10) defines a series of Stieltjes. Series of Stieltjes have known upper and lower bounds in the form of continued fractions or Padé approximations [1]. We use Padé approximations of the power series (2.8).

Let ϵ^{eff} be one of the diagonal elements in the matrix $\boldsymbol{\epsilon}^{\text{eff}} = \epsilon_2\mathbf{F}(z)$. The $\epsilon_{p,q}$ Padé approximant to ϵ^{eff} is defined by the equation

$$\epsilon^{\text{eff}}(z)Q(z) - P(z) = \mathcal{O}(z^{p+q+1}), \tag{2.13}$$

where P and Q are polynomials of degree at most p and q, respectively [1]. This equation gives us an approximation of the effective permittivity by the rational function

$$\epsilon_{p,q} = \frac{P(z)}{Q(z)} = \frac{a_0 + \cdots + a_p z^p}{1 + b_1 z + \cdots + b_q z^q}. \tag{2.14}$$

When $\epsilon_2 > \epsilon_1$ and $N \geq 1$, the N-point upper bounds $\boldsymbol{\epsilon}_N^{\text{U}}$ are obtained by forming the approximations

$$\boldsymbol{\epsilon}_{2M+1}^{\text{U}} = \epsilon_2\boldsymbol{\epsilon}_{M+1,M}(\mathbf{F}), \quad \boldsymbol{\epsilon}_{2M}^{\text{U}} = \epsilon_2\boldsymbol{\epsilon}_{M,M}(\mathbf{F}). \tag{2.15}$$

The inverse of the matrix $\boldsymbol{\epsilon}^{\text{eff}}(\epsilon_1/\epsilon_2, 1)$ is analytic in $\epsilon_1/\epsilon_2 \in \mathbb{C}\backslash]-\infty, 0]$. The analyticity implies that it has a power series expansion in z. Lower bounds on $\boldsymbol{\epsilon}^{\text{eff}}$ are given from Padé approximations of the series

$$\left(\frac{\boldsymbol{\epsilon}^{\text{eff}}}{\epsilon_1}\right)^{-1} = \tilde{\mathbf{F}}(z), \quad \text{where} \quad \tilde{\mathbf{F}}(z) = \sum_{n=0}^{\infty} \tilde{\mathbf{c}}_n z^n. \tag{2.16}$$

The coefficients $\mathbf{c}_n$ and $\tilde{\mathbf{c}}_n$ in the two series are related according to

$$\tilde{\mathbf{c}}_0 = \mathbf{I}, \quad \tilde{\mathbf{c}}_1 = (1 - c_1)\mathbf{I}, \quad \tilde{\mathbf{c}}_n = -\sum_{k=0}^{n-1} \tilde{\mathbf{c}}_k \mathbf{c}_{n-k}. \tag{2.17}$$

The coefficient c_1 is the volume fraction of phase one (2.11) and $\tilde{c}_1$ is the volume fraction of phase two. The N-point lower bounds $\boldsymbol{\epsilon}_N^{\text{L}}$, when $\epsilon_2 > \epsilon_1$ and $N \geq 1$, are obtained from

$$\boldsymbol{\epsilon}_{2M+1}^{\text{L}} = \epsilon_1[\boldsymbol{\epsilon}_{M+1,M}(\tilde{\mathbf{F}})]^{-1}, \quad \boldsymbol{\epsilon}_{2M}^{\text{L}} = \epsilon_1[\boldsymbol{\epsilon}_{M,M}(\tilde{\mathbf{F}})]^{-1}. \tag{2.18}$$

For example the $\boldsymbol{\epsilon}_{1,0}$ Padé approximant of the expansion (2.16) is the harmonic mean

$$\boldsymbol{\epsilon}_1^{\text{L}} = \frac{\epsilon_1}{1 + \tilde{c}_1 z}\mathbf{I} = \left(\frac{f_1}{\epsilon_1} + \frac{f_2}{\epsilon_2}\right)^{-1}\mathbf{I} \tag{2.19}$$

and the $\boldsymbol{\epsilon}_{1,0}$ Padé approximant of (2.8) gives the arithmetic mean

$$\boldsymbol{\epsilon}_1^{\text{U}} = (\epsilon_2 + c_1\epsilon_2 z)\mathbf{I} = (f_1\epsilon_1 + f_2\epsilon_2)\mathbf{I}. \tag{2.20}$$

Wiener [41] first derived these bounds on an effective material parameter. In the same way the $\boldsymbol{\epsilon}_{1,1}$ Padé approximant of the expansion (2.16) is the lower bound

$$\boldsymbol{\epsilon}_2^{\text{L}} = \epsilon_1[\tilde{c}_1\mathbf{I} - \tilde{\mathbf{c}}_2 z][\tilde{c}_1\mathbf{I} - \tilde{\mathbf{c}}_2 z + \tilde{c}_1^2 z\mathbf{I}]^{-1}, \tag{2.21}$$

where $\tilde{\mathbf{c}}_2 = -\mathbf{c}_2 - c_1\tilde{c}_1\mathbf{I}$. The $\boldsymbol{\epsilon}_{1,1}$ Padé approximant of (2.8) gives the upper bound

$$\boldsymbol{\epsilon}_2^{\mathrm{U}} = \epsilon_2[c_1\mathbf{I} - \mathbf{c}_2 z + c_1^2 z\mathbf{I}][c_1\mathbf{I} - \mathbf{c}_2 z]^{-1}. \tag{2.22}$$

These bounds were first derived in [33]; see also [28, 39].

In the isotropic case, $\mathbf{c}_2 = -(c_1\tilde{c}_1/d)\mathbf{I}$, the two-point bounds (2.21) and (2.22) are equivalent to the Hashin–Shtrikman bounds [23], and the bounds $\boldsymbol{\epsilon}_3^{\mathrm{L}}, \boldsymbol{\epsilon}_3^{\mathrm{U}}$ reduce to the Beran bounds [3, 38]. The Padé approximations give a hierarchy of bounds that become progressively narrower as more structural information is used [34, 1, 40, 11]. The bounds (2.21) and (2.22) are optimal, since they are attained for a variety of geometries [7, 32]. In general, the bounds on the effective permittivity (2.15) and (2.18) can be improved by incorporating phase exchange relations. Milton [32] first exploited the phase exchange equality in two dimensions [26] for derivation of bounds, and Bergman [9, 10] first used the phase exchange inequality in three dimensions [36] to improve bounds on the effective permittivity.

2.3. Complex bounds on the permittivity. Let c_n be one of the diagonal elements in $\mathbf{c}_n$. In the general case when the values of the phases are complex, the real segment $l = \{c_n;\, c_n^{\min} \le c_n \le c_n^{\max}\}$ is for fixed values on $c_1, c_2, \ldots, c_{n-1}$ mapped by $\epsilon_n^{\mathrm{L}}(c_n)$ and $\epsilon_n^{\mathrm{U}}(c_n)$ on a circle or a line segment.

The minimum $c_n^{\min}$ and the maximum $c_n^{\max}$ are functions of the lower-order parameters $c_1, c_2, \ldots, c_{n-1}$. The extreme values can be determined by varying the c_n parameter in the n-point bounds and using that the n-point bounds are forbidden to violate the $(n-1)$-point bounds. This procedure was used in [21].

For example, we get complex bounds from the lens-shaped region bounded by

$$\epsilon_2^{\mathrm{L}}(\tilde{c}_2; \epsilon_1, \epsilon_2, \tilde{c}_1), \quad \epsilon_2^{\mathrm{U}}(c_2; \epsilon_1, \epsilon_2, c_1) \tag{2.23}$$

with the structural parameter $\tilde{c}_2$ and c_2 varying between

$$c_2^{\min} = -c_1(1 - c_1), \quad c_2^{\max} = 0. \tag{2.24}$$

Alternatively, we can describe the bounds $\epsilon_n^{\mathrm{L}}(c_n)$ and $\epsilon_n^{\mathrm{U}}(c_n)$ in terms of the points through which the circle passes [8, 33]. Let $\mathrm{Arc}(z_0, z_1, z_2)$ denote the arc of a circle joining the points z_0 and z_1 that when extended passes through z_2. For example, the effective permittivity ϵ^{eff} is in the complex case bounded by the intersection of the circles

$$\mathrm{Arc}(\epsilon_1, \epsilon_1^{\mathrm{L}}, \epsilon_1^{\mathrm{U}}), \quad \mathrm{Arc}(\epsilon_2, \epsilon_1^{\mathrm{L}}, \epsilon_1^{\mathrm{U}}). \tag{2.25}$$

We have $\epsilon_2^{\mathrm{L}} \to \epsilon_1$ and $\epsilon_2^{\mathrm{U}} \to \epsilon_2$ when $c_2 \to -\infty$. It follows that in terms of the structural parameters c_2, the circles are described by

$$\mathrm{Arc}(\epsilon_2^{\mathrm{L}}(-\infty), \epsilon_2^{\mathrm{L}}(c_2^{\min}), \epsilon_2^{\mathrm{L}}(c_2^{\max})), \quad \mathrm{Arc}(\epsilon_2^{\mathrm{U}}(-\infty), \epsilon_2^{\mathrm{U}}(c_2^{\min}), \epsilon_2^{\mathrm{U}}(c_2^{\max})). \tag{2.26}$$

The arcs (2.25) or (2.26), defining the points through which the circles pass, provide a geometrical characterization of the bounds. The alternative representation of the arcs (2.23) gives, in terms of c_2, directly a parameterization of the lens-shaped boundary.

3. Inverse bounds and bulk properties. The task in inverse homogenization is to calculate the structural parameters $\mathbf{c}_n$, or equally, the measure $\mathbf{m}$, given information from experiments.

When only measured values of the effective permittivity are known, the moments cannot be determined. Given a finite number of measurements, there exist in general several geometries that give the same $\boldsymbol{\epsilon}^{\text{eff}}$. Moreover, any measurement contains noise, which limits the accuracy.

The measurements can be on one effective property of the material at different temperatures or in a range of frequencies. It is also possible to get information from measurements of several related parameters such as the permittivity, the permeability, and the thermal conductivity. The important thing is that the microstructure is the same.

Bounds on the volume fraction c_1, using information from measurements, were derived in [31, 14, 21]. In [14], the authors derived bounds on the volume fraction that are valid in the general anisotropic case and tighter bounds on the volume fraction when the material is statistically isotropic.

We focus on the diagonal elements in $\mathbf{c}_n$ and provide a method to derive bounds on any diagonal element c_n. Moreover, we give examples where c_1, c_2, and c_3 are bounded, using information from measurements. We assume that the measurements are on the effective permittivity ϵ^{eff} at different frequencies $\omega_0, \omega_1, \ldots, \omega_n$, although the measurements could very well pertain to several other physical parameters associated with the same microstructure [34].

The bounds on the structural parameters c_n give geometrical information about the composite, but in many cases the composite's effective bulk properties as a function of frequency or temperature is what is desired. The bounds on the structural parameters imply cross-property bounds on the effective properties, which gives bounds on the effective permittivity at all frequencies where the homogenization theory is valid.

3.1. Geometry-independent inverse bounds. The volume fraction $f_1 = c_1$ is bounded between zero and one. The higher-order parameters depend on the geometry, and bounds on c_n are not known a priori. In the general anisotropic case, the parameter c_2 is bounded by

$$-c_1\tilde{c}_1 \leq c_2 \leq 0, \tag{3.1}$$

where $\tilde{c}_1 = 1 - c_1$. This geometry-independent bound on c_2 was proven in [12]. The author uses properties of the scalar measure $m(y)$ to derive the moment constraint

$$0 \leq \int_0^1 y \, \mathrm{d}m(y) \leq f_1 f_2, \tag{3.2}$$

which is equivalent to (3.1); see also [6, 27, 37, 21].

In [21] the author provides a general scheme to derive bounds on the structural parameter c_n, using lower-order parameters; see section 2.3 for the connection to complex bounds. The bounds on the c_n-parameters depend on the lower-order parameters $c_1, \ldots, c_{n-1}$.

Here, we use that c_3 is bounded by $c_3^{\min} \leq c_3 \leq c_3^{\max}$ with [21]

$$c_3^{\min} = \frac{c_2^2}{c_1}, \quad c_3^{\max} = -c_2\left(1 + \frac{c_2}{\tilde{c}_1}\right) \tag{3.3}$$

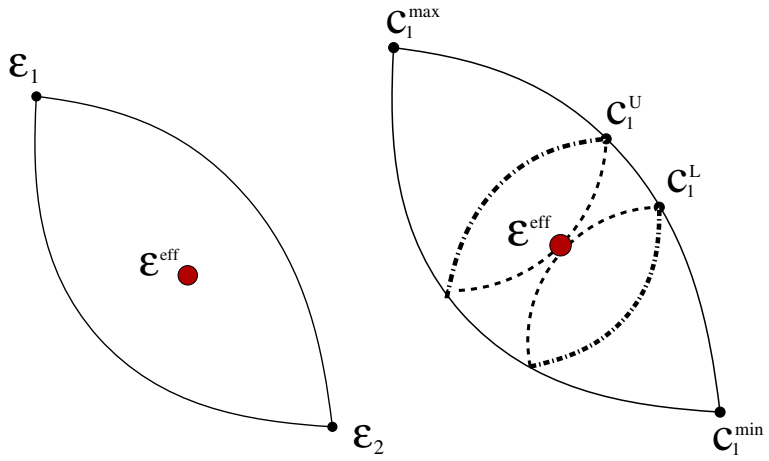


FIG. 1. *Left: The effective permittivity* ϵ^{eff} *is bounded by* $\epsilon_1^{\text{L}}(c_1)$ *and* $\epsilon_1^{\text{U}}(c_1)$. *Right: For some value* $c_1 = c_1^{\text{L}}$, *the effective permittivity* ϵ^{eff} *is on the boundary of* $\epsilon_2^{\text{L}}(c_2; c_1^{\text{L}}, \omega_0)$, $\epsilon_2^{\text{U}}(c_2; c_1^{\text{L}}, \omega_0)$ *and for some value* $c_1 = c_1^{\text{U}}$, *the effective permittivity is on the boundary of* $\epsilon_2^{\text{L}}(c_2; c_1^{\text{U}}, \omega_0)$, $\epsilon_2^{\text{U}}(c_2; c_1^{\text{U}}, \omega_0)$.

and that the structural parameter c_4 is bounded by $c_4^{\min} \leq c_4 \leq c_4^{\max}$, where [21]

$$c_4^{\min} = \frac{c_2^3 + \tilde{c}_1 c_2^2 + c_2 c_3(\tilde{c}_1 - c_1) + c_3(c_3 - c_1\tilde{c}_1)}{c_2 + c_1\tilde{c}_1}, \quad c_4^{\max} \leq \frac{c_3^2}{c_2}. \tag{3.4}$$

3.2. Bounds using one measurement. Assume that the complex value of one effective parameter $\epsilon^{\text{eff}}(\omega_0)$ is measured for some frequency ω_0. We derive bounds on c_1, together with bounds on the effective parameter $\epsilon^{\text{eff}}(\omega_1)$, when $\epsilon_1(\omega_0)$, $\epsilon_2(\omega_0)$, $\epsilon_1(\omega_1)$, and $\epsilon_2(\omega_1)$ are known constants. If the volume fraction c_1 is known, the parameter c_2 is bounded, and so on.

We assume that at least one of the phases has a positive imaginary part. That is, we assume that there are losses somewhere in the composite material. In the case of real values of both the phases, the method developed in [21] can be used to obtain bounds on c_1 and on ϵ^{eff}. In the lossless case, a direct calculation of the inverse of $\epsilon_1^{\text{L}}(c_1)$ and $\epsilon_1^{\text{U}}(c_1)$ is possible. When the measurements are complex-valued, a different approach is needed.

The measured value $\epsilon^{\text{eff}}(\omega_0)$ is inside the lens-shaped region bounded by

$$\epsilon_1^{\text{L}}(c_1; \omega_0) = \frac{1}{1 + \tilde{c}_1 z(\omega_0)}, \quad \epsilon_1^{\text{U}}(c_1; \omega_0) = \epsilon_1(\omega_0) + c_1 \epsilon_2(\omega_0) z(\omega_0), \tag{3.5}$$

with $z(\omega_0) = (\epsilon_1(\omega_0) - \epsilon_2(\omega_0))/\epsilon_2(\omega_0)$, $\tilde{c}_1 = 1 - c_1$, and $0 \leq \tilde{c}_1 \leq 1$. The boundary of the region is depicted in Figure 1.

For some values of c_1 and c_2, the effective parameter $\epsilon^{\text{eff}}(\omega_0)$ is on the curve $\epsilon_2^{\text{U}}(c_2, c_1; \omega_0)$; see Figure 1. The parameters c_1 and c_2 then solve the equation

$$\epsilon^{\text{eff}}(\omega_0) = \epsilon_2(\omega_0) \frac{c_1 - c_2 z(\omega_0) + c_1^2 z(\omega_0)}{c_1 - c_2 z(\omega_0)}, \tag{3.6}$$

with $0 \leq c_1 \leq 1$ and $-c_1\tilde{c}_1 \leq c_2 \leq 0$. We show below that (3.6) has one solution (c_1, c_2), except for trivial cases.

At the minimum volume fraction $c_1 = 0$ and at the maximum volume fraction $c_1 = 1$, the ϵ_2^{U}-bound reduces to

$$\epsilon_2^{\mathrm{U}}(0, c_2) = \epsilon_2^{\mathrm{U}}(0, 0) = \epsilon_1^{\mathrm{U}}(0) = \epsilon_2, \quad \epsilon_2^{\mathrm{U}}(1, c_2) = \epsilon_2^{\mathrm{U}}(1, 0) = \epsilon_1^{\mathrm{U}}(1) = \epsilon_1, \tag{3.7}$$

which implies that (3.6) has the solutions

$$\epsilon_2 = \epsilon_2^{\mathrm{U}}(0, 0) = \epsilon^{\mathrm{eff}}, \quad \epsilon_1 = \epsilon_2^{\mathrm{U}}(1, 0) = \epsilon^{\mathrm{eff}}. \tag{3.8}$$

By multiplying (3.6) with the denominator in ϵ_2^{U} we obtain

$$(c_1 - c_2 z)\epsilon^{\mathrm{eff}} = \epsilon_2(c_1 - c_2 z + c_1^2 z^2). \tag{3.9}$$

We assume that $\epsilon^{\mathrm{eff}} \neq \epsilon_1$ and look for solutions to (3.9) when $0 \leq c_1 \leq 1$ and $-c_1\tilde{c}_1 \leq c_2 \leq 0$. Taking the real and imaginary part of (3.9), which is quadratic in c_1 and linear in c_2, gives one solution (c_1, c_2), except for the trivial solution $(c_1, c_2) = (0, 0)$.

The calculated value on c_1 is a lower bound $c_1^{\mathrm{L}}(\omega_0)$ on the volume fraction c_1. Explicitly, the volume fraction is bounded from below by

$$c_1^{\mathrm{L}} = \Im(z)\frac{(\Im(\epsilon^{\mathrm{eff}}) - \Im(\epsilon_2))^2 + (\Re(\epsilon^{\mathrm{eff}}) - \Re(\epsilon_2))^2}{|z|^2(\Im(\epsilon^{\mathrm{eff}})\Re(\epsilon_2) - \Re(\epsilon^{\mathrm{eff}})\Im(\epsilon_2))}. \tag{3.10}$$

In the same way, for some values of $\tilde{c}_1$ and $\tilde{c}_2$, the effective parameter $\epsilon^{\mathrm{eff}}(\omega_0)$ is on the curve $\epsilon_2^{\mathrm{L}}(\tilde{c}_2, \tilde{c}_1; \omega_0)$. That is, we solve the equation

$$\epsilon^{\mathrm{eff}}(\omega_0) = \epsilon_1(\omega_0)\frac{\tilde{c}_1 - \tilde{c}_2 z(\omega_0)}{\tilde{c}_1 - \tilde{c}_2 z(\omega_0) + \tilde{c}_1^2 z(\omega_0)} \tag{3.11}$$

when $0 \leq \tilde{c}_1 \leq 1$ and $-\tilde{c}_1(1 - \tilde{c}_1) \leq \tilde{c}_2 \leq 0$. Equation (3.11) has one solution (c_1, c_2), except for the trivial cases below.

At the endpoints $(c_1, c_2) = (0, 0)$ and $(c_1, c_2) = (1, 0)$, (3.11) has the solutions

$$\epsilon_1 = \epsilon_2^{\mathrm{L}}(0, 0) = \epsilon^{\mathrm{eff}}, \quad \epsilon_2 = \epsilon_2^{\mathrm{L}}(1, 0) = \epsilon^{\mathrm{eff}}. \tag{3.12}$$

Assume that $\epsilon^{\mathrm{eff}} \neq \epsilon_2$, and multiply (3.11) with the denominator in ϵ_2^{L}. The resulting equation has one solution, except for the trivial solution $(\tilde{c}_1, \tilde{c}_2) = (0, 0)$.

The solution to the equation $\epsilon^{\mathrm{eff}}(\omega_0) = \epsilon_2^{\mathrm{L}}(\tilde{c}_1, \tilde{c}_2)$ and the relation $c_1 = 1 - \tilde{c}_1$ give an upper bound $c_1^{\mathrm{U}}(\omega_0)$ on the volume fraction c_1. Explicitly, the volume fraction is bounded from above by

$$c_1^{\mathrm{U}} = 1 - \Im(z)\frac{(\Im(\epsilon^{\mathrm{eff}}) - \Im(\epsilon_1))^2 + (\Re(\epsilon^{\mathrm{eff}}) - \Re(\epsilon_1))^2}{|z|^2(\Re(\epsilon^{\mathrm{eff}})\Im(\epsilon_1) - \Im(\epsilon^{\mathrm{eff}})\Re(\epsilon_1))}. \tag{3.13}$$

The derived bounds (3.10) and (3.13) on the volume fraction are equivalent to the bounds in [14]. Here we use a different method, which seems to be easier to generalize.

If $c_1 = c_1^{\mathrm{L}}$, the measured value $\epsilon^{\mathrm{eff}}(\omega_0)$ is equal to ϵ_2^{U} for some value of c_2. If $c_1 = c_1^{\mathrm{U}}$, the effective permittivity $\epsilon^{\mathrm{eff}}(\omega_0)$ is equal to ϵ_2^{L} for some value on c_2. The effective permittivity is bounded by the one-point bounds $\epsilon_1^{\mathrm{L}}(c_1)$ and $\epsilon_1^{\mathrm{U}}(c_1)$. From the calculations above and Figure 1, it follows that the effective permittivity also is bounded by the two-point bounds

$$\epsilon_2^{\mathrm{L}}(c_2, c_1^{\mathrm{L}}), \quad \text{with} \quad -c_1^{\mathrm{L}}(1 - c_1^{\mathrm{L}}) \leq c_2 \leq 0 \tag{3.14}$$

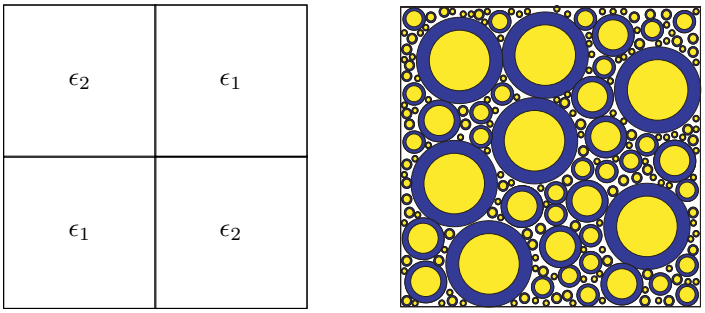


FIG. 2. *Left: The checkerboard structure, a two-dimensional and periodic problem. Right: The Hashin structure. Coated spheres that are composed of a spherical core of permittivity ϵ_2 are surrounded by a concentric shell of permittivity ϵ_1.*

and

$$\epsilon_2^{\mathrm{U}}(c_2, c_1^{\mathrm{U}}), \quad \text{with} \quad -c_1^{\mathrm{U}}(1 - c_1^{\mathrm{U}}) \leq c_2 \leq 0. \tag{3.15}$$

These bounds can, for example, be used to check the volume fraction in experiments when it is difficult to determine the volume fraction from direct measurements. If we measure the lossy permittivity for more than one frequency, the minimum of the calculated bounds on c_1 is the optimal.

3.2.1. Asymptotic behavior. Write c_2 on the form $c_2 = -\alpha c_1 \tilde{c}_1$, $0 \leq \alpha \leq 1$, and let $\epsilon_1 = 1$ and $\epsilon_2 = 1 + \delta w$, where w is a complex number with nonzero imaginary part and modulus one. Using the expansion (2.8), the asymptotic behavior when $\delta \to 0$ is

$$c_1^{\mathrm{U}} - c_1^{\mathrm{L}} = c_1 \hat{c}_1 \alpha (1 - \alpha) \delta^2 + \mathcal{O}(\delta^3). \tag{3.16}$$

For a fixed δ, the difference is small when the c_2 parameter is close to the endpoints (3.1) and when the volume fraction $c_1 = f_1$ is close to its endpoints.

In the case of real-valued phases, the parameter c_1 is bounded by [21]

$$c_1^{\mathrm{L}} = \frac{1/\epsilon^{\mathrm{eff}} - 1/\epsilon_2}{1/\epsilon_1 - 1/\epsilon_2}, \quad c_1^{\mathrm{U}} = \frac{\epsilon_2 - \epsilon^{\mathrm{eff}}}{\epsilon_2 - \epsilon_1}. \tag{3.17}$$

To proceed, let $\epsilon_1 = 1$ and $\epsilon_2 = 1 + \delta$. Using the expansion (2.8), the asymptotic behavior when $\delta \to 0$ is in the lossless case given by

$$c_1^{\mathrm{U}} - c_1^{\mathrm{L}} = c_1 \hat{c}_1 \delta + \mathcal{O}(\delta^2). \tag{3.18}$$

The convergence is faster in the complex-valued case, which in many cases of interest implies much tighter bounds on the volume fraction. One interpretation of the result is that a measurement of a complex value contains more information compared to a measurement of a real value.

3.2.2. Examples. As a first illustration of the theory presented above, assume that one of the phases is a frequency-independent material $\epsilon_1(\omega) = 3$ in the chosen range of frequencies. Moreover, phase two is lossy and measured at the frequencies ω_0, ω_1, and ω_2. We use the checkerboard structure and the Hashin structure; see Figure 2 to exemplify the method.

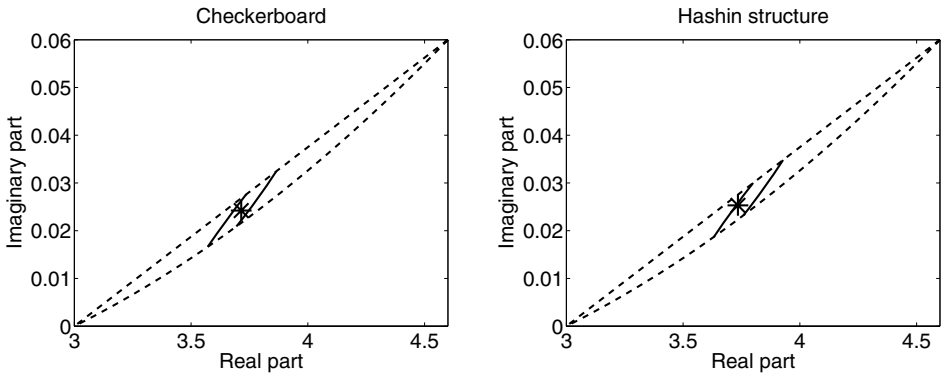


FIG. 3. *The star to the left is the effective permittivity in the checkerboard case and the star to the right corresponds to* ϵ^{eff} *for the Hashin structure. In both figures, the dashed lines* $\epsilon_1^{\text{L}}(c_1)$ *and* $\epsilon_1^{\text{U}}(c_1)$ *bound* $\epsilon^{\text{eff}}(\omega_1)$ *and the solid lines are the tighter bounds* $\epsilon_2^{\text{L}}(c_2; c_1^{\text{L}})$ *and* $\epsilon_2^{\text{U}}(c_2; c_1^{\text{U}})$.

Phase two has the value $\epsilon_2(\omega_0) = 4.1 + 4.5\text{i}$ at frequency ω_0. The checkerboard structure has the exact effective permittivity [34]

$$\epsilon_{\text{C}}^{\text{eff}}(\omega) = \sqrt{\epsilon_1(\omega)\epsilon_2(\omega)}. \tag{3.19}$$

It is interesting to notice that the checkerboard structure corresponds exactly to Bruggemans formula [34] at the percolation threshold $c_1 = 0.5$.

As described above, the solutions of the equations $\epsilon_{\text{C}}^{\text{eff}} = \epsilon_2^{\text{L}}$ and $\epsilon_{\text{C}}^{\text{eff}} = \epsilon_2^{\text{U}}$ bound the volume fraction c_1. Figure 3 shows the bounds on $\epsilon_{\text{C}}^{\text{eff}}(\omega_1)$ when $\epsilon_2(\omega_1) = 4.6{+}0.06\text{i}$ is known and the bounds on c_1 are calculated to $c_1^{\text{L}}(\omega_0) = 0.46$ and $c_1^{\text{U}}(\omega_0) = 0.54$. The exact value on the volume fraction is $c_1 = 0.5$.

The Hashin structure [34] (see Figure 2) in d dimensions has the effective permittivity

$$\epsilon_{\text{H}}^{\text{eff}}(\omega) = \epsilon_1(\omega)\frac{(d-1)c_1(\epsilon_1(\omega) - \epsilon_2(\omega)) + d\epsilon_2(\omega)}{d\epsilon_1(\omega) + c_1(\epsilon_2(\omega) - \epsilon_1(\omega))}. \tag{3.20}$$

We consider the three-dimensional case, $d = 3$, with the volume fraction $c_1 = 0.5$. Using the values of ϵ_1, ϵ_2, and $\epsilon_{\text{H}}^{\text{eff}}$ at $\omega = \omega_0$ the bounds on c_1 are calculated to $c_1^{\text{L}}(\omega_0) = 0.42$ and $c_1^{\text{U}}(\omega_0) = 0.50$. Figure 3 shows the bounds on $\epsilon_{\text{H}}^{\text{eff}}(\omega_1)$ when $\epsilon_2(\omega_1)$ is known.

3.2.3. Bounds when the volume fraction is known. If the volume fraction c_1 is known, we obtain in the same way bounds on c_2. The measured value $\epsilon^{\text{eff}}(\omega_0)$ is bounded by the lens-shaped region $\epsilon_2^{\text{L}}(c_2; \omega_0)$ and $\epsilon_2^{\text{U}}(c_2; \omega_0)$, with $-c_1\tilde{c}_1 \le c_2 \le 0$.

For some values of c_2 and c_3, the effective parameter $\epsilon^{\text{eff}}(\omega_0)$ is on the boundary of $\epsilon_3^{\text{U}}(c_2, c_3; \omega_0)$, which is given by the Padé approximation $\epsilon_{1,1}$ of the series (2.8). On the curve, the parameters c_2 and c_3 satisfy the equation

$$\epsilon^{\text{eff}}(\omega_0) = \epsilon_2\frac{c_2 + c_1c_2z + c_2^2z^2 - c_3z(1 + c_1z)}{c_2 - c_3z}, \tag{3.21}$$

with $-c_1\tilde{c}_1 \le c_2 \le 0$ and $c_2^2/c_1 \le c_3 \le -c_2(1 + c_2/\tilde{c}_1)$.

At the minimum, $c_2 = -c_1\tilde{c}_1$, (3.21) has the solution

$$\epsilon_1^{\text{L}}(c_1) = \epsilon_2^{\text{U}}(c_2^{\min}) = \epsilon_3^{\text{U}}(c_2^{\min}, c_3^{\min}(c_2^{\min})) = \epsilon^{\text{eff}}, \tag{3.22}$$

where $c_3^{\min}(c_2^{\min}) = c_1(1-c_1)^2$, and at the maximum $c_2 = 0$, the solution to the equation is

$$\epsilon_1^{\mathrm{U}}(c_1) = \epsilon_2^{\mathrm{U}}(c_2^{\max}) = \epsilon_3^{\mathrm{U}}(c_2^{\max}, c_3^{\max}(c_2^{\max})) = \epsilon^{\mathrm{eff}}, \tag{3.23}$$

where $c_3^{\max}(c_2^{\max}) = 0$. By multipling (3.21) with the denominator in ϵ_3^{U} an equation quadratic in c_2 and linear in c_3 is obtained. Assume that $\epsilon^{\mathrm{eff}} \neq \epsilon_1^{\mathrm{L}}(c_1)$. Taking the real and imaginary part gives one solution (c_2, c_3), except for the trivial solution when $(c_2, c_3) = (0, 0)$. The calculated value on c_2 is an upper bound $c_2^{\mathrm{U}}(\omega_0)$ on the structural parameter c_2.

Analogously, for some values of $\tilde{c}_2$ and $\tilde{c}_3$, the effective parameter $\epsilon^{\mathrm{eff}}(\omega_0)$ is located on the boundary of $\epsilon_3^{\mathrm{L}}(\tilde{c}_2, \tilde{c}_3; \omega_0)$, which is given by the Padé approximation $\epsilon_{1,1}$ of the series (2.16). That is, the equation

$$\epsilon^{\mathrm{eff}}(\omega_0) = \epsilon_1 \frac{\tilde{c}_2 - \tilde{c}_3 z}{\tilde{c}_2 + \tilde{c}_1\tilde{c}_2 z + \tilde{c}_2^2 z^2 - \tilde{c}_3 z(1 + \tilde{c}_1 z)} \tag{3.24}$$

is solved with respect to $\tilde{c}_2$ and $\tilde{c}_3$. Using that the coefficients c_n and $\tilde{c}_n$ are related by (2.17), and solving the equation $\epsilon^{\mathrm{eff}}(\omega_0) = \epsilon_3^{\mathrm{L}}(c_2, c_3)$, gives a lower bound $c_2^{\mathrm{L}}(\omega_0)$ on the structural parameter c_2. As before, the equation has one solution, except for the cases when $\epsilon^{\mathrm{eff}} = \epsilon_1^{\mathrm{L}}(c_1)$ and when $\epsilon^{\mathrm{eff}} = \epsilon_1^{\mathrm{U}}(c_1)$.

It is possible to derive explicit formulas for c_2^{L} and c_2^{U}, but they contain many terms and will for this reason not be presented.

The effective permittivity is bounded by the two-point bounds $\epsilon_2^{\mathrm{L}}(c_2)$ and $\epsilon_1^{\mathrm{U}}(c_2)$. We have shown that the effective permittivity also is bounded by the three-point bounds

$$\epsilon_3^{\mathrm{L}}(c_3, c_2^{\mathrm{U}}), \quad \text{with} \quad \frac{c_2^{\mathrm{U}}}{c_1} \leq c_3 \leq -c_2^{\mathrm{U}}\left(1 + \frac{c_2^{\mathrm{U}}}{1-c_1}\right) \tag{3.25}$$

and

$$\epsilon_3^{\mathrm{U}}(c_3, c_2^{\mathrm{L}}), \quad \text{with} \quad \frac{c_2^{\mathrm{L}}}{c_1} \leq c_3 \leq -c_2^{\mathrm{L}}\left(1 + \frac{c_2^{\mathrm{L}}}{1-c_1}\right), \tag{3.26}$$

where c_2^{U} is calculated from (3.21) and c_2^{L} is the solution to (3.24). The bounds on c_3 are given by (3.3).

In many cases of interest, the composite is known to be isotropic, $c_2 = -c_1\hat{c}_1/d$. The bounds on c_2 can then be used to check experimental data. If the volume fraction c_1 is known and $-c_1\hat{c}_1/d$ does not belong to the interval $[c_2^{\mathrm{L}}, c_2^{\mathrm{U}}]$, the experimental value on ϵ^{eff} is inconsistent with the bounds.

3.2.4. Examples. The checkerboard structure and the Hashin structure, with the same values on the phases as before, are used to illustrate the method.

The checkerboard has volume fraction $c_1 = 0.5$, which is assumed to be known. Figure 4 shows bounds on $\epsilon^{\mathrm{eff}}(\omega_1)$ when the bounds on c_2 are calculated to $c_2^{\mathrm{L}}(\omega_0) = -0.135$ and $c_2^{\mathrm{U}}(\omega_0) = -0.115$. The checkerboard problem is two-dimensional and isotropic. The second moment, (2.12), with $c_1 = 0.5$ is then exactly $c_2 = -1/8 = -0.125$.

The Hashin structure is three-dimensional and isotropic. Using $c_1 = 0.5$, the second moment is $c_2 = -1/12 \approx -0.0833$. In this case the solution of the equations (3.21) and (3.24), when $c_1 = 0.5$ is known, determines c_2 numerically. The lower

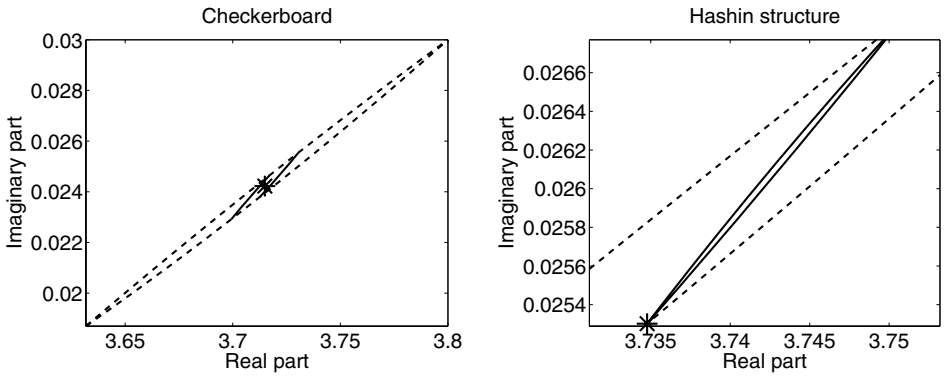


FIG. 4. *The star to the left is the effective permittivity in the checkerboard case and the star to the right corresponds to* ϵ^{eff} *for the Hashin structure. In both figures, the dashed lines* $\epsilon_2^{\text{L}}(c_2)$ *and* $\epsilon_2^{\text{U}}(c_2)$ *bound* $\epsilon^{\text{eff}}(\omega_1)$ *and the solid lines are the tighter bounds* $\epsilon_3^{\text{L}}(c_3; c_2^{\text{U}})$ *and* $\epsilon_2^{\text{U}}(c_3; c_2^{\text{L}})$.

bound and the upper bound on c_2 have 15 digits in common when the equations are solved with Mathematica (www.wolfram.com). In the following section, we use the approximative value $c_2 = c_2^{\text{U}}(\omega_0) = c_2^{\text{L}}(\omega_0) = -0.0833$. Figure 4 shows bounds on $\epsilon^{\text{eff}}(\omega_1)$ when $\epsilon_2(\omega_1) = 4.6 + 0.06\text{i}$ is known.

3.2.5. Bounds on isotropic materials. If the volume fraction c_1, together with the c_2 parameter, is known (for example, if the material is isotropic, $\mathbf{c}_2 = -(c_1\tilde{c}_1/d)\mathbf{I}$), the equations

$$\epsilon^{\text{eff}}(\omega_0) = \epsilon_4^{\text{U}}(c_3, c_4; \omega_0), \quad \epsilon^{\text{eff}}(\omega_0) = \epsilon_4^{\text{L}}(\tilde{c}_3, \tilde{c}_4; \omega_0) \tag{3.27}$$

give us bounds on c_3. In general, if the structural parameters $c_1, c_2, \ldots, c_n$ are known, we obtain bounds on c_{n+1} from the equations

$$\epsilon^{\text{eff}}(\omega_0) = \epsilon_{n+1}^{\text{U}}(c_{n+1}, c_{n+2}; \omega_0), \quad \epsilon^{\text{eff}}(\omega_0) = \epsilon_{n+1}^{\text{L}}(\tilde{c}_{n+1}, \tilde{c}_{n+2}; \omega_0). \tag{3.28}$$

We can also get bounds on one structural parameter c_n if $c_1, c_2, \ldots, c_{n-1}$ and c_{n+1} are known. For example, if the material is known to be isotropic, $c_2 = -c_1(1-c_1)/d$, the Hashin–Shtrikman bounds give us tighter bounds on the volume fraction than the solution to (3.6).

3.2.6. Examples. The bounds on $c_1(\omega_0)$ for the checkerboard structure above were calculated to $c_1^{\text{L}}(\omega_0) = 0.46$ and $c_1^{\text{U}}(\omega_0) = 0.54$. We now use that $c_2 = -0.125$ and solve $\epsilon^{\text{eff}}(\omega_0) = \epsilon_3^{\text{L}}$ and $\epsilon^{\text{eff}}(\omega_0) = \epsilon_3^{\text{U}}$ with respect to c_1 and c_3. Excluding trivial solutions, we get the bounds

$$c_1^{\text{L}}(\omega_0) = 0.492, \quad c_1^{\text{U}}(\omega_0) = 0.508. \tag{3.29}$$

When the composite is known to be isotropic and the volume fraction is known, the effective permittivity is bounded by the three-point bounds $\epsilon_3^{\text{L}}(c_3)$ and $\epsilon_3^{\text{U}}(c_3)$. The effective value is also bounded by the four-point bounds

$$\epsilon_4^{\text{L}}(c_4, c_3^{\text{L}}), \quad \text{with} \quad c_4^{\min}(c_3^{\text{L}}) \le c_4 \le c_4^{\max}(c_3^{\text{L}}) \tag{3.30}$$

and

$$\epsilon_4^{\text{U}}(c_4, c_3^{\text{U}}), \quad \text{with} \quad c_4^{\min}(c_3^{\text{U}}) \le c_4 \le c_4^{\max}(c_3^{\text{U}}), \tag{3.31}$$

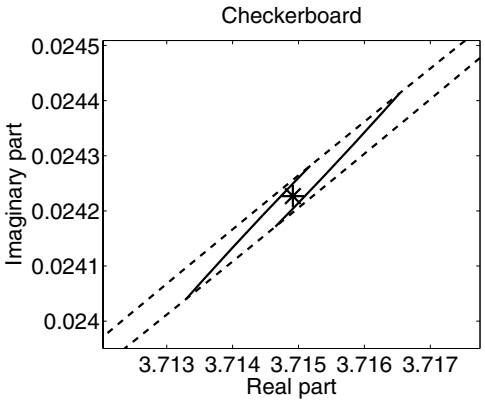


FIG. 5. *The star is the effective permittivity* $\epsilon^{\text{eff}}(\omega_1)$ *in the checkerboard case. The dashed lines* $\epsilon_3^{\text{L}}(c_3)$ *and* $\epsilon_3^{\text{U}}(c_3)$ *bound* $\epsilon^{\text{eff}}(\omega_1)$ *and the solid lines are the tighter bounds* $\epsilon_4^{\text{L}}(c_4; c_3^{\text{L}})$ *and* $\epsilon_4^{\text{U}}(c_4; c_3^{\text{U}})$.

where the bounds on c_4 are given by (3.4). We use that the checkerboard is isotropic and that the volume fraction is $c_1 = 0.5$. Using the same values on the phases as above, the bounds on c_3 are calculated to $c_3^{\text{L}}(\omega_0) = 0.0601$ and $c_3^{\text{U}}(\omega_0) = 0.0649$, respectively. The geometry-independent bounds (3.3) are in this case $c_3^{\min} = 0.0315$ and $c_3^{\max} = 0.09375$.

The exact value on c_3 can be identified from a Taylor expansion of $\epsilon^{\text{eff}} = \sqrt{\epsilon_1\epsilon_2}$ when $\epsilon_1 = 1$ and $\epsilon_2 = 1 + \eta$, $\eta < 1$. The effective permittivity ϵ^{eff} is then

$$\epsilon^{\text{eff}}(1, 1+\eta) = 1 + \frac{1}{2}\eta - \frac{1}{8}\eta^2 + \frac{1}{16}\eta^3 - \frac{5}{128}\eta^4 + \cdots . \tag{3.32}$$

The bounds on c_3 are tight, and the arithmetic mean $(c_3^{\text{L}}(\omega_0)+c_3^{\text{U}}(\omega_0))/2$ provides an accurate approximation of $c_3 = 1/16 = 0.0625$. Figure 5 shows the bounds on $\epsilon^{\text{eff}}(\omega_1)$ when the volume fraction is $c_1 = 0.5$ and the composite is known to be isotropic, $c_2 = -0.125$.

The Hashin structure is three-dimensional and isotropic. Using the same values as above, the solution of $\epsilon_4^{\text{U}} = \epsilon_{\text{H}}^{\text{L}}$ gives the lower bound $c_3^{\text{L}} = c_3^{\max}$. This solution determines c_3 numerically. The lower bound c_3^{L} and the maximum $c_3^{\max}$ have 16 digits in common when the equations are solved with Mathematica.

The properties $\epsilon_3^{\text{U}}(c_3^{\max}) = \epsilon_2^{\text{L}}(c_2, c_1)$ and $\epsilon_3^{\text{L}}(c_3^{\max}) = \epsilon_2^{\text{L}}(c_2, c_1)$ imply that $\epsilon^{\text{eff}} = \epsilon_2^{\text{L}}(c_2, c_1)$.

When the composite is isotropic, the lower bound ϵ_2^{L} is equivalent to the Maxwell–Garnett formula [30, 34]. This formula, commonly used by experimentalists, is a good approximation formula if c_3 is close to $c_3^{\max}$.

3.3. Bounds using two measurements. We cannot determine bounds on more than one structural parameter with information from one measurement. If we have two measurements, which give us different bounds on c_1, it is also possible to get bounds on c_2 without any assumptions on the microstructure. Geometrically, we fail to get bounds on c_2 from one measurement, because the effective permittivity is (by construction) on the boundary of the $\epsilon_2^{\text{L}}/\epsilon_2^{\text{U}}$-bounds, when $c_1 = c_1^{\text{L}}$ or $c_1 = c_1^{\text{U}}$.

Assume that the measurement of $\epsilon^{\text{eff}}(\omega_0)$ gives us tighter bounds $c_1^{\text{L}}(\omega_0) \le c_1 \le c_1^{\text{U}}(\omega_0)$ than the measurement of $\epsilon^{\text{eff}}(\omega_1)$. If we use the tighter bounds $c_1^{\text{L}}(\omega_0) \le c_1 \le c_1^{\text{U}}(\omega_0)$, together with the measurement $\epsilon^{\text{eff}}(\omega_1)$, we avoid the boundary and can

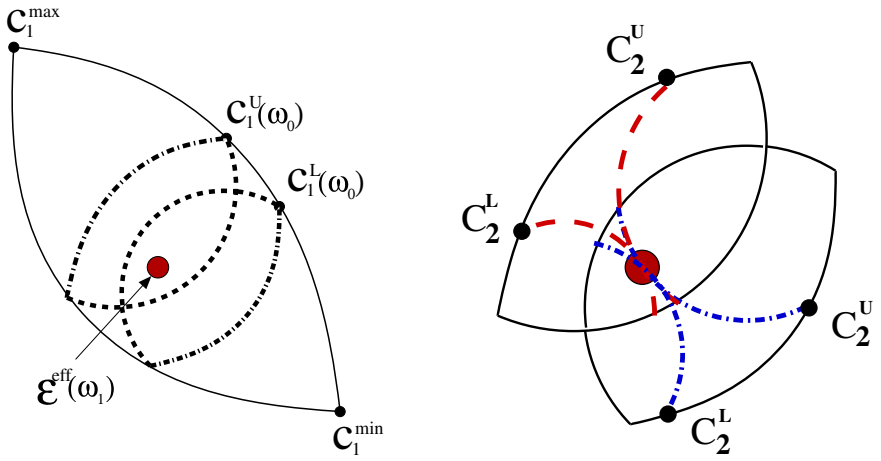


FIG. 6. *Left: The union of the regions* $\epsilon_2^{\mathrm{L}}(c_2;c_1^{\mathrm{L}})$, $\epsilon_2^{\mathrm{U}}(c_2;c_1^{\mathrm{L}})$ *and* $\epsilon_2^{\mathrm{L}}(c_2;c_1^{\mathrm{U}})$, $\epsilon_2^{\mathrm{U}}(c_2;c_1^{\mathrm{U}})$ *bound* $\epsilon^{\mathrm{eff}}(\omega_1)$. *Right: For some values on* $c_2 = c_2(c_1)$, *the effective permittivity* $\epsilon^{\mathrm{eff}}(\omega_1)$ *is on the boundary of* $\epsilon_3^{\mathrm{L}}(c_3;c_2(c_1),c_1,\omega_1)$, $\epsilon_2^{\mathrm{U}}(c_3;c_2(c_1),c_1,\omega_1)$. *The two lens-shaped regions correspond to two different values on the volume fraction* c_1.

continue to bound c_2. This simple observation is the key to the construction of the bounds on any structural parameter.

To bound the c_1-dependent parameter c_2 with a fixed value on c_1 the equations

$$\epsilon^{\mathrm{eff}}(\omega_1) = \epsilon_3^{\mathrm{U}}(c_3,c_2;c_1(\omega_0)), \quad \epsilon^{\mathrm{eff}}(\omega_1) = \epsilon_3^{\mathrm{L}}(c_3,c_2;c_1(\omega_0)) \tag{3.33}$$

are solved in the range $c_1^{\mathrm{L}}(\omega_0) \le c_1 \le c_1^{\mathrm{U}}(\omega_0)$. By construction, the two lens-shaped regions $\epsilon_2^{\mathrm{L}}(c_2;c_1^{\mathrm{L}},\omega_0)$, $\epsilon_2^{\mathrm{U}}(c_2;c_1^{\mathrm{L}},\omega_0)$ and $\epsilon_2^{\mathrm{L}}(c_2;c_1^{\mathrm{U}},\omega_0)$, $\epsilon_2^{\mathrm{U}}(c_2;c_1^{\mathrm{U}},\omega_0)$ intersect; see Figure 6. From the bound ϵ_3^{U}, we get an upper bound $c_2^{\mathrm{U}}(c_1)$ on c_2, and the lower bound ϵ_3^{L} provides a lower bound $c_2^{\mathrm{U}}(c_1)$ on c_2.

We can now construct three-point bounds on ϵ^{eff} by forming

$$\epsilon_3^{\mathrm{L}}(c_3,c_2^{\mathrm{U}}(c_1),c_1), \quad \epsilon_3^{\mathrm{U}}(c_3,c_2^{\mathrm{L}}(c_1),c_1), \tag{3.34}$$

with $c_1 \in [c_1^{\mathrm{L}}(\omega_0),c_1^{\mathrm{U}}(\omega_0)]$ and $c_3 \in [c_3^{\min},c_3^{\max}]$. The c_1-dependent maximum $c_3^{\max}$ and the minimum $c_3^{\max}$ are taken from the expression (3.3).

From the derivation of the maximum $c_3^{\max}$ and the minimum $c_3^{\min}$ in [21] we have the equalities $\epsilon_3^{\mathrm{L}} = \epsilon_2^{\mathrm{L}}$ when $c_3 = c_3^{\max}$ and $\epsilon_3^{\mathrm{L}} = \epsilon_2^{\mathrm{U}}$ when $c_3 = c_3^{\min}$. In the same way the upper bound ϵ_3^{U} can be used to limit the c_3-parameter. We obtain the equalities $\epsilon_3^{\mathrm{U}} = \epsilon_2^{\mathrm{L}}$ when $c_3 = c_3^{\max}$ and $\epsilon_3^{\mathrm{U}} = \epsilon_2^{\mathrm{U}}$ when $c_3 = c_3^{\min}$. Using these properties, the bounding region in (3.34) that depends on two variables c_1 and c_3 can be expressed as a set of bounds, depending on one single variable. The new bounds are

$$\epsilon_3^{\mathrm{U}}(c_3;c_2^{\mathrm{L}}(c_1^{\mathrm{U}}),c_1^{\mathrm{U}}), \quad \epsilon_3^{\mathrm{U}}(c_3;c_2^{\mathrm{L}}(c_1^{\mathrm{L}}),c_1^{\mathrm{L}}), \quad \epsilon_2^{\mathrm{U}}(c_1,c_2^{\mathrm{L}}(c_1)), \quad \epsilon_2^{\mathrm{L}}(c_1,c_2^{\mathrm{L}}(c_1)) \tag{3.35}$$

and

$$\epsilon_3^{\mathrm{L}}(c_3;c_2^{\mathrm{U}}(c_1^{\mathrm{U}}),c_1^{\mathrm{U}}), \quad \epsilon_3^{\mathrm{L}}(c_3;c_2^{\mathrm{U}}(c_1^{\mathrm{L}}),c_1^{\mathrm{L}}), \quad \epsilon_2^{\mathrm{U}}(c_1,c_2^{\mathrm{U}}(c_1)), \quad \epsilon_2^{\mathrm{L}}(c_1,c_2^{\mathrm{U}}(c_1)), \tag{3.36}$$

where the two-point bounds depend on $c_1 \in [c_1^{\mathrm{L}}(\omega_0),c_1^{\mathrm{U}}(\omega_0)]$ and the three-point bounds depend on $c_3 \in [c_3^{\min},c_3^{\max}]$. If some of the structural parameters are known,

for example, if the volume fraction is known and the material is isotropic, the two measurements give bounds on the higher-order moments c_3 and c_4.

The upper bound $c_2^{\mathrm{U}}(c_1)$ and the lower bound $c_2^{\mathrm{L}}(c_1)$ are both second-degree polynomials in c_1, which are easily maximized and minimized. Global, c_1-independent, bounds on c_2 are defined as

$$c_2^{\mathrm{L}}(\omega_1) = \min_{c_1 \in [c_1^{\mathrm{L}}, c_1^{\mathrm{U}}]} \{c_2^{\mathrm{L}}(c_1)\}, \quad c_2^{\mathrm{U}}(\omega_1) = \max_{c_1 \in [c_1^{\mathrm{L}}, c_1^{\mathrm{U}}]} \{c_2^{\mathrm{U}}(c_1)\}. \tag{3.37}$$

The global bounds on c_2 can be used to simplify the above formulas at the expense of less tight bounds.

3.4. The checkerboard. We give an example of the method when no structural information is known using the checkerboard structure. Assume, as before, that $\epsilon_2(\omega_0) = 4.1 + 4.5\mathrm{i}$ and $\epsilon_2(\omega_1) = 4.6 + 0.06\mathrm{i}$ are known and that $\epsilon_1 = 3$ independent of the frequency ω. Moreover, we assume that $\epsilon^{\mathrm{eff}}(\omega_0)$ and $\epsilon^{\mathrm{eff}}(\omega_1)$ are measured and seek bounds on $\epsilon^{\mathrm{eff}}(\omega_2)$ when $\epsilon_2(\omega_2) = 3.7 + 0.04\mathrm{i}$ is known.

The second measurement on frequency ω_1 gives the tightest bounds on $c_1 = 0.5$,

$$c_1^{\mathrm{L}}(\omega_1) = 0.494, \quad c_1^{\mathrm{U}}(\omega_1) = 0.506. \tag{3.38}$$

We use the measurement of the effective permittivity on the frequency ω_0 to bound c_2. The solutions to the equations $\epsilon^{\mathrm{eff}}(\omega_0) = \epsilon_3^{\mathrm{L}}$ and $\epsilon^{\mathrm{eff}}(\omega_0) = \epsilon_3^{\mathrm{U}}$ when $c_1 \in [c_1^{\mathrm{L}}(\omega_1), c_1^{\mathrm{U}}(\omega_1)]$ are

$$c_2^{\mathrm{L}}(c_1) = 1.09296 - 6.0343c_1 + 7.15922c_1^2 \tag{3.39}$$

and

$$c_2^{\mathrm{U}}(c_1) = -2.21787 + 7.28412c_1 - 6.15921c_1^2. \tag{3.40}$$

These functions have no stationary point when $c_1 \in [0.494, 0.506]$. The endpoints give the global bounds on $c_2 = -0.125$,

$$c_2^{\mathrm{L}}(\omega_0) = -0.141, \quad c_2^{\mathrm{U}}(\omega_0) = -0.108. \tag{3.41}$$

The bounds (3.35) and (3.36) that bound $\epsilon^{\mathrm{eff}}(\omega_1)$ are depicted in Figure 7.

3.5. An anisotropic example. Using the same material parameters as above, we also give an example in the anisotropic and periodic case; see Figure 8.

We use FEMLAB (www.comsol.com) to numerically calculate the solution to the local problem (2.2), (2.3). At the frequencies ω_0 and ω_1, the results are

$$\epsilon^{\mathrm{eff}}(\omega_0) = 3.9426 + 0.9852\mathrm{i}, \quad \epsilon^{\mathrm{eff}}(\omega_1) = 3.5147 + 0.01554\mathrm{i}. \tag{3.42}$$

The second measurement at the frequency ω_1 gives the tightest bounds on c_1,

$$c_1^{\mathrm{L}}(\omega_1) = 0.5941, \quad c_1^{\mathrm{U}}(\omega_1) = 0.6007. \tag{3.43}$$

We use the measurement of the effective permittivity on frequency ω_0 to bound c_2. The solutions to the equations $\epsilon^{\mathrm{eff}}(\omega_0) = \epsilon_3^{\mathrm{L}}$ and $\epsilon^{\mathrm{eff}}(\omega_0) = \epsilon_3^{\mathrm{U}}$ when $c_1 \in [c_1^{\mathrm{L}}(\omega_1), c_1^{\mathrm{U}}(\omega_1)]$ are

$$c_2^{\mathrm{L}}(c_1) = 7.07395 - 26.20256c_1 + 23.50343c_1^2 \tag{3.44}$$

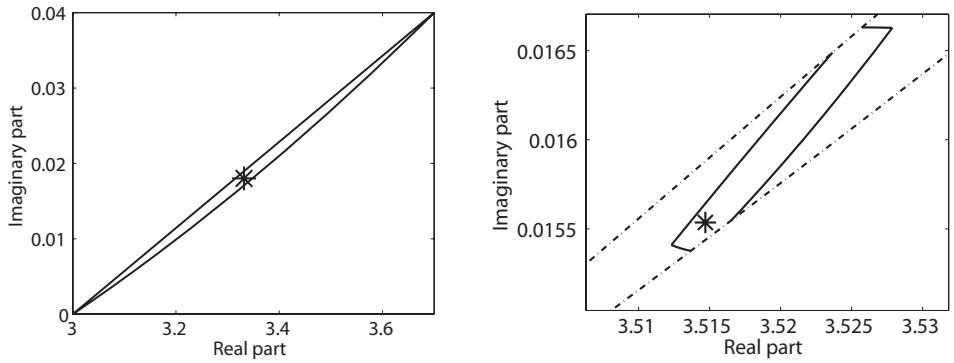


FIG. 7. *Left: The star is the location of the effective permittivity* $\epsilon^{\text{eff}}(\omega_1)$ *in the checkerboard case. The solid lines* (3.35) *and* (3.36) *bound* $\epsilon^{\text{eff}}(\omega_1)$. *Right: The star corresponds to* ϵ^{eff} *for the rods. The dash-dotted lines* $\epsilon_2^{\text{L}}(c_2; c_1^{\text{L}})$ *and* $\epsilon_2^{\text{U}}(c_2; c_1^{\text{U}})$ *bound* $\epsilon^{\text{eff}}(\omega_1)$ *and the solid lines give the tighter bounds* (3.35) *and* (3.36).

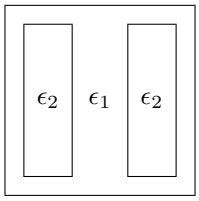


FIG. 8. *The geometry used to generate the result shown in Figures* 7 *and* 9. *Two rods with length* 0.8 *and width* 0.25 *are located, a distant* 0.3 *apart, in a unit square. The volume fraction is then* $c_1 = 0.6$. *The applied field is oriented perpendicularly to the rods.*

and

$$c_2^{\text{U}}(c_1) = -2.46585 + 6.37039c_1 - 4.27488c_1^2. \tag{3.45}$$

These functions have no stationary point when $c_1 \in [0.5941, 0.6007]$. The endpoints give the global bounds

$$c_2^{\text{L}}(\omega_0) = -0.1974, \quad c_2^{\text{U}}(\omega_0) = -0.1817. \tag{3.46}$$

The bounds on $\epsilon^{\text{eff}}(\omega_2)$ when $\epsilon_2(\omega_2) = 3.7 + 0.04\text{i}$ are known are depicted in Figure 7. The effective permittivity is numerically calculated to $\epsilon^{\text{eff}}(\omega_2) = 3.253 + 0.01306\text{i}$.

In practice, the effective permittivity (3.42) is the result of measurements and cannot in general be given with this accuracy. A computer program that takes into account that measurements have errors has been written. If we assume that the error in the measurements of $\epsilon^{\text{eff}}(\omega)$ is 1%, the bounds on the volume fraction are numerically computed to

$$0.57 \leq c_1 \leq 0.62. \tag{3.47}$$

In a separate paper [20], the method derived here will be used to analyze data from real measurements.

3.5.1. Bounds when the volume fraction is known. Assume that we have one measurement at ω_0 and one measurement at ω_2 (that here is numerically calcu-

lated in FEMLAB). The effective permittivity at ω_2 is

$$\epsilon^{\text{eff}}(\omega_2) = 3.253 + 0.01306\text{i}. \tag{3.48}$$

The measurement at frequency ω_2 gives the tightest bounds on c_1,

$$c_1^{\text{L}}(\omega_2) = 0.5984, \quad c_1^{\text{U}}(\omega_2) = 0.6002. \tag{3.49}$$

The bounds on c_1 are in this case very tight. The arithmetic mean of $c_1^{\text{L}}(\omega_2)$ and $c_1^{\text{U}}(\omega_2)$ is then approximately $c_1^{\text{app}}(\omega_2) = 0.6$, which is the exact value on the volume fraction.

If $c_1 = 0.6$ is used, the same schedule as above can be used to bound the parameters c_2 and c_3. The solution to the equations $\epsilon^{\text{eff}}(\omega_2) = \epsilon_3^{\text{L}}$ and $\epsilon^{\text{eff}}(\omega_2) = \epsilon_3^{\text{U}}$, with $c_1 = 0.6$, gives the tightest bounds on c_2,

$$c_2^{\text{L}}(\omega_2) = -0.18413, \quad c_1^{\text{U}}(\omega_2) = -0.18403. \tag{3.50}$$

The solutions to the equations $\epsilon^{\text{eff}}(\omega_0) = \epsilon_4^{\text{L}}$ and $\epsilon^{\text{eff}}(\omega_0) = \epsilon_4^{\text{U}}$, with $c_2 \in [c_2^{\text{L}}(\omega_1), c_2^{\text{U}}(\omega_1)]$, are

$$c_3^{\text{L}}(c_2) = 1.37544 + 12.75996c_2 + 31.54795c_2^2 \tag{3.51}$$

and

$$c_3^{\text{U}}(c_2) = -7.60752 - 84.64558c_2 - 232.47525c_2^2. \tag{3.52}$$

These functions have no stationary point when $c_2 \in [c_2^{\text{L}}(\omega_1), c_2^{\text{U}}(\omega_1)]$. The endpoints give the global bounds

$$c_3^{\text{L}}(\omega_0) = 0.0955, \quad c_3^{\text{U}}(\omega_0) = 0.0966. \tag{3.53}$$

The bounds on $\epsilon^{\text{eff}}(\omega_2)$ were tight when the volume fraction was unknown, and they are now even tighter. We use a composite with larger contrast to illustrate the bounds. Assume that $\epsilon_1(\omega_3) = 3 + 0.1\text{i}$ and $\epsilon_2(\omega_3) = 2 + 20\text{i}$ are known. The bounds $\epsilon_4^{\text{L}}(c_4, c_3^{\text{L}}(c_2))$ and $\epsilon_4^{\text{U}}(c_4, c_3^{\text{L}}(c_2))$ on the effective permittivity $\epsilon^{\text{eff}}(\omega_3)$ are depicted in Figure 9. The effective permittivity is numerically calculated to $\epsilon^{\text{eff}}(\omega_3) = 5.409 + 1.038\text{i}$.

The geometry and the values on the phases were previously used in [21], where the value on the volume fraction c_1 and the anisotropy c_2 were assumed to be known. Here we obtain almost as tight bounds as in [21] by using the values of two measurements of a bulk property.

The bounds on c_2 from the measurement on ω_2 are close. If we use the arithmetic mean of $c_2^{\text{L}}(\omega_2)$ and $c_2^{\text{U}}(\omega_2)$ as an approximation, the same schedule can be used to bound the parameters c_3 and c_4.

4. Discussion and conclusions. We have developed a method to calculate inverse bounds on the structural parameters from measurements of lossy two-component composites. For example, measurements can be used to determine the frequency-dependent effective permittivity.

If no structural information is known, data from two measurements determine bounds on the volume fraction and on the isotropy parameter. The bounds on the structural parameters are used to bound the permittivity at some frequency of interest or a related effective property such as the electrical and thermal conductivity, magnetism, diffusion, and flow in porous media.

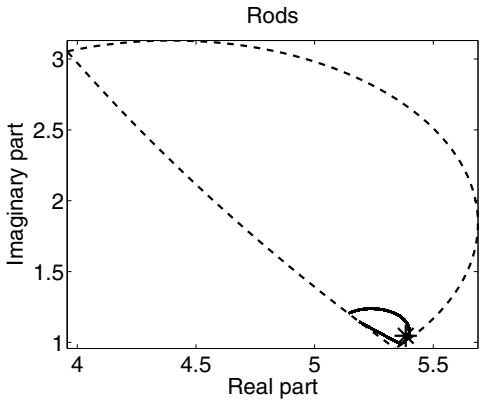


FIG. 9. *The star is the effective permittivity* $\epsilon^{\text{eff}}(\omega_3)$ *bounded by the dashed lines* $\epsilon_3^{\text{L}}(c_3; c_2^{\text{U}})$ *and* $\epsilon_3^{\text{U}}(c_3; c_2^{\text{L}})$. *The solid lines are the tighter bounds* $\epsilon_4^{\text{L}}(c_4, c_3^{\text{L}}(c_2))$ *and* $\epsilon_4^{\text{U}}(c_4, c_3^{\text{L}}(c_2))$.

In the case when some of the structural parameters are known, for example, if the composite is known to be isotropic and the volume fraction is known, the same schedule can be used to bound higher-order moments. The method can be extended to bound higher-order moments, provided that we have information from more measurements of the bulk parameters.

Numerical experiments, with reasonable values for the permittivity, were used to illustrate the method.

Acknowledgments. The author is grateful to Daniel Sjöberg and Gerhard Kristensson for many helpful discussions and comments on different parts of this paper.

REFERENCES

[1] G. A. BAKER, *Essentials of Padé Approximants*, Academic Press, New York, 1975.
[2] A. BENSOUSSAN, J. L. LIONS, AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, Stud. Math. Appl. 5, North–Holland, Amsterdam, 1978.
[3] M. J. BERAN, *Use of the variational approach to determine bounds for the effective permittivity in random media*, Nuovo Cimento, 38 (1965), pp. 771–782.
[4] M. J. BERAN, *Statistical Continuum Theories*, Interscience, New York, 1968.
[5] D. J. BERGMAN, *Variational bounds on some bulk properties of a two-phase composite material*, Phys. Rev. B, 14 (1976), pp. 1531–1542.
[6] D. J. BERGMAN, *The dielectric constant of a composite material—a problem in classical physics*, Phys. Rep., 43 (1978), pp. 377–407.
[7] D. J. BERGMAN, *Exactly solvable microscopic geometries and rigorous bounds for the complex dielectric constant of a two-component composite material*, Phys. Rev. Lett., 44 (1980), pp. 1285–1287.
[8] D. J. BERGMAN, *Bounds for the complex dielectric constant of a two-component composite material*, Phys. Rev. B, 23 (1981), pp. 3058–3065.
[9] D. J. BERGMAN, *Resonances in the bulk properties of composite media—theory and applications*, in Macroscopic Properties of Disordered Media, Lecture Notes in Phys. 154, R. Burridge, S. Childress, and G. Papanicolaou, eds., Springer-Verlag, Berlin, 1982, pp. 10–37.
[10] D. J. BERGMAN, *Rigorous bounds for the complex dielectric constant of a two component composite*, Ann. Physics, 138 (1982), pp. 78–114.
[11] D. J. BERGMAN, *Hierarchies of Stieltjes functions and their application to the calculation of bounds for the dielectric constant of a two-component composite medium*, SIAM J. Appl. Math., 53 (1993), pp. 915–930.
[12] O. P. BRUNO, *The effective conductivity of strongly heterogeneous composites*, Proc. Roy. Soc. London Ser. A, 433 (1991), pp. 353–381.

[13] E. Cherkaeva, *Inverse homogenization for evaluation of effective properties of a mixture*, Inverse Problems, 17 (2001), pp. 1203–1218.
[14] E. Cherkaeva and K. M. Golden, *Inverse bounds for microstructural parameters of composite media derived from complex permittivity measurements*, Waves Random Media, 8 (1998), pp. 437–450.
[15] E. Cherkaeva and A. C. Tripp, *Inverse conductivity for inaccurate measurements*, Inverse Problems, 12 (1996), pp. 869–883.
[16] A. R. Day and M. F. Thorpe, *The spectral function of random resistor networks*, J. Phys.–Condensed Matter, 8 (1996), pp. 4389–4409.
[17] A. R. Day and M. F. Thorpe, *The spectral function of composites: The inverse problem*, J. Phys.–Condensed Matter, 11 (1999), pp. 2551–2568.
[18] A. R. Day, M. F. Thorpe, A. R. Grant, and A. J. Sievers, *The spectral function of a composite from reflectance data*, Phys. B, 279 (2000), pp. 17–20.
[19] B. R. Djordjević, J. H. Hetherington, and M. F. Thorpe, *Spectral function for a conducting sheet containing circular inclusions*, Phys. Rev. B, 53 (1996), pp. 14862–14871.
[20] C. Engström, *Structural information of composites from complex-valued measured bulk properties*, Inverse Problems, submitted.
[21] C. Engström, *Bounds on the effective tensor and the structural parameters for anisotropic two-phase composite material*, J. Phys. D: Appl. Phys., 38 (2005), pp. 3695–3702.
[22] K. Golden and G. Papanicolaou, *Bounds for effective parameters of heterogeneous media by analytic continuation*, Comm. Math. Phys., 90 (1983), pp. 473–491.
[23] Z. Hashin and S. Shtrikman, *A variational approach to the theory of the effective magnetic permeability of multiphase materials*, J. Appl. Phys., 33 (1962), pp. 3125–3131.
[24] K. Hinsen and B. U. Felderhof, *Dielectric constant of a suspension of uniform spheres*, Phys. Rev. B, 46 (1992), pp. 12955–12963.
[25] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.
[26] J. B. Keller, *A theorem on the conductivity of a composite medium*, J. Math. Phys., 5 (1964), pp. 548–549.
[27] R. Lipton, *Optimal inequalities for gradients of solutions of elliptic equations occurring in two-phase heat conductors*, SIAM J. Math. Anal., 32 (2001), pp. 1081–1093.
[28] K. A. Lurie and A. V. Cherkaev, *Exact estimates of conductivity of composites formed by two isotropically conducting media taken in prescribed proportion*, Proc. Roy. Soc. Edinburgh Sect. A, 99 (1984), pp. 71–87.
[29] H. Ma, B. Zhang, W. Y. Tam, and P. Sheng, *Dielectric-constant evaluation from microstructures*, Phys. Rev. B, 61 (2000), pp. 962–966.
[30] J. C. Maxwell, *A Treatise on Electricity and Magnetism*, Vol. 1, Dover, New York, 1954.
[31] R. C. McPhedran, D. R. McKenzie, and G. W. Milton, *Extraction of structural information from measured transport properties of composites*, Appl. Phys. A, 29 (1982), pp. 19–27.
[32] G. W. Milton, *Bounds on the complex dielectric constant of a composite material*, Appl. Phys. Lett., 37 (1980), pp. 300–302.
[33] G. W. Milton, *Bounds on the transport and optical properties of two-component composite material*, J. Appl. Phys., 52 (1981), pp. 5294–5304.
[34] G. W. Milton, *The Theory of Composites*, Cambridge University Press, Cambridge, UK, 2002.
[35] S. Prager, *Improved variational bounds on some bulk properties of a two-phase random medium*, J. Chem. Phys., 50 (1969), pp. 4305–4312.
[36] K. Schulgasser, *On a phase interchange relationship for composite materials*, J. Math. Phys., 17 (1976), pp. 378–381.
[37] A. K. Sen and S. Torquato, *Effective conductivity of anisotropic two-phase composite media*, Phys. Rev. B, 39 (1989), pp. 4504–4515.
[38] N. R. Silnutzer, *Effective Constants of Statistically Homogeneous Materials*, Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, 1972.
[39] L. Tartar, *Estimations fines de coefficients homoqénéisés*, in Ennio De Giorgi Colloquium, Res. Notes in Math. 125, P. Krée, ed., Pitman, Boston, MA, 1985, pp. 168–187.
[40] S. Torquato, *Effective electrical conductivity of two-phase disordered composite media*, J. Appl. Phys., 58 (1985), pp. 3790–3797.
[41] O. Wiener, *Die Theorie des Mischkörpers für das Feld des stationären Strömung*, Abh. Math.-Physichen Klasse Königl. Sächsh. Gesel. Wissen, 32 (1912), pp. 509–604.

SIAM J. APPL. MATH.
Vol. 67, No. 1, pp. 214–234

COMPUTATIONAL MODELING OF TEXTURE FORMATION AND OPTICAL PERFORMANCE OF LIQUID CRYSTAL FILMS ON PATTERNED SURFACES*

DAE KUN HWANG† AND ALEJANDRO D. REY†

Abstract. An integrated microstructural-optical model based on the tensorial Landau–de Gennes liquid crystal theory, the Matrix–Berreman optical model, and the finite-difference time-domain (FDTD) optical method is used to investigate texture formation and polarized light propagation in thin nematic liquid crystal (NLC) films for various anchoring boundary conditions mimicking surface conditions of an existing liquid crystal (LC)-based biosensor device used to detect biological binding events. The integrated mathematical model of the optical device describes the signal generation process of the biosensor based on LC vision. The FDTD optical method predicts two important optical signatures of the transmitted polarized light: oscillations and nonsymmetric optical signals. However, the approximate Matrix–Berreman optical method cannot predict these important optical responses when strong lateral orientation gradients are present. The model predictions are found to be in good agreement with actual experimental results, and can be used to detect interfacial LC orientation due to bound biomolecules.

Key words. texture formation, FDTD

AMS subject classification. 65A05

DOI. 10.1137/060649045

1. Introduction.

Liquid crystals (LCs) are anisotropic electroptical materials [1] widely used in displays, light valves, and more recently in biosensor applications [1, 2, 3, 4]. This paper presents a mathematical and computational study of uniaxial rod-like low-molar mass nematic liquid crystal (NLC) [5] films for biosensor applications [3, 4, 6, 7].

Liquid crystal films are soft materials where weak substrate forces can alter orientation states and generate topological defects [1, 2]. Furthermore, LC films are thermodynamically stable and possess long-range orientational order and optical anisotropy [5, 8]. Thus LC films exhibit unique optical textures when observed under cross-polars due to spatial heterogeneities of macroscopic orientation [5, 8]; in this paper, orientation refers to the average molecular orientation of the rod-like molecules composing the NLC phase, and it is described by the unit vector or director $\mathbf{n}$; the director $\mathbf{n}$ is the optic axis [5]. The unique combination of soft material, sensitivity to substrate chemistry and geometry, and optical anisotropy provides unique opportunities in the development of LC-based biosensors [3, 4, 6, 7].

Surface treatments that affect physicochemical surface conditions of substrates such as surface topology and chemical compositions lead to defect-free LC structures needed for display applications [9]. One popular method of substrate preparation for obtaining defect-free LC structures is mechanical rubbing of glass plate substrates in one direction to create a sinusoidal topology; the rubbing direction creates an

*Received by the editors January 4, 2006; accepted for publication (in revised form) June 15, 2006; published electronically December 1, 2006. This work is supported by a grant from the Donors of the Petroleum Research Fund (PRF) administered by the American Chemical Society.
http://www.siam.org/journals/siap/67-1/64904.html
†Department of Chemical Engineering, McGill University, 3610 University Street, Montreal, QC, Canada H3A 2B2 (dae.hwang@mail.mcgill.ca, alejandro.rey@mcgill.ca). The first author acknowledges support from Fonds de recherché sur la nature et les technologies of Quebec and the Eugene Lamonthe Fund of the Department of Chemical Engineering, McGill University.

easy-axis $\mathbf{E}$ or preferred orientation [9], and the director $\mathbf{n}$ aligns along $\mathbf{E}$; strong anchoring denotes the state at the surface $\mathbf{n} = \mathbf{E}$. Since desirable LC structures on various substrates have been obtained using many different procedures and physiochemical surface treatments, LCs have been successfully used in display applications. In addition to display applications, Skaife and Abbott have demonstrated another successful application of LCs biosensors in order to detect biological binding events on nano-structured surfaces supporting LC films [3, 6]. The basis of the sensor is the presence of uniform orientation in the absence of biomolecular surface-covering. On the other hand, the presence of surface-bound proteins or viruses modifies the aligning properties of the surface, creating textures or spatial heterogeneities in the optic axis that are easily detectable and quantifiable using light transmission under cross-polars [3, 6]. Thus, surface-bound biomolecules can be detected and their surface density quantified through measurement of the optical output [3, 6]. Denoting by $\mathbf{k}$ the unit surface normal, substrates in contact with NLCs can induce homeotropic (or normal, $\mathbf{E} = \mathbf{k}$), oblique (or tilted, $0 < (\mathbf{E} \cdot \mathbf{k})^2 < 1$), degenerate planar (tangential, $(\mathbf{E} \cdot \mathbf{k}) = 0$), and uniform planar surface orientation ($\mathbf{E} \cdot \mathbf{k} = 0, \mathbf{E} = \mathbf{t}_o$, and $\mathbf{t}_o$ a tangential unit vector) [5, 9]. Recent work by Abbott and coworkers [4, 7], demonstrates that a uniform planar orientation of LCs can be obtained by a special protein deposition on functionalized nano-structured surfaces. The preferred tangential orientation of the surface director ($\mathbf{n} = \mathbf{t}_o$) was then determined by measuring the modulation of transmitted optical intensity upon rotating the LC films with respect to fixed cross-polars [7]. Sample rotation under fixed cross-polars provides a simple and useful way to detect specific planar orientations on protein-covered substrates. The optical principle operating here is based on the fact that maximum light transmittance through cross-polars in a NLC film is obtained when $\mathbf{n}$ is at 45° from the cross-polars. Hence, to find $\mathbf{t}_o$ in a substrate, the sample is rotated under fixed cross-polars until the maximum optical transmittance is found. This paper uses computational optical modeling to predict a uniform planar surface orientation on partially covered substrates, as observed in experiments [4, 7].

Despite the great advantages of the LC-based biosensors [3, 4, 6, 7] and their potential uses, the sensor functionalities and fundamental relationships between optical responses and complex surface-driven LC texture formation processes are not fully understood. The integration of surface-induced texture formation of LCs with its optical responses has not been fully explored. Optical computations of light propagation and texture formation modeling studies considering multiscale phenomena provide a better understanding of the LC-based biosensor functionalities and eventually may lead to simulation-based biosensor design and optimization.

In this paper we simulate texture formation in NLC films using the tensor Landau–de Gennes liquid crystal theory [5] and compute light transmittance through cross-polars using the Matrix–Berreman method and the FDTD method, as described below. The geometry and surface conditions replicate the experiments [4, 7]. We study texture formation of low-molar NLCs in a thin film between two patterned surfaces containing two distinct vertical regions: (i) a nonprinted protein region, with strong planar anchoring (i.e., $\mathbf{n} \cdot \mathbf{k} = \mathbf{0}$) on its lower surface and strong homeotropic anchoring (i.e., $\mathbf{n} \cdot \mathbf{k} = \mathbf{1}$) on its upper surface, and (ii) a printed protein region, with homeotropic anchoring on its upper surface (i.e., $\mathbf{n} \cdot \mathbf{k} = \mathbf{1}$) and one preferred tangential easy-axis (i.e., $\mathbf{n} = \mathbf{t_o}$) on its lower surface where printed protein is present.

Two popular computational optical methods for solving the Maxwell equations have been applied in order to describe light propagation through NLC films: (i) the

Berreman method [10, 11, 12, 13, 14, 15, 16] and (ii) the finite-difference time-domain (FDTD) method [17, 18, 19, 20]. The Berreman method is an approximate matrix-type method based on the stratified approach, and the FDTD method is a direct numerical simulation of the Maxwell equations. Despite the wide use of the Berreman method in computational optical studies for LC displays, this method has one major limitation due to the assumption that variation of the dielectric tensor occurs only in the direction of light propagation. Thus the Berreman method is best suited to one dimensional problems [10, 11, 12, 13, 14, 15, 16]. Thus, the Berreman method may not be appropriate for studying optical responses in the LC-based biosensor, where the dielectric tensor varies over small length scales in multiple directions.

Application of the FDTD method in LC films, especially for LC displays, is relatively recent compared with the Berreman method [17, 18, 19, 20, 21, 22]. In contrast to the Berreman method, detailed optical responses and important optical features of LC structures observed in advanced LC displays are successfully predicted by the FDTD method [19, 20, 21, 22]. However, the performance of the FDTD method in optical studies of textured NLCs with multiscale heterogeneities remains to be explored and quantified. Our previous studies [23, 24, 25] on computational optics on textured NLC films containing wedge, twist, and twist loop defects show that the FDTD method has excellent abilities to accurately compute light transmittance in heterogeneous films. In this paper we extend our previous work [23, 24, 25] and simulate the light transmittance of an NLC thin film with surface and bulk heterogeneities of relevance to biosensors.

The objectives of this paper are the following:

(a) to simulate transient texture formation in NLC thin films with a substrate containing a sequence of nonprinted-protein and printed-protein regions;
(b) to characterize the computational optical responses of the predicted orientation structures using the Berreman and FDTD methods;
(c) to evaluate the FDTD and Berreman methods;
(d) to determine the preferred alignment of NLCs on a printed-protein region in NLC films

The organization of this paper is as follows. Section 2 presents the Landau–de Gennes LC theory and governing equations for describing orientation structures in NLCs in thin films. Section 3 presents the main features of the Berreman and FDTD optical methods. Section 4 presents and discusses the predicted results. Section 5 presents the conclusions.

2. Theory and governing equations.

2.1. Description of orientation and alignment of NLCs. The multiscale description of the microstructure of NLCs is characterized by the second moment of the orientation distribution function, referred to as a second-order symmetric traceless tensor $\mathbf{Q}$ [5], defined as

$$\mathbf{Q} = \mathrm{S}\left(\mathbf{nn} - \frac{1}{3}\delta\right) + \frac{1}{3\mathrm{P}}(\mathbf{mm} - \mathbf{ll}), \tag{2.1}$$

with the following restrictions:

$$\mathbf{Q} = \mathbf{Q}^{\mathrm{T}}, \tag{2.2}$$

$$\mathrm{tr}(\mathbf{Q}) = 0, \tag{2.3}$$

$$-\frac{1}{2} \leq S \leq 1, \tag{2.4}$$

$$-\frac{3}{2} \leq P \leq \frac{3}{2}, \tag{2.5}$$

$$\mathbf{n} \cdot \mathbf{n} = \mathbf{m} \cdot \mathbf{m} = \mathbf{l} \cdot \mathbf{l} = 1, \tag{2.6}$$

$$\mathbf{nn} + \mathbf{mm} + \mathbf{ll} = \delta = \begin{Bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{Bmatrix}, \tag{2.7}$$

where $\mathbf{n}$, $\mathbf{m}$, $\mathbf{l}$ are referred as the uniaxial director and the first and second biaxial directors forming the orthogonal director triad, which defines the macroscopic orientation of NLCs. The scalar order parameters S and P are measures of the molecular alignment; the magnitude of the uniaxial scalar order parameter S is a degree of the molecular alignment along the uniaxial director $\mathbf{n}$, and the magnitude of the biaxial scalar order parameter P is a degree of the molecular alignment along the first biaxial director. On the principal axis, the tensor order parameter $\mathbf{Q}$ is

$$\mathbf{Q} = \begin{pmatrix} -\frac{1}{3}(S-P) & 0 & 0 \\ 0 & -\frac{1}{3}(S+P) & 0 \\ 0 & 0 & \frac{2}{3}S \end{pmatrix}, \tag{2.8}$$

where $S = \frac{3}{2}(\mathbf{n} \cdot \mathbf{Q} \cdot \mathbf{n})$ and $P = \frac{3}{2}(\mathbf{m} \cdot \mathbf{Q} \cdot \mathbf{m} - \mathbf{l} \cdot \mathbf{Q} \cdot \mathbf{l})$. Depending on the values of the parameters S and P, the tensor $\mathbf{Q}$ is able to describe three states: isotropic ($S = 0, P = 0$), uniaxial ($S \neq 0, P = 0$), and biaxial ($S \neq 0, P \neq 0$).

2.2. Landau–de Gennes model for NLCs. According to the Landau–de Gennes model [2, 5, 26, 27], the total free energy density f of NLCs in the absence of external fields is expressed as the sum of three contributions, isotropic (f_i), homogeneous (f_h), and gradient (f_g) contributions:

$$f = f_i\,(T, P) + f_h\,(T, \mathbf{Q}) + f_g\,(T, \mathbf{Q}, \nabla\mathbf{Q})\,, \tag{2.9}$$

where f_i represents the free energy of the isotropic state and is a function of conventional thermodynamic parameters such as temperature and pressure while independent of $\mathbf{Q}$; $f_h\,(T, \mathbf{Q})$ is the homogeneous contribution and captures the isotropic↔ nematic phase transition, given by [2, 5, 26, 27]

$$f_s = \frac{1}{2}a(T - T^*)Q_{\alpha\beta}Q_{\beta\alpha} - \frac{1}{3}bQ_{\alpha\beta}Q_{\beta\gamma}Q_{\gamma\alpha} + \frac{1}{4}c(Q_{\alpha\beta}Q_{\beta\alpha})^2, \tag{2.10}$$

where $\{\alpha, \beta, \gamma, \delta\} = 1, 2, 3$ denote the components along the three orthogonal axes in a Cartesian coordinate system; a, b, and c are constants, and T^* is the isotropic↔ nematic transition temperature.

The gradient f_g term is due to long-range elastic effects and expressed in terms of the gradient of $\mathbf{Q}$ [2, 5, 26, 27]:

$$\begin{aligned} f_g =& \frac{1}{2}L_1\nabla_\alpha Q_{\beta\gamma}\nabla_\alpha Q_{\beta\gamma} + \frac{1}{2}L_2\nabla_\alpha Q_{\alpha\gamma}\nabla_\beta Q_{\beta\gamma} \\ &+ \frac{1}{2}L_3 Q_{\alpha\beta}\nabla_\alpha Q_{\gamma\delta}\nabla_\beta Q_{\gamma\delta}, \end{aligned} \tag{2.11}$$

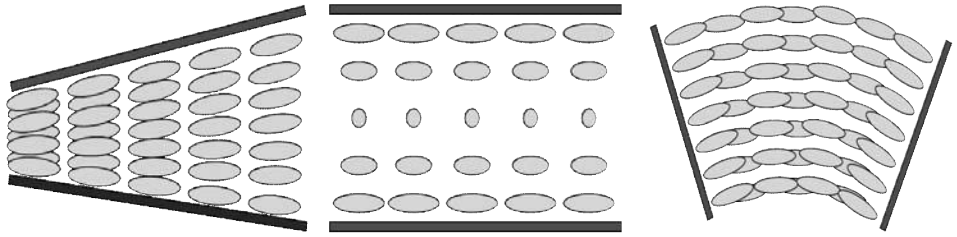


FIG. 2.1. *Schematics of the three elastic deformations of rod-like uniaxial nematics:* (a) *splay mode,* (b) *twist mode, and* (c) *bend mode, corresponding to elastic constants* K_{11}*,* K_{22}*, and* K_{33}*, respectively.*

where L_1, L_2, and L_3 are phenomenological parameters of LCs. To ensure stability, the Landau coefficients L_1, L_2, and L_3 are restricted (under uniaxiality) as follows [2, 5, 26, 27]:

$$2L_1 + L_2 - \frac{2}{3}SL_3 > 0, \quad 2L_1 - \frac{2}{3}SL_3 > 0, \quad 2L_1 + L_2 + \frac{4}{3}SL_3 > 0. \tag{2.12}$$

Under elastic isotropy (one constant approximation), $L_2 = L_3 = 0$; under the constant uniaxial scalar order parameter condition, (2.11) yields the Frank–Oseen gradient energy density [5, 8], based on the director model. The relation between the Landau elastic constants of the tensor model and the Frank elastic constants of the vector model is [28, 29]

$$L_1 = \frac{3K_{22} - K_{11} + K_{33}}{6S^2}, \quad L2 = \frac{K_{11} - K_{22}}{S^2}, \quad L3 = \frac{K_{33} - K_{11}}{2S^3}, \tag{2.13}$$

where K_{11}, K_{22}, K_{33} are the splay, twist, and bend elastic constants, respectively. These elastic modes are shown in Figure 2.1. Planar distortions contain no twist. Nonplanar distortions are trimodal.

Using the Doi model of LCs in conjunction with the Landau–de Gennes model, the dimensionless free energy density equations (2.10) and (2.11) are [26, 30]

$$\widetilde{f_h} = \frac{1}{2}\left(1 - \frac{U}{3}\right) Q_{\alpha\beta}Q_{\beta\alpha} - \frac{U}{3}Q_{\alpha\beta}Q_{\beta\gamma}Q_{\gamma\alpha} + \frac{U}{4}(Q_{\alpha\beta}Q_{\beta\alpha})^2, \tag{2.14}$$

$$\begin{aligned}\widetilde{f}_g = &\left(\frac{\xi}{X}\right)^2 \left(\frac{1}{2}\widetilde{\nabla}_\alpha Q_{\beta\gamma}\widetilde{\nabla}_\alpha Q_{\beta\gamma} + \frac{1}{2}\widetilde{L}_1\widetilde{\nabla}_\alpha Q_{\alpha\gamma}\widetilde{\nabla}_\beta Q_{\beta\gamma}\right.\\ &\left. + \frac{1}{2}\widetilde{L}_2 Q_{\alpha\beta}\widetilde{\nabla}_\alpha Q_{\gamma\delta}\widetilde{\nabla}_\beta Q_{\gamma\delta}\right)\end{aligned} \tag{2.15}$$

$$\widetilde{f} = \frac{f}{\varphi kT}, \quad U = 3\frac{T^*}{T}, \quad a = \varphi k, \quad b = c = \varphi kTU, \quad \xi = \sqrt{\frac{L_1}{\varphi kT}}, \tag{2.16}$$

$$\widetilde{L}_2 = \frac{L_2}{L_1}, \quad \widetilde{L}_3 = \frac{L_3}{L_1}, \quad \widetilde{\nabla} = X\nabla, \tag{2.17}$$

where φ and k are the concentration and Boltzman's constant, respectively. The nematic potential U controls the stability such that: $U < 8/3$ corresponds to the

isotropic phase, $8/3 \leq U \leq 3$ to the biphasic isotropic-nematic equilibrium, and $U > 3$ to the uniaxial nematic phase. The internal length scale ξ is the characteristic scale for changes in S and is of the order of a defect core [5]. X indicates the external length scale of the system, which in this paper is the half length of SPR region (see Figure 4.1).

The microstructure evolution, in the absence of flow, is given by a standard gradient flow dynamic equation [26, 30]:

$$-\gamma(\mathbf{Q})\frac{\partial \mathbf{Q}}{\partial t} == \left(\frac{\delta f}{\delta \mathbf{Q}}\right)^{[s]} = \left(\frac{\partial f}{\partial \mathbf{Q}} - \nabla \cdot \frac{\partial f}{\partial \nabla \mathbf{Q}}\right)^{[s]}, \tag{2.18}$$

where γ is the rotational viscosity and the superscript $[s]$ denotes a symmetric and traceless tensor. Substituting (2.14) and (2.15) into (2.18) yields the dynamical equation for $\mathbf{Q}$:

$$\begin{aligned} -\frac{\partial \mathrm{Q}_{ij}}{\partial \widetilde{t}} &= \left(1 - \frac{U}{3}\right) Q_{ij} - U\left(Q_{i\alpha}Q_{\alpha j} - \frac{1}{3Q_{\beta\alpha}Q_{\alpha\beta}\delta_{ij}}\right) + UQ_{\alpha\beta}Q_{\beta\alpha}Q_{ij} \\ &\quad - R\widetilde{\nabla}_k\widetilde{\nabla}_k Q_{ij} - R\tilde{L}_2\left(\frac{1}{2}(\widetilde{\nabla}_i\widetilde{\nabla}_\alpha Q_{\alpha j} + \widetilde{\nabla}_j\widetilde{\nabla}_\alpha Q_{\alpha i}) - \frac{1}{3}\widetilde{\nabla}_\beta\widetilde{\nabla}_\alpha Q_{\alpha\beta}\delta_{ij}\right) \\ &\quad - R\tilde{L}_3(\widetilde{\nabla}_k Q_{\alpha k}\widetilde{\nabla}_\alpha Q_{ij} + Q_{\alpha k}\widetilde{\nabla}_k\widetilde{\nabla}_\alpha Q_{ij}) \\ &\quad + R\tilde{L}_3\left(\widetilde{\nabla}_i Q_{\gamma\delta}\widetilde{\nabla}_j Q_{\gamma\delta} - \frac{1}{3\widetilde{\nabla}_\alpha Q_{\gamma\delta}\widetilde{\nabla}_\alpha Q_{\gamma\delta}\delta_{ij}}\right), \end{aligned} \tag{2.19}$$

where $\widetilde{t} = \varphi k T^* t/\gamma$ is the dimensionless time and

$$R = \left(\frac{\xi}{X}\right)^2 \tag{2.20}$$

is the square of the internal/external length scale ratio. The dimensionless numbers $\wp$ that control the dynamics of $\mathbf{Q}$ are

$$\wp : \left\{U, R, \tilde{L}_2, \tilde{L}_3\right\}. \tag{2.21}$$

As mentioned above, U controls the stability, and for a stable homogenous uniaxial nematic phase the relation between U and the equilibrium scalar order parameter S_{eq} is

$$S_{eq} = \frac{1}{4} + \frac{3}{4}\sqrt{1 - \frac{8}{3U}}. \tag{2.22}$$

For low molar mass NLCs, we estimate $\xi = 10$–20nm [5, 31], and for films in the micron range this gives $X/\xi \approx 100$. Hence using this model to simulate realistic materials and realistic geometries gives rise to a PDE system with a small parameter. This small parameter is responsible for the ability of the model to capture topological defects and changes in the scalar order parameter close to bounding surfaces [32, 33]. The $R = 0$ limit of tensor equation (2.20), with $S = S_{eq}$ and $P = 0$, yields the dynamic version of the Frank–Oseen director model [5]. Lastly, $\tilde{L}_2$ and $\tilde{L}_3$ denote elastic anisotropy in the system. For low molar mass nematics, such as 5CB (4-pentyl-4'-cyanobiphenly) the values for these dimensionless elastic constants are $\tilde{L}_2 = 0.85, \tilde{L}_3 = 0.87$ [34, 35]. In the material system under study here, elastic anisotropy is not a significant effect.

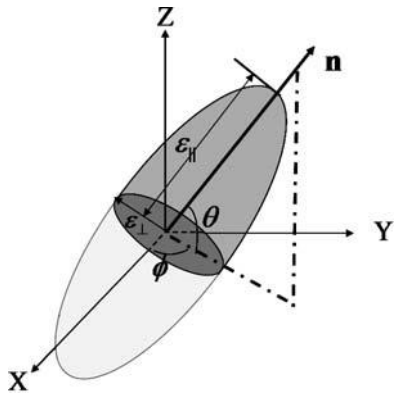


FIG. 3.1. *Schematic diagram of the optical component of the dielectric tensor ε, ordinary dielectric constant $\varepsilon_{\perp}$ and the extraordinary dielectric constant $\varepsilon_{\parallel}$, with respect to the optic axis $\mathbf{n}$ described by the two Euler angles, azimuthal ϕ and polar θ, for a uniaxial rod-like NLCs in a rectangular (x,y,z) coordinate system. The unit vector $\mathbf{n}$ also represents the local director.*

3. Optical modeling of LCs. NLCs are optically transparent and anisotropic materials. These properties provide unique and interesting optical features of light propagation through NLC films. The Matrix–Berreman and FDTD method have been successfully applied to describe classical defects observed in NLCs such as wedge, twist, and loop defects [23, 24, 25]. In this section, we discuss the main features of the two optical methods of direct relevance to the objectives of the paper; other details are found in various references and texts [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 36].

The Maxwell curl equations and constitutive equations in the absence of current and for nonmagnetic materials are given by

$$\frac{\partial \mathbf{D}}{\partial t} = \nabla \times \mathbf{H}, \quad \frac{\partial \mathbf{B}}{\partial t} = \nabla \times \mathbf{E}, \quad \mathbf{D} = \varepsilon \mathbf{E}, \quad \mathbf{B} = \mu_o \mathbf{H}, \tag{3.1}$$

where $\mathbf{D}$ is the electric flux, $\mathbf{E}$ and $\mathbf{H}$ are the electric and magnetic fields, and μ_o is the magnetic permeability in a vacuum. Figure 3.1 shows the optic axis, two dielectric constants, and Euler angles of rod-like uniaxial NLCs. The optically anisotropic properties of NLCs are included in the dielectric tensor ε. The Euler azimuthal ϕ and polar angle θ describe the direction of the optic axis $\mathbf{n}$ for rod-like uniaxial NLCs in a rectangular coordinate system. The optic axis coincides with the local director $\mathbf{n}$ of uniaxial NLCs. The six components of the symmetric dielectric tensor ε for uniaxial NLCs in two dimensions are given by

$$\varepsilon(x,z) = \begin{pmatrix} \varepsilon_{\perp} + \Delta\varepsilon \cos^2\theta \cos^2\phi & \Delta\varepsilon \cos^2\theta \sin\phi \cos\phi & \Delta\varepsilon \sin\theta \cos\theta \cos\phi \\ \Delta\varepsilon \cos^2\theta \sin\phi \cos\phi & \varepsilon_{\perp} + \Delta\varepsilon \cos^2\theta \sin^2\phi & \Delta\varepsilon \sin\theta \cos\theta \sin\phi \\ \Delta\varepsilon \sin\theta \cos\theta \cos\phi & \Delta\varepsilon \sin\theta \cos\theta \sin\phi & \varepsilon_{\perp} + \Delta\varepsilon \sin^2\theta \end{pmatrix}, \tag{3.2}$$

where $\Delta\varepsilon = \varepsilon_{\perp} - \varepsilon_{\parallel}$, $\varepsilon_{\perp}$ is the ordinary and $\varepsilon_{\parallel}$ the extraordinary dielectric constant. The Maxwell equations (3.1) in two dimensions can be decoupled into two separated waves, ordinary (transverse magnetic) and extraordinary (transverse electric), known as TM and TE mode fields, under the condition that the dielectric tensor is dependent only on one of Euler angles. For example, LCs exhibit only planar structures, $\theta = 0$; otherwise, the decoupling of the Maxwell equations into TM and TE fields is not feasible [20].

3.1. Berreman method [10, 11, 12, 13, 14, 15]. Reformulation of the Maxwell equations (3.1) into four linear differential equations is possible under the assumption that variation of the dielectric tensor occurs only in the light propagation direction and is negligible in the transverse directions, where the Euler angles are only functions of z or the spatial gradient of the Euler angles with respect to x is moderate in a long range. The four linear differential equations are given by [10]

$$\frac{d\psi}{dz} = -i\frac{\omega}{c}\Delta(z)\psi, \qquad \psi = (E_x, H_y, E_y, -H_x)^T, \tag{3.3}$$

where c is the light propagation velocity in vacuum, ω is the angular frequency, and E_x, E_y, H_x, and H_y are electric and magnetic components. Optical properties of NLCs, polarizations due to local molecular orientation of NLCs, and multiple reflections due to the presence of different media are imposed in the 4×4 matrix $\Delta(z)$. Thus, the space-dependent $\Delta(z)$ are mainly functions of $\varepsilon(z)$.

The global computational domain of the NLC film including substrates is divided into local cubic lattices. Each local lattice is assumed to be a homogenous medium whose dielectric tensor is uniform ε_i; the subscript i indicates the local cubic lattice. Solution vectors for transmitted waves ψ_t and reflected waves ψ_r after the incident waves ψ_i travel through the global computational domain from z_0 to z_n are obtained by solving the linear equation (3.3). The solution vectors are obtained as follows:

$$\psi_{\mathbf{t}}(z_n) = \mathbf{F}(i,n)(\psi_{\mathbf{i}}(z_o) + \psi_{\mathbf{r}}(z_o)), \tag{3.4}$$

$$\mathbf{F}(i,n) = \mathbf{p}_{i+n}(\varepsilon_{i+n}(h))\mathbf{p}_{i+n-1}(\varepsilon_{i+n}(h))\dots\mathbf{p}_{i+1}(\varepsilon_{i+1}(h))\mathbf{p}_i(\varepsilon_i(h)), \tag{3.5}$$

$$\mathbf{p}_i(h) = \exp\left[-i\left(\frac{\omega}{c}\right)\Delta h\right], \tag{3.6}$$

where $\mathbf{F}$ is the global transfer matrix, $\mathbf{p}_i$ is the local transfer matrix for each local cubic lattice i, and h is the thickness of each local lattice. $\mathbf{F}$ is just the multiplication of each local transfer matrix $\mathbf{p}_i$. Thus, the main computational cost and challenge of the Berreman method is to obtain solutions of the local transfer matrices. The exponent matrix $\mathbf{p}_i$ can be expressed using a Taylor series as follows:

$$\begin{aligned}\mathbf{p}_i(h) &= \exp\left(-i\frac{\omega}{c}\mathbf{\Delta}h\right)\\ &= \mathbf{I} + \left(-i\frac{\omega h}{c}\mathbf{\Delta}\right) + \frac{1}{2!}\left(-i\frac{\omega h}{c}\right)^2\mathbf{\Delta}^2 + \frac{1}{3!}\left(-i\frac{\omega h}{c}\right)^3\mathbf{\Delta}^3 + \cdots,\end{aligned} \tag{3.7}$$

where $\mathbf{I}$ is the unit matrix. The higher term of (3.7) can be neglected in the case when the thickness h of each local lattice is sufficiently small and thus each lattice is a uniform medium with $\varepsilon_{i+n}(h)$. In addition, analytical expressions of the local matrix $\mathbf{p}_i$ can be obtained in case the optic axis has a planar structure where the dielectric tensor is dependent only on a single Euler angle. The detailed expression of analytical solutions of the local matrix $\mathbf{P}$ can be found in the literature [14, 15, 16]. In this study, the dielectric tensor is a function of both Euler angles, azimuthal ϕ and polar angle θ. The local transfer matrix is obtained using (3.7) with the assistance of the built-in function called "expm" in Matlab 7 [37].

In this paper we consider a monochromatic incident wave, and the solution vectors are obtained using (3.5) after considering transmission through ideal crossed polarizers.

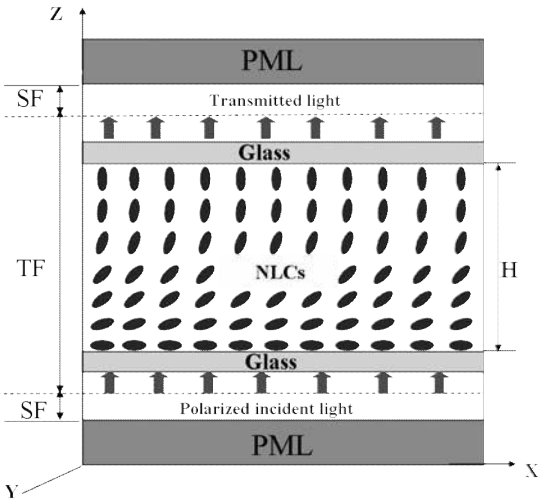


FIG. 3.2. *Schematic of the computational domain used in the FDTD method. H indicates the thickness of an NLC film, and the arrows indicate the normal propagation direction of polarized light. The polarizer and analyzer are placed along the x and y directions. PMLs are placed in the z direction, and periodic conditions are used in the x direction. Oval shapes indicate the average LC molecules' orientation, known as director* $\mathbf{n}$.

3.2. FDTD method [17, 18, 19, 20, 21, 22, 36]. The Maxwell equations (3.1), rewritten in a rescaled form using $\widetilde{\mathbf{D}} = \frac{1}{\sqrt{\varepsilon_o \mu_o}}\mathbf{D}$ and $\widetilde{\mathbf{E}} = \sqrt{\frac{\varepsilon_o}{\mu_o}}\mathbf{E}$, are given by [38]

$$\frac{\partial \widetilde{\mathbf{D}}}{\partial t} = \frac{1}{\sqrt{\varepsilon_o \mu_o}} \nabla \times \mathbf{H}, \quad \widetilde{\mathbf{D}} = \varepsilon^* \widetilde{\mathbf{E}}, \quad \frac{\partial \mathbf{H}}{\partial \mathbf{t}} = -\frac{\mathbf{1}}{\sqrt{\varepsilon_\mathbf{o} \mu_\mathbf{o}}} \nabla \times \widetilde{\mathbf{E}}. \tag{3.8}$$

One of the main challenges and important issues of the FDTD method in solving the Maxwell equations is to implement boundary layer conditions in order to truncate the computational space. Any artificial and nonphysical reflections of outgoing waves, which arise from truncation of the computational space, back into the domain of interest causes contamination of solutions. Thus, it is highly desirable to prevent outgoing waves leaving the domain of interest from reflecting back into the domain. Berenger [39] first introduced a perfectly matched layer (PML) as boundary layers in order to truncate the computational space and to absorb waves leaving the computational domain of interest and entering into the PML without any reflections. However, some difficulties arise in implementing the PML in the case of dielectric anisotropic media and high dimensions. Simpler and more effective formulations of material-independent PML have been developed and successfully implemented in LC application with high performance [38, 40, 41, 42]. Among the formulations, a simplified formulation of the PML called *unsplit PML* is used in this study [38]. One of the benefits of rescaling the Maxwell equations as shown in (3.8) is that it provides a simpler implementation of the unsplit PML method, regardless of level of the complexity in the optical properties of the medium [38]. Figure 3.2 shows a schematic of the computation domain and the three main components of the optical system: (1) the thin NLC film, (2) two supporting glass substrates, and (3) two PML layers used in the FDTD method. The PML layers are used in order to truncate the computational

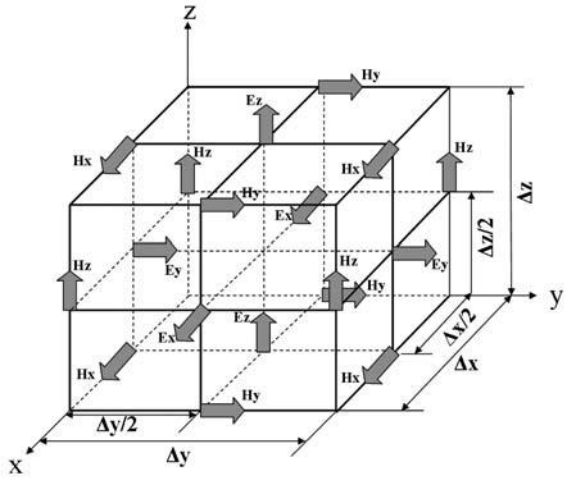


FIG. 3.3. *Schematic of a cubic unit lattice of the Yee cell.*

domain in the z direction. Lateral periodic boundary conditions are used in order to truncate the computational domain in the x direction. The periodic boundary conditions in the x direction are justified when assuming repetition of the system in the x direction. The entire computational domain is subdivided into staggered cubic lattices, and the lattices follow a Yee cell configuration [43]. The Yee cell shown in Figure 3.3 consists of electrical and magnetic components in a staggered lattice in which the electric and magnetic components are located by the half space, so that each **E** and **H** component is surrounded by the four circulating **H** and **E** components. A second-order central finite difference is used to discretize the normalized (3.8) in both space and time on the basis of the Yee algorithm and a fully explicit leap frog scheme, respectively [36]. One example of the finite-difference expression of Dy and Hy components in the rectangular (x, z) coordinate system is given by

$$Dy|_{x,z}^{n+1} = Dx|_{x,z}^{n} + \frac{\Delta t}{\sqrt{\varepsilon_o \mu_o}} \left(\frac{Hx|_{x,z+1/2}^{n+1/2} - Hx|_{x,z-1/2}^{n+1/2}}{\Delta z} - \frac{Hz|_{x+1/2,z}^{n+1/2} - Hz|_{x-1/2,z}^{n+1/2}}{\Delta x} \right), \tag{3.9}$$

$$\mathbf{E}|_{x,z}^{n+1} = \varepsilon^*(x,z)^{-1}\, \mathbf{D}|_{x,z}^{n+1}\,, \tag{3.10}$$

$$Hy|_{x+1/2,z+1/2}^{n+3/2} = Hy|_{x+1/2,z+1/2}^{n+1/2} + \frac{\Delta t}{\sqrt{\varepsilon_o \mu_o}} \left(\frac{Ez|_{x+1,z+1/2}^{n+1} - Ez|_{x,z+1/2}^{n+1}}{\Delta x} - \frac{Ex|_{x+1/2,z+1}^{n+1} - Ex|_{x+1/2,z}^{n+1}}{\Delta z} \right), \tag{3.11}$$

where Δx, Δz and Δt indicate space and time increments, respectively, and n indicates the time step. The **D** field is advanced in time by the half time increment based on the previous advanced **H** field according to (3.9). Then, the **E** field is obtained with the inverse of the dielectric tensor and the previously obtained **D** field. The updated **E** field is used to obtain a new **H** field according to (3.11). This procedure continues

until an initial transient period vanishes and the steady-solution vectors **E** and **H** are obtained, after considering transmission through the ideal crossed polarizers.

In the initial time step, a known monochromatic incident wave is introduced into the computation domain using the total field and scattered field (TF/SF) formulation at the lower interface between the TF/SF regions, as shown in Figure 3.2. The TF/SF formulation at the lower interface is expressed by [36]

$$\mathbf{E_{total}} = \mathbf{E_{inc}} + \mathbf{E_{scat}}, \quad \mathbf{H_{total}} = \mathbf{H_{inc}} + \mathbf{H_{scat}} \quad \text{for total field region,} \tag{3.12}$$

$$\mathbf{E_{scat}} = \mathbf{E_{total}} - \mathbf{E_{inc}}, \quad \mathbf{H_{scat}} = \mathbf{H_{total}} - \mathbf{H_{inc}} \quad \text{for scattered field region,} \tag{3.13}$$

where $\mathbf{E}_{inc}$ indicates the incident electric components and $\mathbf{E}_{scat}$ indicates the electric component due to reflection, scattering, and retardation of incident waves induced by the presence of different media and spatial variations in the NLC orientation. The electric and magnetic fields are advanced in time and space according to (3.9) and (3.11). Only scattered electric and magnetic components are allowed to leave from the total region and enter into the scattered region. Then the scattered waves enter into PML layers, which have fictitious exponential conductivities $\sigma(z)$. In the PML layers, all of the $\mathbf{E}_{scat}$ and $\mathbf{H}_{scat}$ are absorbed exponentially, regardless of any propagation directions of the entering scattered waves. Therefore, no reflections of the waves into the computational domain of interest occur.

4. Results and discussion. In this section, we present details of computational issues and numerical results of texture formation in a NLC thin film on a patterned surface based on the Landau–de Gennes theory and its optical responses using the FDTD and Berreman methods. The optical results are validated using the experimental data of [7].

4.1. Texture formation. In the present paper, we are interested in computing texture formation in a low-molar mass NLC film such as 5CB, used in biosensor applications [3, 4, 6, 7], where the elastic anisotropy effects on the director field conformation are moderate. Hence a one-elastic-constant approximation ($L_2 = L_3 = 0$) is used.

Figure 4.1 shows a schematic of the computational geometry of a thin NLC film used for computing texture formation. The two dimensional computational domain

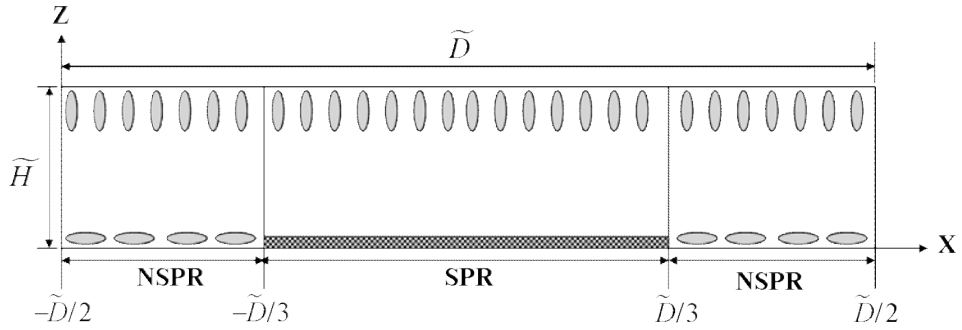


FIG. 4.1. *Schematic of the computational geometry of LC film.* $\widetilde{H}$ *and* $\widetilde{D}$ *indicate dimensionless thickness and length of the film, respectively. The NSPR includes the region* $-1.5 \le x \le -1$ *and* $1 \le x \le 1.5$ *for both* $z = -0.2$ *and* $z = 0.2$*, and the SPR covers* $-1 < x < 1$*,* $z = -0.2$*. Oval shapes on the surface represent fixed anchoring boundary conditions. The check-patterned box indicates stamped proteins where various boundary conditions are used.*

of the thin NLC film shown in Figure 4.1 consists of the area enclosed by the external boundaries, of dimensionless thickness $\widetilde{H} = 0.4$ and dimensionless length $\widetilde{D} = 3$. The computational domain also consists of three distinct regions: two nonstamped protein regions (NSPR) ($-\widetilde{D}/2 \le x \le -\widetilde{D}/3,\ \widetilde{D}/2 \le x \le \widetilde{D}/3$) and one stamped protein region (SPR) ($-\widetilde{D}/3 < x < \widetilde{D}/3$) in Figure 4.1. The two dimensional computational domain replicates the experimental geometry used in [4, 7], so as to keep the same thickness-to-length ratio, the same NSPR-to-SPR length ratio, and same boundary conditions as in the experiments of [4, 7].

Periodic boundary conditions are employed in the x direction, which are consistent with the lateral boundary conditions used for optical modeling. The tensor order parameter $\mathbf{Q}(x, z, t)$ on the boundary ($x = -\widetilde{D}/2, \widetilde{D}/2$) is given by

$$\begin{aligned} &\mathbf{Q}|_{x=-\widetilde{D}/2} = \mathbf{Q}|_{x=\widetilde{D}/2}, \qquad \left.\frac{\partial \mathbf{Q}}{\partial x}\right|_{x=-\widetilde{D}/2} = \left.\frac{\partial \mathbf{Q}}{\partial x}\right|_{x=\widetilde{D}/2}, \\ &\left.\frac{\partial \mathbf{Q}}{\partial y}\right|_{x=-\widetilde{D}/2} = \left.\frac{\partial \mathbf{Q}}{\partial y}\right|_{x=\widetilde{D}/2}. \end{aligned} \tag{4.1}$$

For the remaining boundaries in the z direction, the Dirichlet boundary conditions are implemented, as shown in Figure 4.1. The upper surface ($z = \widetilde{H}/2$) provides a strong homeotropic boundary condition in both SPR and NPR regions; the lower surface ($z = -\widetilde{H}/2$) in the NSPR region provides a strong planar boundary condition parallel to the x direction. Hence, the tensor order parameter $\mathbf{Q}$ on these boundaries is given by

$$\mathbf{Q}_b = S_{eq}\left(\mathbf{n}_b\mathbf{n}_b - \frac{1}{3}\delta\right), \tag{4.2}$$

$$\mathbf{n}_b = (0, 0, 1), \qquad z = \frac{\widetilde{H}}{2}, \quad -\frac{\widetilde{D}}{2} < x < \frac{\widetilde{D}}{2}, \tag{4.3}$$

$$\mathbf{n}_b = (1, 0, 0), \qquad z = -\frac{\widetilde{H}}{2}, \quad -\frac{\widetilde{D}}{2} \le x \le \frac{\widetilde{D}}{3}, \quad \frac{\widetilde{D}}{3} \le x \le \frac{\widetilde{D}}{2}, \tag{4.4}$$

where $\mathbf{n}_b$ is the prescribed uniaxial director or optic axis at $\tilde{z} = \pm\tilde{H}/2$. Since NLCs on the SPR lower surface exhibit a preferred orientation, which we are interested in capturing, we vary boundary conditions on the surface with a strong planar assumption as follows:

$$(1, 0, 0) \le \mathbf{n}_b \le (0, 1, 0), \qquad z = -\frac{\widetilde{H}}{2}, \quad -\frac{\widetilde{D}}{3} \le x \le -\frac{\widetilde{D}}{3}. \tag{4.5}$$

The surface director Euler angles are

$$\theta = 0, \quad 0 \le \phi \le \frac{\pi}{2}, \quad z = -\frac{\widetilde{H}}{2}, \quad -\frac{\widetilde{D}}{3} \le x \le -\frac{\widetilde{D}}{3}. \tag{4.6}$$

The system is initially quenched from isotropic state, and the initial conditions are

$$\begin{aligned} &\mathbf{Q}_{in}(x, z, \widetilde{t} = 0) = S_{in}\left(\mathbf{nn} - \frac{1}{3}\delta\right), \qquad S_{in} \approx 0, \\ &-\frac{\widetilde{H}}{2} < z < \frac{\widetilde{H}}{2}, \qquad -\frac{\widetilde{D}}{2} < x < \frac{\widetilde{D}}{2}. \end{aligned} \tag{4.7}$$

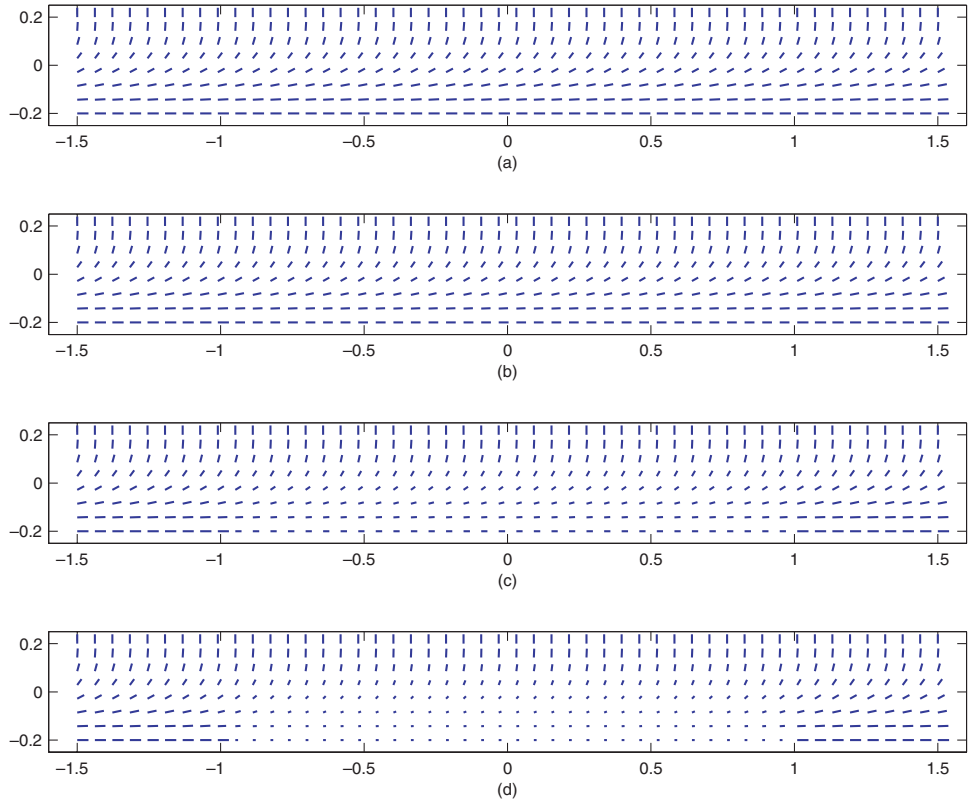


FIG. 4.2. *Computed director profiles of the liquid crystal films for the anchoring boundary conditions on the lower surface of the SPR:* (a) $\phi = 0$, (b) $\phi = \pi/12$, (c) $\phi = \pi/6$, *and* (d) $\phi = \pi/4$. *The values of the axes of the x-z plane are equal to the dimensionless length and thickness.*

The nematic potential $U = 3\frac{T^*}{T}$ is set to be 3.5, which corresponds to S_{eq}=0.615. The spatio-temporal behavior of $\mathbf{n}$ is predicted by solving (2.18), subject to (4.5)–(4.7) for various different anchoring conditions on the lower surface of the SPR, according to (4.5), (4.6).

Figure 4.2 shows computed visualizations of the steady-state director profiles corresponding to the following director boundary conditions on the SPR lower surface region: (a) $\phi = 0$, (b) $\phi = \pi/12$, (c) $\phi = \pi/6$, and (d) $\phi = \pi/4$, respectively, for $\frac{\xi}{H}$=0.01. The NLC structure in Figure 4.2(a) exhibits hybrid alignment and a planar director field with splay and bend distortions. Figures 4.2(b–d) show a nonplanar director field with splay-twist-bend modes. The twist mode occurs in a stripe around the center region ($x \approx 0$). The splay-bend-twist distortions are concentrated close to the interface between the adjacent NSPR and SPR regions.

Figure 4.3 shows computed director profiles corresponding to increasingly larger values of the anchoring boundary conditions: (a) $\phi = \pi/3$, (b) $\phi = \pi/2.57$, (c) $\phi = \pi/2.25$, and (d) $\phi = \pi/2$. As expected, in the vicinity of the interface between two adjacent NSPR and SPR regions, deformation of the NLCs increases with increases in the twist angle ϕ.

In partial summary, as the twist angle ϕ in the SPR increases, the steady-state texture evolves from a diffuse planar splay-bend mode to a sharp nonplanar splay-

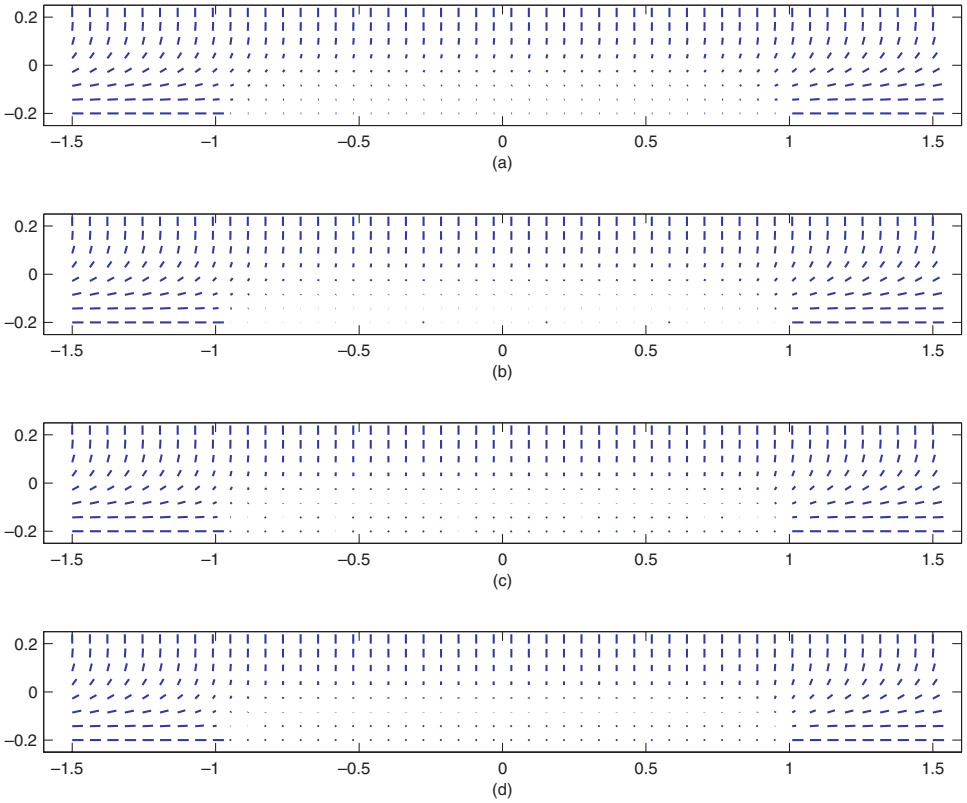


FIG. 4.3. *Computed director profiles of the LC films for the anchoring boundary conditions on the lower surface of the SPR:* (a) $\phi = \pi/3$, (b) $\phi = \pi/2.57$, (c) $\phi = \pi/2.25$, *and* (d) $\phi = \pi/2$. *The values of the axes of the x-z plane are equal to the dimensionless length and thickness.*

twist-bend mode, exhibiting a box-like region over the SPR with uniform escape into the third dimension ("y" axis). The NLC director profiles shown in Figures 4.2 and 4.3, as well as results from other anchoring conditions (not shown for brevity), are used for optical texture modeling, shown next.

4.2. Optical texture formation. The FDTD and Berreman methods for optical computation are applied to predict optical response of the obtained NLC textures. The computational domain for the NLC structure consists of 577×77 rectangular cubic lattices for the Berreman method, and rectangular Yee cells for the FDTD method. A grid size of each cell is set to $\lambda/30$, which is selected in order to prevent numerical dispersion for a mean refractive index close to 1.6. The selected grid size is $\Delta z = \Delta x = 20$nm, and the dimensionless time step is equal to $\Delta t = 3.3356 \times 10^{-17}$, which is selected for numerical stability. Extra cubic lattices for the supporting glass layers in the Berreman method and extra Yee cells are added for the glass layers and the PML layers in the FDTD along the light propagation direction. A linearly polarized monochromatic plane wave along the x direction with free wavelength $\lambda = 600$ is introduced into the computational domain. Only normal incidence is considered. The refractive indices of 5CB are equal to 1.71 for the extraordinary index, no, and 1.53 for the ordinary index, ne [44]. These values correspond to the selected wavelength $\lambda = 600$. The supporting refractive index is equal to 1.52.

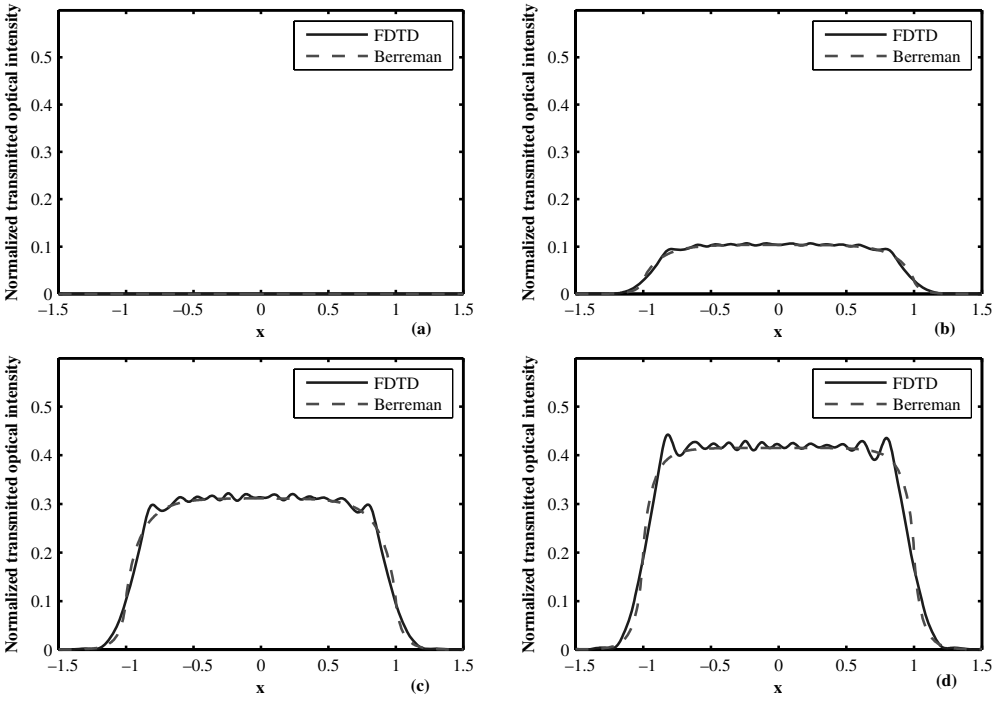


FIG. 4.4. *Normalized optical intensities of transmitted polarized light using the FDTD and Berreman method in the case of* (a) $\phi = 0$, (b) $\phi = \pi/12$, (c) $\phi = \pi/6$, *and* (d) $\phi = \pi/4$ *anchoring boundary conditions corresponding to NLS structures in the corresponding panels of Figure* 4.2. *The normalized intensities are a function of distance along the x direction.*

Figure 4.4 shows the normalized optical intensity of transmitted light through the NLC textures (see Figure 4.2 for the following anchoring conditions in the SPR: (a) $\phi = 0$, (b) $\phi = \pi/12$, (c) $\phi = \pi/6$, and (d) $\phi = \pi/4$, computed using the FDTD and Berreman methods). The optical intensities are plotted as a function of distance in the x direction after considering the presence of the ideal analyzer parallel to the y direction. For $\phi = \pi, 0$, as shown in Figure 4.4(a), the transmitted light from the NLC film is completely extinguished by the analyzer placed along the y direction, so that zero magnitude of the transmitted light is predicted in this case, and the NLC film appears completely dark. This is expected because there is no deviation of the azimuthal angle with respect to the polarizer placed parallel to x-axis; when the optic axis of NLCs uniformly aligns parallel or perpendicular to either polarizer or analyzer (here we consider cross-polars), the NLC film appears dark. Hence, as the surface azimuthal angle deviates from the polarizer in the SPR region (which means the deviation of NLC orientation with respect to the polarizer increases), the intensity of the transmitted light increases in the SPR, but in the NSPR region the intensity vanishes, as shown in Figure 4.2(b–c). Consequently, the maximum optical intensity is predicted by both optical methods when $\phi = \pi/4$ (Figure 4.4(d)).

In weakly heterogeneous textures, the FDTD and Berreman methods predict similar transmitted light intensity magnitudes under cross-polars. However, disagreement between the two methods for sharply textured LC films is significant, as shown in Figure 4.2. This disagreement originates from the optical intensity oscillations predicted by the FDTD method, due to scattering effects from lateral optic axis gradients

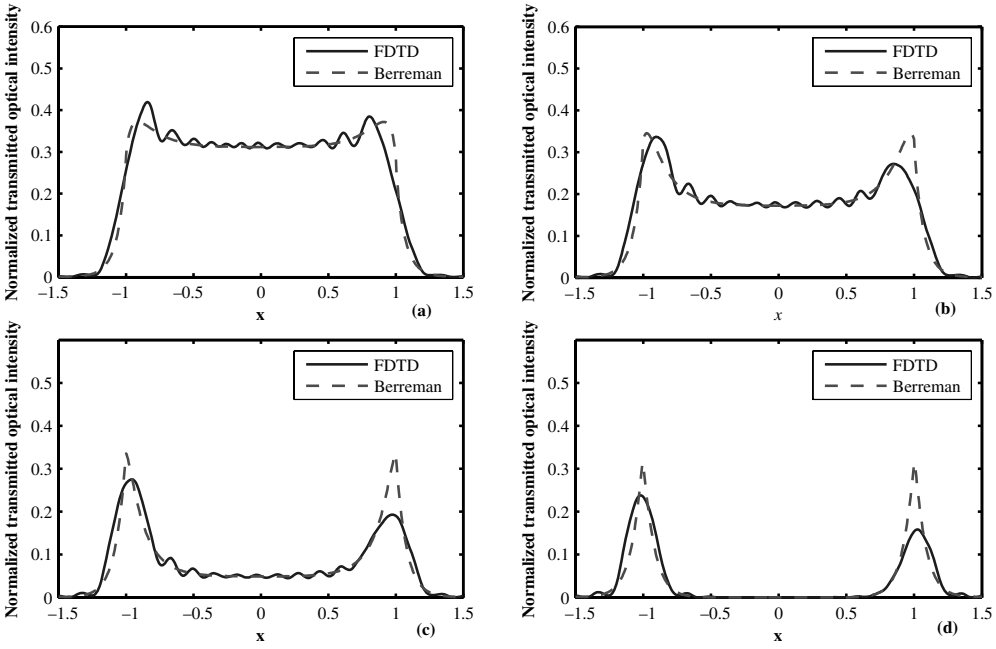


FIG. 4.5. *Normalized optical intensities of transmitted polarized light using the FDTD and Berreman method in the case of* (a) $\phi = \pi/3$, (b) $\phi = \pi/2.57$, (c) $\phi = \pi/2.25$, *and* (d) $\phi = \pi/2$ *anchoring boundary conditions corresponding to NLS structures in the corresponding panels of Figure* 4.3. *The normalized intensities are a function of distance along the x direction.*

$(\partial \mathbf{n}/\partial \mathrm{x})$ in the adjacent two interfaces regions between the NSPR and SPR. On the other hand, the Berreman method fails to capture this effect because lateral gradients in the optic axis are not taken into account [20, 21, 22, 23, 24, 25]. From this result, it is to be expected that the magnitude of deviation between the two optical methods will increase as lateral gradients $(\partial \mathbf{n}/\partial x)$ increase.

Figure 4.5 shows the normalized optical intensity of transmitted light through the NLC textures (see Figure 4.3), computed using the FDTD and Berreman methods, for the following anchoring conditions in the SPR; (a) $\phi = \pi/3$, (b) $\phi = \pi/2.57$, (c) $\phi = \pi/2.25$, and (d) $\phi = \pi/2$. Once again, the intensities of the transmitted light in the SPR region continually decay as the azimuthal angle increases above$\phi = \pi/4$, as shown in Figure 4.5. On the other hand, the degree of the disagreement in the optical intensities between both methods continually increases, due to an increase in the lateral gradient of the optic axis near the interface between the NSPR and SPR. The results show that the maximum level of the disagreement corresponds to $\phi = \pi/2$ (Figure 4.5(d)), where two surface defects are present on the two interfaces between the NSPR and SPR. An additional difference between the two predicted outputs is found by considering the symmetry properties of optical signals. FDTD always predicts asymmetric optical signals, while Berreman's method predicts a symmetric optical signal:

$$\begin{aligned} &\text{Berreman method:} \quad B(x) = B(-x), \\ &\text{FDTD method:} \quad F(x) \neq F(-x), \end{aligned}$$

where $B(x)$ is the Berreman optical signal and $F(x)$ is the FDTD signal, shown by Figures 4.5(a–d).

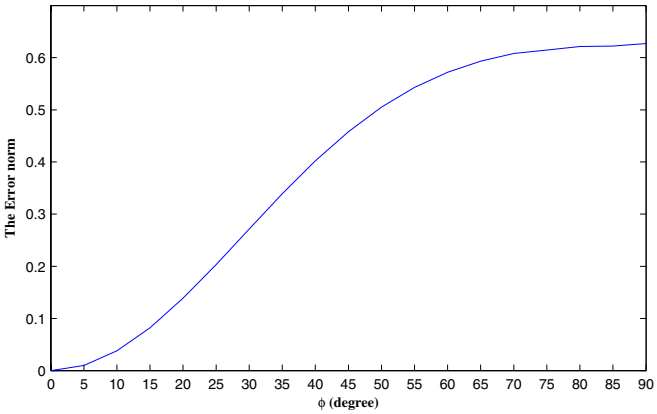


FIG. 4.6. *Error norm $\|E(\phi)\|$ based on the difference of the optical response between the two optical methods, FDTD and Berreman, as a function of ϕ azimuthal anchoring conditions.*

The norm of the asymmetry function $|F(x) - F(-x)|$ increases with increases in the lateral gradient orientation, and for $\phi = \pi/2$, when two surface defects are present, the two peaks of the optical signal differ by 0.1, as shown in Figure 4.5(d). This asymmetric optical feature is due to nonsymmetric orientation gradients along the x direction. The asymmetric optical feature due to nonsymmetric orientation gradients has been reported in the literature [20, 21, 22]. Next we assess the difference between optical signals predicted by the two methods as a function of the azimuthal angle ϕ on the lower surface containing the SPR, by introducing the error norm $\|E(\phi)\|$:

$$\|E(\phi)\| = \sqrt{\sum_{i=1}^{m} |\mathrm{F}(i) - \mathrm{B}(i)|^2}, \tag{4.8}$$

where **F** and **B** are the optical vector solutions of the FDTD and Berreman, respectively; i is the discretized location along the x dimension; and m is equal to 577.

Figure 4.6 shows the error norm $\|E(\phi)\|$ as a function of the azimuthal angle. The error norm $\|E(\phi)\|$ initially increases exponentially with ϕ and eventually reaches a maximum at $\phi = \pi/2$, where there is maximum in the lateral gradient $\left|\frac{\partial \mathbf{n}}{\partial x}\right|$. The major contribution of the deviations between the two optical methods arises from the interfaces between the NSPR and SPR, where strong variations of the lateral gradients are present. Next we consider the practical problem of how to use optical transmission to find surface orientation; this is an inverse problem that arises in the actual use of the LC biosensor. Modulation of transmitted optical intensity under fixed cross-polar upon gradual rotation of an NLC film is one of methods used to determine the degree of alignment and texture type in NLC films [45, 46]. In this work using the accurate FDTD method we have simulated the effect of sample rotation under fixed cross-polars on the optical output in order to determine the preferred orientation in the SPR region; the simulation results are validated with experiments [4, 7]. The simulation steps are as follows:

1. specify the anchoring condition on the SPR by selecting ϕ;
2. solve the texture equation and obtain the director field;
3. with fixed cross-polars, rotate the computational domain by a small angle ψ;

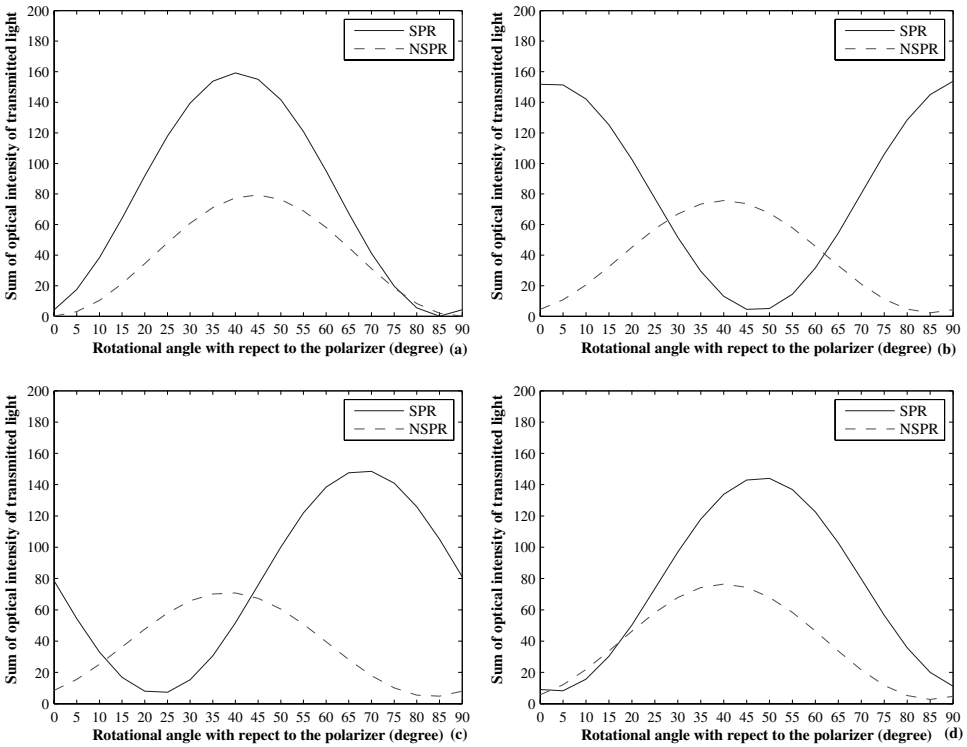


FIG. 4.7. *Modulation of the transmitted intensities of polarized light during rotation with respect to polarizer for the NSPR and SPR regions of the NLC structures corresponding to anchoring boundary conditions for the SPR:* (a) $\phi = \pi/36$, (b) $\phi = \pi/4$, (c) $\phi = \pi/2.57$, *and* (d) $\phi = \pi/2$. *Dashed lines indicate the sum of normalized intensities of transmitted light in the NSPR from* $-1.5 \le x \le -1$ *and* $1 \le x \le 1.5$ *at each rotation angle from* 0 *to* 90°. *Solid lines indicate the sum of normalized intensities of transmitted light in the SPR from* $-1 < x < 1$, *at each rotation angle from* 0 *to* 90°.

4. compute $\mathbf{F}(i), \mathbf{B}(i)$ and $\|E(\phi, \psi)\|$;
5. increase rotation angle and repeat step 4 until total rotation is $\pi/2$;
6. repeat steps 1–5 for $0 \le \phi \le \pi/2$.

The key point in this technique is that the cross-polars are fixed and the sample is rotated.

Figure 4.7 shows optical modulation of the NLS structures for (a) $\phi = \pi/36$, (b) $\phi = \pi/4$, (c) $\phi = \pi/2.57$, and (d) $\phi = \pi/2$ under sample rotation ($0 \le \psi \le \pi/2$) between fixed cross-polars using the FDTD. Comparing the dashed profiles in Figures 4.7(a–d), it is seen that the maximum optical intensity of polarized light appears near $\psi = \pi/4$ and that the profiles are nearly independent of ϕ. The magnitude of ϕ on the SPR region has no effect on $\mathbf{F}$ in the NSPR. Comparing the full-line profiles in Figures 4.7(a–d), it is seen that the extremum in optical intensity is a strong function of ϕ. The profiles shown in Figure 4.7, their specific features, and their response to changes in the azimuthal angle are in good agreement with experimental results [4, 7, 45, 46].

Figure 4.8 shows the extrema in the optical output in the SPR region in terms of ψ as a function of ϕ. Each dot is found from the maximum in optical transmission; for example, in Figure 4.7(a), the maximum corresponds to ψ =0.74, $\phi = 0$. The maximum optical intensities are predicted near $\psi = \pi/4$, for both $\phi = 0$ and $\phi = \pi/2$. Comparing the predicted results of the modulation pattern of the transmitted optical

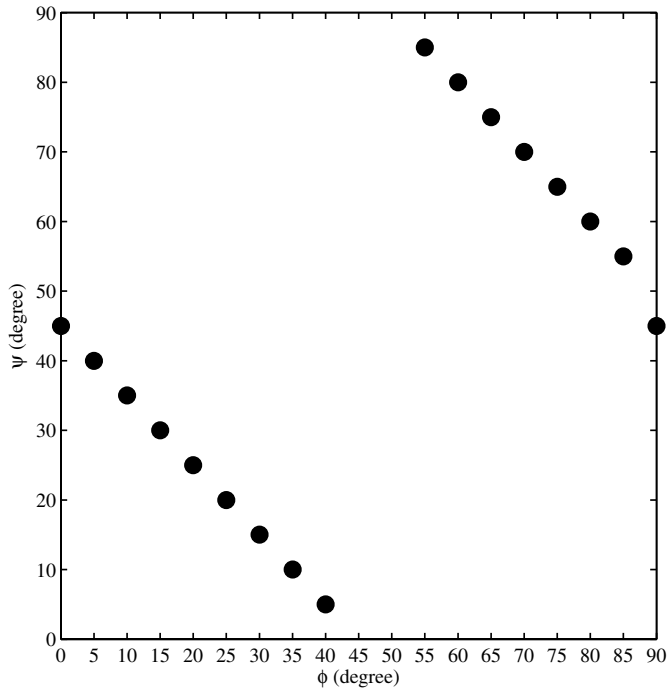


FIG. 4.8. *Extrema in the optical output in the SPR region in terms of ψ as a function of ϕ. ψ and ϕ represent the rotational angle (0 to 90°) and anchoring conditions, respectively, in the SPR region $-1 \leq x \leq 1$.*

intensity with the extrema shown in Figures 4.7 and 4.8 by the FDTD method with the experimental results [4, 7], a preferred orientation of NLCs on the SPR lower surface is confirmed to be near 50 to 55° with the respect to the polarizer.

5. Conclusions. A computational model based on the classical Landau–de Gennes tensor theory for LCs has been developed and implemented in order to describe orientation and surface defects in an NLC film exhibiting two distinct regions due to surface condition changes arising from protein deposition. Optical models using the FDTD and Berreman methods are evaluated in order to predict optical behavior of the heterogeneous films and to determine possible surface anchoring conditions.

Texture formation for complex surface conditions was predicted and characterized as a function of azimuthal anchoring conditions on a lower surface. These predicted textures were then used for optical modeling. Two significant optical features, oscillations and nonsymmetric optical signals, are predicted by the FDTD method but are absent in the Matrix–Berreman method due to its inability to capture effects arising from gradients of the optic axis in the lateral direction. The oscillation's amplitude and magnitude of the nonsymmetric optical response increase with increasing magnitude in the lateral orientation gradients. The optical simulations indicate that the magnitude of the lateral orientation gradient and its symmetry are important factors in the optical behavior of textured NLC films, and hence the FDTD method is more appropriate than the Berreman method for textured samples.

The optical responses of textured LC films in contact with partially covered substrates containing adsorbed proteins were simulated in order to determine the preferred orientation of the optic axis on the protein-covered section of the substrate. Prediction of modulation of transmitted polarized light under fixed cross-polars during sample rotation was used to define a possible preferred orientation on the lower surface where the printed proteins are present. The optical simulations were in good agreement with experimental results [4, 7], indicating that the preferred orientation of the optic axis on the protein patch is approximately 50 to 55° with the respect to the polarizer.

The integrated microstructure-optical simulation model based on the Landau–de Gennes-FDTD method provides for firm foundations on which to further develop biosensors based on LC vision.

REFERENCES

[1] S. Singh, *Liquid Crystals: Fundamentals*, World Scientific, London, 2002.

[2] M. Kleman and O. D. Lavrentovich, *Soft Matter Physics: An Introduction*, Springer-Verlag, London, 2002.

[3] J. J. Skaife and N. L. Abbott, *Quantitative interpretation of the optical texture of liquid crystals caused by specific binding of immunoglobulins to surface-bound antigens*, Langmuir, 16 (2000), pp. 3529–3536.

[4] M. L. Tingey, S. Wilyana, E. J. Snodgrass, and N. L. Abbott, *Imaging of affinity microcontact printed proteins by using liquid crystals*, Langmuir, 20 (2004), pp. 6818–6826.

[5] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, Oxford University Press, Oxford, UK, 1993.

[6] J. J. Skaife and N. L. Abbott, *Influence of molecular-level interactions on the orientations of liquid crystals supported on nano-structured surfaces presenting specifically bound proteins*, Langmuir, 17 (2001), pp. 5595–5604.

[7] M. L. Tingey, E. J. Snodgrass, and N. L. Abbott, *Patterned orientations of liquid crystals on affinity microcontact printed proteins*, Adv. Mater., 16 (2004), pp. 1331–1336.

[8] S. Chandrasekhar, *Liquid Crystals*, Cambridge University Press, Cambridge, UK, 1997.

[9] A. A. Sonin, *The Surface Physics of Liquid Crystals*, Gordon and Breach, Amsterdam, 1995.

[10] D. W. Berreman, *Optics in stratified and anisotropic media:* 4×4*-matrix formulation*, J. Opt. Soc. Amer., 62 (1972), pp. 502-510.

[11] P. Yeh, *Extended Jones matrix method*, J. Opt. Soc. Amer., 72 (1982), pp. 507–513.

[12] D. K. Yang and X. D. Mi, *Modelling of the reflection of cholesteric liquid crystals using the Jones matrix*, J. Phys. D, 33 (2000), pp. 672–676.

[13] C. Gu and P. Yeh, *Extended Jones matrix method and its application in the analysis of compensators for liquid crystal displays*, Displays, 20 (1999), pp. 237–257.

[14] K. H. Yang, *Elimination of the Fabry-Perot effect in the* 4×4 *matrix method for inhomogeneous uniaxial media*, J. Appl. Phys., 68 (1990), pp. 1550–1554.

[15] J. R. Park, G. Ryu, J. Byun, H. Hwang, S. T. Kim, and I. Kim, *Numerical modeling and simulation of a cholesteric liquid crystal polarizer*, Opt. Rev., 9 (2002), pp. 207–212.

[16] H. Wohler, G. Hass, M. Fritsch, and D. A. Mlynski, *Faster* 4×4 *matrix method for uniaxial inhomogeneous media*, J. Opt. Soc. Amer. A, 5 (1988), pp. 1554–1557.

[17] B. Witzigmann, P. Regli, and W. Fichtner, *Rigorous electromagnetic simulation of liquid crystal displays*, J. Opt. Soc. Amer. A., 15 (1998), pp. 753–757.

[18] C. M. Titus, P. J. Bos, J. R. Kelly, and E. C. Gartland, *Comparison of analytical calculations to finite-difference time-domain simulations of one-dimensional spatially varying anisotropic liquid crystal structures*, Japan. J. Appl. Phys., 38 (1999), pp. 1488–1494.

[19] T. Scharf and C. Bohley, *Light propagation through alignment-patterned liquid crystal gratings*, Molecular Crystals and Liquid Crystals, 375 (2002), pp. 491–500.

[20] E. E. Kriezis and S. J. Elston, *Light wave propagation in liquid crystal displays by the* 2*-D finite-difference time-domain method*, Opt. Commun., 177 (2000), pp. 69–77.

[21] E. E. Kriezis, S. K. Filippov, and S. J. Elston, *Light propagation in domain walls in ferroelectric liquid crystal devices by the finite-difference time-domain method*, J. Optim. A. Pure Appl. Optim., 2 (2000), pp. 27–33.

[22] E. E. Kriezis and S. J. Elston, *Finite-difference time domain method for light wave propagation within liquid crystal devices*, Opt. Commun., 165 (1999), pp. 99–105.
[23] D. K. Hwang and A. D. Rey, *Light propagation in textured liquid crystals*, Liquid Cryst., 32 (2005), pp. 483–497.
[24] D. K. Hwang and A. D. Rey, *Computational modeling of light propagation of twist disclinations in liquid crystals based on the finite-difference time-domain (FDTD) method*, Appl. Optics, 44 (2005), pp. 4513–4522.
[25] D. K. Hwang and A. D. Rey, *Computational studies of optical textures of twist disclinations loops in liquid crystal films using the finite-difference time-domain method*, J. Opt. Soc. Amer. A., 23 (2006), pp. 483–496.
[26] G. Gupta, D. K. Hwang, and A. D. Rey, *Optical and structural modeling of disclination lattices in carbonaceous mesophases*, J. Chem. Phys., 122 (2005), paper 034902.
[27] J. Yan and A. D. Rey, *Modeling elastic and viscous effect on the texture of ribbon-shaped carbonaceous mesophase fibers*, Carbon, 41 (2003), pp. 105–121.
[28] A. N. Beris and B. J. Edwards, *Thermodynamics of Flowing Systems*, Clarendon Press, Oxford, UK, 1994.
[29] D. W. Berreman and S. Meiboom, *Tensor represnentation of Oseen-Frank strain energy in unaxial cholesterics*, Phy. Rev. A., 30 (1984), pp. 1955–1959.
[30] M. Doi and S. F. Edwards, *Theory of Polymer Dynamics*, Oxford University Press, New York, 1987.
[31] D. Andrienko, M. Tasinkevych, and S. Dietrich, *Effective pair interactions between colloidal particles at a nematic-isotropic interface*, Europhys. Lett., 70 (2005), pp. 95–101.
[32] T. Z. Qian and P. Sheng, *Orientational state and phase transitions induced by microtextured substrates*, Phys. Rev. E., 55 (1997), pp. 7111–7120.
[33] G. Skacej, A. L. Alexe-Ionescu, G. Barbero, and S. Zumer, *Surface-induced nematic order variation: Intrinsic anchoring and subsurface director deformations*, Phys. Rev. E., 57 (1998), pp. 1780–1788.
[34] R. D Polak, G. P. Crawford, B. C. Kostival, J. W. Doane, and S. Zumer, *Optical determination of the saddle-splay elastic K_{24} in nematic liquid crystals*, Phys. Rev. E., 49 (1994), R978–R981.
[35] G. P. Crawford, R. Ondris-Crawford, S. Zumer, and J. W. Doane, *Anchoring and orientational wetting transitions of confined liquid crystals*, Phys. Rev. Lett., 70 (1993), pp. 1838–1841.
[36] A. Taflove, *Review of the formulation and applications of the FDTD for numerical modeling of electromagnetic wave interactions with arbitrary structures*, Wave Motion, 10 (1998), pp. 547–582.
[37] Matlab 7, The Mathworks, Natick, MA, 2005.
[38] D. M. Sullivan, *An unsplit step 3-D PML for use with the FDTD method*, IEEE Microwave and Guided Wave Letters, 7 (1997), pp. 184–186.
[39] J. P. Berenger, *A perfectly matched layer for the absorption of electromagnetic-waves*, J. Comput. Phys., 114 (1994), pp. 185–200.
[40] A. P. Zhao, J. Juntunen, and A. V. Raisanen, *Material independent PML absorbers for arbitrary anisotropic dielectric media*, Electron. Lett., 33 (1997), pp. 1535–1536.
[41] A. P. Zhao and M. A. Rinne, *Theoretical proof of the material independent PML absorbers used for arbitrary anisotropic media*, Electron. Lett., 34 (1998), pp. 48–49.
[42] S. G. Garcia, I. V. Perez, R. G. Martin, and B. G. Olmedo, *Application of the PML absorbing boundary condition to dielectric anisotropic media*, Electron. Lett., 32 (1996), pp. 1270–1271.
[43] K. S. Yee, *Numerical solution of initial boundary value problems involving Maxwells equations in isotropic media*, IEEE Trans. Antennas and Propagation, 14 (1966), pp. 302–307.
[44] S. T. Wu, C. S. Wu, M. Warenghem, and M. Lsmaili, *Refractive index dispersions of liquid crystals*, Opt. Eng., 32 (1993), pp. 1775–1780.
[45] S. R. Kim, R. R. Shah, and N. L. Abbott, *Orientations of liquid crystals on mechanically rubbed films of bovine serum albumin: A possible substrate for biomolecular assays based on liquid crystals*, Anal. Chem., 72 (2000), pp. 4646–4653.
[46] V. K. Gupta and N. L. Abbott, *Uniform anchoring of nematic liquid crystals on self-assembled monolayers formed from alkanethiolds on obliquely deposited films of gold*, Langmuir, 12 (1996), pp. 2587–2593.

SIAM J. APPL. MATH.
Vol. 67, No. 1, pp. 235–259

ON BUDAEV AND BOGY'S APPROACH TO DIFFRACTION BY THE 2D TRACTION-FREE ELASTIC WEDGE*

V. V. KAMOTSKI†, L. JU. FRADKIN‡, B. A. SAMOKISH§, V. A. BOROVIKOV¶, AND V. M. BABICH‖

Abstract. Several semianalytical approaches are now available for describing diffraction of a plane wave by the 2D (two-dimensional) traction free isotropic elastic wedge. In this paper we follow Budaev and Bogy, who reformulated the original diffraction problem as a singular integral one. This comprises two algebraic and two singular integral equations. Each integral equation involves two unknowns, a function and a constant. We discuss the underlying integral operators and develop a new semianalytical scheme for solving the integral equations. We investigate the properties of the solution obtained and argue that it is the solution of the original diffraction problem. We describe a comprehensive code verification and validation program.

Key words. diffraction, elastic wedge, Sommerfeld transform, singular integral problem

AMS subject classifications. 35J05, 35L05, 44A15, 45Exx, 47B35

DOI. 10.1137/050637297

1. Introduction.

Evaluation of the wave fields diffracted by an isotropic elastic wedge is a challenging problem. A review of various attempts to solve it over the past fifty years is given, e.g., in [5]. It appears that a purely analytical solution is impossible, and instead there have been two major semianalytical approaches developed so far. These are based on

1. a representation of the displacement in the form of a single layer potential—the superposition of the fields radiated by imaginary point sources situated on the faces of the wedge. Their Fourier transforms satisfy integral equations, which can be solved numerically (see, e.g., [12, 11, 10, 13, 14] and references therein), or
2. a representation of the elastodynamic potentials in the form of the Sommerfeld integral—the superposition of plane waves propagating in all (including complex) directions. The amplitudes of the plane waves belong to a certain class of analytical functions and satisfy a system of functional equations. Budaev [4] has reformulated this problem as a singular integral one, involving a

*Received by the editors August 1, 2005; accepted for publication (in revised form) May 22, 2006; published electronically December 5, 2006. This work was conducted at London South Bank University funded by the Industrial Management Committee of the U.K. Nuclear Licensees under the IMC contract PC/GNSR/5129. Partial funding has been provided by EPSRC under the grant GR/R13142 and by London South Bank University. This work is partially based on work that appeared in the reports [2] and [3].

http://www.siam.org/journals/siap/67-1/63729.html

†St. Petersburg Department of Steklov Mathematical Institute, 27 Fontanka, St. Petersburg 191023, Russia. Current address: Bath Institute of Complex Systems and Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK (v.kamotski@maths.bath.ac.uk).

‡Waves and Fields Research Group, ECCE, FESBE, London South Bank University, 103 Borough Rd., London SE1 0AA, UK (fradkil@lsbu.ac.uk).

§Department of Mathematics and Mechanics, St. Petersburg University, 28, Universitetski Prospekt, Petergof, St. Petersburg 198504, Russia (esoch@esoch.mail.iephb.ru).

¶Institute of Problems in Mechanics, Russian Academy of Sciences, 101/1 Prospect Vernadskogo, Moscow 117526, Russia (root@borovik.msk.ru).

‖St. Petersburg Department of Steklov Mathematical Institute, 27 Fontanka, St. Petersburg 191023, Russia (babich@pdmi.ras.ru).

combination of algebraic and singular integral equations. Budaev and Bogy [5, 6, 9] have offered a numerical schedule for solving the problem, and for the incident Rayleigh wave they calculated the Rayleigh reflection and transmission coefficients. However, the schedule has never been given a transparent description.

In sections 2 and 3 we outline our own semianalytical recipe for solution of the singular integral problem, and in section 4 we verify and validate the resulting code. In Appendix A we describe the nomenclature, and in other appendices, we offer the necessary theoretical considerations, formulas, and numerical options.

2. Statement of the problem and the Sommerfeld amplitudes. Let us briefly present the full statement of the original diffraction problem. We seek the elastodynamic potentials $\psi_i = \psi_i(kr, \theta)$ that satisfy the Helmholtz equations in the two-dimensional (2D) wedge of angle 2α with traction-free faces; that is, we address the boundary value problem

$$\Delta\psi_0 + \gamma^2 k^2 \psi_0 = 0, \qquad \Delta\psi_1 + k^2\psi_1 = 0, \qquad |\theta| < \alpha, \tag{2.1}$$

$$\left[\frac{2}{r}\frac{\partial^2\psi_0}{\partial\theta\partial r} + \frac{1}{r^2}\frac{\partial^2\psi_1}{\partial\theta^2} - \frac{\partial^2\psi_1}{\partial r^2} + \frac{1}{r}\frac{\partial\psi_1}{\partial r} - \frac{2}{r^2}\frac{\partial\psi_0}{\partial\theta}\right] = 0, \qquad |\theta| = \alpha, \tag{2.2}$$

$$\frac{1}{\gamma^2}\left[\frac{1}{r^2}\frac{\partial^2\psi_0}{\partial\theta^2} + \frac{\partial^2\psi_0}{\partial r^2} + \frac{1}{r}\frac{\partial\psi_0}{\partial r}\right] - 2\left[\frac{\partial^2\psi_0}{\partial r^2} + \frac{1}{r}\frac{\partial^2\psi_1}{\partial\theta\partial r} - \frac{1}{r^2}\frac{\partial\psi_1}{\partial\theta}\right] = 0, \qquad |\theta| = \alpha. \tag{2.3}$$

Above and everywhere below, the parameter k is the shear wave number; $\gamma = c_S/c_P$ is the ratio of the shear and compressional speeds c_S and c_P, and the subscript i takes values 0 or 1. The geometry of the problem is shown in Figure 2.1. Given an incident wave, we seek the scattered potentials satisfying the radiation conditions at infinity (analogous to the ones in [17, Theorem 4.1]]; also see [15, Appendix C]) and bounded elastic energy condition at the wedge tip.

Note that the potentials are related to displacement $\mathbf{u} = \mathbf{u}(\mathbf{x})$ via $\mathbf{u} = \nabla\psi_0 + \nabla^\perp\psi_1$, where the nabla operators are $\nabla = (\partial_x, \partial_y)$, $\nabla^\perp = (\partial_y, -\partial_x)$. Note too that

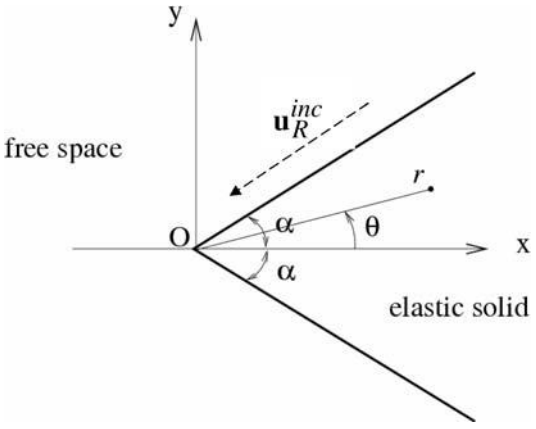


FIG. 2.1. *Geometry of the traction-free elastic wedge.*

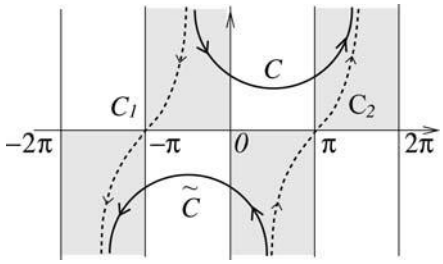


FIG. 2.2. *The Sommerfeld contours in the complex ω-plane.*

in view of this representation the Helmholtz equations imply

$$\psi_0(\mathbf{x}) = -(\gamma k)^{-2}\nabla\cdot\mathbf{u}(\mathbf{x}), \qquad \psi_1(\mathbf{x}) = -k^{-2}\nabla^{\perp}\cdot\mathbf{u}(\mathbf{x}). \tag{2.4}$$

It follows that $\psi_i(kr,\theta)$ are uniquely defined by $\mathbf{u}(\mathbf{x})$, and since for the corresponding Lamé problem in $\mathbf{u}(\mathbf{x})$ the existence and uniqueness results have been proven [16]. [17, Theorem 3.1], the above diffraction problem as formulated in terms of the potentials has a unique solution too. Moreover, due to (2.4) all the analytical properties of $\mathbf{u}(\mathbf{x})$ established in [17] imply the corresponding properties of $\psi_i(kr,\theta)$.

The solutions $\psi_i(kr,\theta)$ of the Helmholtz equations can be represented in the form of the Sommerfeld integrals

$$\psi_i(kr,\theta) = \int_{C\bigcup\widetilde{C}} \Psi_i(\omega+\theta)\mathrm{e}^{\mathrm{i}\gamma kr\cos\,\omega}\mathrm{d}\omega, \tag{2.5}$$

which can be rewritten as

$$\psi_i(kr,\theta) = \int_{C} \Big[\Psi_i(\omega+\theta) - \Psi_i(-\omega+\theta)\Big]\mathrm{e}^{\mathrm{i}\gamma kr\cos\,\omega}\mathrm{d}\omega. \tag{2.6}$$

Here the Π-shaped contour C runs from $-\pi/2+\mathrm{i}\infty$ to $3\pi/2+\mathrm{i}\infty$. The contour $\widetilde{C}$ is the reflection of C with respect to the origin, and the full integration contour $C\bigcup\widetilde{C}$ is shown in Figure 2.2. One can justify (2.6) by applying the Laplace transform in kr to $\psi_i(kr,\theta)$, changing the Laplace variable s to ω, such that $s=\mathrm{i}\cos\,\omega$, and exploiting the behavior of $\psi_i(kr,\theta)$ both in the vicinity of the wedge tip and at infinity (for details see [15, Appendix A]). It follows that the analytic functions $\Psi_i(\omega)$ possess a finite number of singularities in any vertical strip and are regular at the imaginary infinity. Moreover, in Appendix B we show that the tip behavior of $\psi_i(kr,\theta)$ determines the asymptotic expansion of $\Psi_i(\omega)$ at the imaginary infinity and provides us with an asymptotic estimate

$$\Psi_i(\omega) = O(\mathrm{e}^{-\mathrm{Re}\;p\;|\,\mathrm{Im}\;\omega|}), \quad \mathrm{Re}\;p>-1, \quad |\,\mathrm{Im}\;\omega|\to\infty, \tag{2.7}$$

which assures the convergence of integrals (2.6). Note that these integrals are invariant under the transformation $\Psi_i(\omega)\to\Psi_i(\omega)+\mathrm{const}$. In Appendix B we also show that as $\mathrm{Im}\;\omega\to\infty$, the asymptotic expansion of the preexponential factor in (2.6) contains a constant term. Below, we choose $\Psi_i(\omega)$, $i=0,1$ so that the constant terms in their asymptotic expansions valid as $\mathrm{Im}\;\omega\to\infty$ and $\mathrm{Im}\;\omega\to-\infty$ differ only by sign. We call those $\Psi_i(\omega)$ the Sommerfeld amplitudes.

Introducing the closed contour $C\bigcup(-\mathrm{C}_2)\bigcup\widetilde{C}\bigcup(-C_1)$, the Sommerfeld integrals (2.5) can be evaluated as the contributions of the poles and branch points inside that

contour plus the sum of integrals over the steepest descent contours C_1 and C_2. The former describe all bulk and surface waves arising in the problem, as well as the head waves, and the latter can be evaluated using the steepest descent method to provide a description of the tip diffracted body waves. It follows that all physically meaningful poles and branch points of $\Psi_i(\omega+\theta)$ must be located between the contours C_1 and C_2, and in this region the Sommerfeld amplitudes $\Psi_i(\omega)$ should contain no physically meaningless singularities. Since inside the wedge we have $|\theta| \le \alpha$, this means that all physically meaningful singularities—and only those—must lie at a finite distance from the horizontal axis between the contours C_1 and C_2 as shifted horizontally by $-\alpha$ and α, respectively; that is, within the Malyuzhinets region

$$\left\{\omega: \ -\frac{\pi}{2}-\alpha-2\tan^{-1}(\mathrm{e}^{-\operatorname{Im}\ \omega}) \le \operatorname{Re}\ \omega \le \frac{3\pi}{2}+\alpha-2\tan^{-1}(\mathrm{e}^{-\operatorname{Im}\ \omega})\right\}. \tag{2.8}$$

To summarize, the scattered field can be fully described once an efficient algorithm is produced for calculating the Sommerfeld amplitudes $\Psi_i(\omega)$ in the region (2.8). We proceed with this task.

3. Problem reformulation. In view of its symmetry with respect to the polar angle, the original problem naturally splits into "symmetric" and "antisymmetric," corresponding respectively to the symmetric and antisymmetric parts of the incident wave. These involve functions $\Psi_i^{\pm}(\omega)$ such that we have

$$\Psi_i^{\pm}(\omega) = \frac{1}{2}[\Psi_i(\omega) + (-1)^{i+1}\Psi_i(-\omega)]. \tag{3.1}$$

We follow the approach pioneered in elastodynamics by [4] and first substitute (2.6) into the boundary conditions (2.2) and (2.3) to obtain the system of functional equations. We then employ the singular integral transforms to reformulate the problem as a system of algebraic and singular integral equations.

3.1. Functional equations. Let us start with the symmetric problem. The boundary conditions (2.2) and (2.3) imply

$$\begin{aligned}&\int_C \gamma^2 a_1(\omega)\Big[\Psi_0^+(\omega+\theta)-\Psi_0^+(\omega-\theta)\Big]\mathrm{e}^{\mathrm{i}\gamma kr\ \cos\ \omega}\mathrm{d}\omega\\ &\qquad -\int_C a_2(\omega)\Big[\Psi_1^+(\omega+\theta)-\Psi_1^+(\omega-\theta)\Big]\mathrm{e}^{\mathrm{i}kr\cos\ \omega}\mathrm{d}\omega = 0,\end{aligned} \tag{3.2}$$

$$\begin{aligned}&\int_C \gamma^2\left[a_3(\omega)+\frac{1}{\gamma^2}\right]\Big[\Psi_0^+(\omega+\theta)+\Psi_0^+(\omega-\theta)\Big]\mathrm{e}^{\mathrm{i}\gamma kr\cos\ \omega}\mathrm{d}\omega\\ &\qquad -\int_C a_1(\omega)\Big[\Psi_1^+(\omega+\theta)+\Psi_1^+(\omega-\theta)\Big]\mathrm{e}^{\mathrm{i}kr\cos\ \omega}\mathrm{d}\omega = 0,\end{aligned} \tag{3.3}$$

where $a_1(\omega) = \sin 2\omega, a_2(\omega) = -\cos 2\omega$, and $a_3(\omega) = -2\cos^2\omega$. Introducing in the first terms of both (3.2) and (3.3) the new integration variable $\breve{\omega}$ such that $\cos\breve{\omega} = \gamma\cos\omega$, dropping the check, and transforming the contour of integration back to C, the equations acquire the form

$$\int_C f(\omega)\mathrm{e}^{\mathrm{i}kr\cos\ \omega}\mathrm{d}\omega = 0, \tag{3.4}$$

where, due to (2.7), as $|\operatorname{Im}\ \omega| \to \infty$, $f(\omega) = O(\exp(\operatorname{Im}\ \omega\ [2-\operatorname{Re}\ p]))$, with $\operatorname{Re}\ p > -1$. Using the Malyuzhinets theorem [20], f is an odd trigonometric polynomial of

the second order. It follows that the pair $\Psi_i^+(\omega)$ satisfies (3.2) and (3.3) if and only if it satisfies the functional equations

$$t_{11}\Big\{\Psi_0^+[g(\omega)+\alpha]+\Psi_0^+[g(\omega)-\alpha]\Big\}+t_{12}\Big\{\Psi_1^+(\omega+\alpha)+\Psi_1^+(\omega-\alpha)\Big\}=Q_1^+,$$

$$t_{21}\Big\{\Psi_0^+[g(\omega)+\alpha]-\Psi_0^+[g(\omega)-\alpha]\Big\}+t_{22}\Big\{\Psi_1^+(\omega+\alpha)-\Psi_1^+(\omega-\alpha)\Big\}=Q_2^+, \tag{3.5}$$

where $t_{11}=\cos 2\omega \sin\omega/\sqrt{\gamma^2-\cos^2\omega}$, $t_{12}=t_{21}=\sin 2\omega$, $t_{22}=-\cos 2\omega$, and we have

$$Q_j^+=c_{j1}^+\sin\omega+c_{j2}^+\sin 2\omega, \tag{3.6}$$

with c_{jk}^+, $j,k=1,2$, unknown constants. The function $g(\omega)=\cos^{-1}(\gamma^{-1}\cos\omega)$ relates the shear incidence angles to compressional reflection angles, and its branch cuts are chosen so that the deformed contour of integration $\breve{C}=\{\breve{\omega}=g(\omega):\omega\in C\}$ may be transformed back to C without touching them. They are the segments

$$[-\theta_h+\pi n,\theta_h+\pi n],\quad \theta_h=\cos^{-1}\gamma,\quad n\ -\ \text{integer}. \tag{3.7}$$

The branch of $g(\omega)$ is chosen so that it has the properties

$$\begin{aligned} &g\left(\frac{\pi}{2}\right)=\frac{\pi}{2},\\ &g(\omega+\pi n)=g(\omega)+\pi n,\\ &g(\omega)\simeq\omega-\mathrm{i}\ln\gamma+O\big(\mathrm{e}^{-2|\mathrm{Im}\omega|}\big)\ \text{ as } \mathrm{Im}\,\omega\to\infty. \end{aligned} \tag{3.8}$$

In order to investigate restrictions on c_{ij}^+ let us substitute expansions (B.7) of the Sommerfeld amplitudes into (3.5) and equate the coefficients of the leading asymptotic terms in the resulting equations. First, we note that in the symmetric case, (B.7) contains no constant terms, and therefore there are no $\exp(-2\mathrm{i}\omega)$ terms in the left-hand sides of these equations. This implies that $c_{12}^+=c_{22}^+=0$. Secondly, since the tip asymptotics of $\psi_i^+(kr,\theta)$ contain the terms with the exponent 1, these sides contain the $\exp(-\mathrm{i}\omega)$ terms. Equating the coefficients of the $\exp(-\mathrm{i}\omega)$ terms, we obtain

$$(\gamma\Psi_{0m}^+ +\mathrm{i}\Psi_{1m}^+)\cos\alpha=\frac{\mathrm{i}}{2}c_{11}^+,\qquad -(\gamma\Psi_{0m}^+ +\mathrm{i}\Psi_{1m}^+)\sin\alpha=\frac{\mathrm{i}}{2}c_{21}^+. \tag{3.9}$$

Hence we have

$$c_{21}^+=-c_{11}^+\tan\alpha. \tag{3.10}$$

Note that if $\gamma\Psi_{0m}^+ +\mathrm{i}\Psi_{1m}^+=0$, then $c_{21}^+=-c_{11}^+=0$.

It follows that the right-hand sides in (3.5) *might be*—and, as we show in sections 4 and 5, *are*—nonzero, so that we have

$$Q_1^+=c_1^+\sin\omega,\qquad Q_2^+=-c_1^+\tan\alpha\sin\omega, \tag{3.11}$$

where from now on, for simplicity, we use the notation $c_1^+=c_{11}^+$.

The antisymmetric problem can be treated similarly, with one minor modification: For all wedge angles α, expansion (B.7) might contain a nonzero constant term Ψ_{00}^-, and therefore the functional equations might contain a second order term. However, this term can be eliminated by subtracting Ψ_{00}^- from the Sommerfeld amplitude $\Psi_0^-(\omega)$, redefining it in the process. The corresponding functional equations are

$$t_{21}\Big\{\Psi_0^-[g(\omega)+\alpha]+\Psi_0^-[g(\omega)-\alpha]\Big\}+t_{22}\Big\{\Psi_1^-(\omega+\alpha)+\Psi_1^-(\omega-\alpha)\Big\}=Q_1^-,$$

$$t_{11}\Big\{\Psi_0^-[g(\omega)+\alpha]-\Psi_0^-[g(\omega)-\alpha]\Big\}+t_{12}\Big\{\Psi_1^-(\omega+\alpha)-\Psi_1^-(\omega-\alpha)\Big\}=Q_2^-, \tag{3.12}$$

with

$$Q_1^- = c_1^- \sin\omega, \qquad Q_1^- = c_1^- \tan\alpha \sin\omega. \tag{3.13}$$

We note that the above reasoning involves only the asymptotic terms with $p_m^- = 0$ or else with $p_m^\pm = 1$ and $N_m^\pm = 1$, and therefore applies to all wedge angles under consideration. We note too that Budaev and Bogy [5] have made several attempts to establish restrictions on the constants. They first used the arguments of the type outlined above in [7]. By excluding from consideration the terms with $p_m^\pm = 1$, it is easy to reach the erroneous conclusion that all constants $c_{jk}^\pm$ are zero. In the static problems, such exclusion is justified, because the terms describe body translations. By contrast, in the dynamic problems, their presence is indicative of nontrivial phenomena.

We proceed by discussing the singularities of the Sommerfeld amplitudes. First, we assume that the incident wave is plane or Rayleigh, so that it manifests itself in $\Psi_i^\pm(\omega)$ in the form of terms which contain simple poles $\theta_{i\ell}$ in the strip $|\operatorname{Re}\ \omega| \le \alpha$. The functional equations (3.5) and (3.12) can be recast as

$$\begin{pmatrix} \Psi_0^\pm(g(\omega)+\alpha) \\ \Psi_1^\pm(\omega+\alpha) \end{pmatrix} = \pm \begin{pmatrix} \mathrm{r}_{11}(\omega) & \mathrm{r}_{12}(\omega) \\ \mathrm{r}_{21}(\omega) & \mathrm{r}_{22}(\omega) \end{pmatrix} \begin{pmatrix} \Psi_0^\pm(g(\omega)-\alpha) \\ \Psi_1^\pm(\omega-\alpha) \end{pmatrix} + c_1^\pm \frac{\sqrt{\gamma^2 - \cos^2\omega}}{\Delta(\omega)} \begin{pmatrix} e_1^\pm(\omega) \\ e_2^\pm(\omega) \end{pmatrix}, \tag{3.14}$$

where the reflection coefficients for the traction-free elastic half space $\mathrm{r}_{jk}(\omega)$, $j,k = 1,2$, as well as the Rayleigh function $\Delta(\omega)$, and functions $e_j^\pm(\omega)$ are given in Appendix A. The system (3.14) can be used to effect the analytical continuation from the strip $|\operatorname{Re}\ \omega| \le \alpha$ and thus find all poles $\theta_{i\ell}$ of the Sommerfeld amplitudes, with their respective residues, which are located in the strip $\operatorname{Re}\ \omega \,\epsilon\, I = [\pi/2-\alpha,\ \pi/2+\alpha]$. The rationale behind the choice of the latter strip is clarified below. The poles are incidence and reflection angles of the respective incident, reflected, and multiply reflected waves, and their residues describe the amplitudes of these waves—see [5, (17) and (18)]. The first index in $\theta_{i\ell}$ refers to the mode of the wave, and the second to its place in a sequence of all incident and (multiply) reflected waves (see [15, Appendix D]).

Let us now again follow the above authors and introduce the decomposition

$$\Psi_i^\pm(\omega) = \widehat{\Psi}_i^\pm(\omega) + \widetilde{\Psi}_i^\pm(\omega), \tag{3.15}$$

where the unknown $\widetilde{\Psi}_i^\pm(\omega)$ is regular in the strip $\operatorname{Re}\ \omega \,\epsilon\, I$, and the known $\widehat{\Psi}_i^\pm$ is

$$\widehat{\Psi}_i^\pm(\omega) = \sum_\ell \operatorname{Res}\ (\Psi_i^\pm; \theta_{i\ell}) \sigma(\omega - \theta_{i\ell}), \quad \operatorname{Re}\ \theta_{i\ell} \,\epsilon\, I. \tag{3.16}$$

Above, an otherwise arbitrary function $\sigma(\omega)$ should be chosen to be analytic everywhere inside the strip $\operatorname{Re}\ \omega \,\epsilon\, I$, except for a simple pole at zero, where it has the residue 1. The weakest restriction we can impose on behavior of $\sigma(\omega)$ at the imaginary infinity is that it grows more slowly than $\exp(|\operatorname{Im}\ \omega|\pi/2\alpha)$. Instead, we impose a stronger restriction—that it behaves as the amplitudes in (2.7). If $\sigma(\omega)$ possesses singularities which lie outside the strip $\operatorname{Re}\ \omega \,\epsilon\, I$, the functions $\widehat{\Psi}_i^\pm(\omega)$ contain extra poles which describe waves that are outgoing at physical reflection angles but have nonphysical amplitudes. This causes no complication, since the corresponding singular terms in $\widehat{\Psi}_i^\pm(\omega)$ and $\widetilde{\Psi}_i^\pm(\omega)$ mutually cancel.

Next we substitute decomposition (3.15) into the functional equations (3.5) and then (3.12) to obtain the following inhomogeneous systems of equations for the regular components of the Sommerfeld amplitudes:

$$\Big\{\widetilde{\Psi}_0^+[g(\omega)+\alpha]+\widetilde{\Psi}_0^+[g(\omega)-\alpha]\Big\}+B\Big[\widetilde{\Psi}_1^+(\omega+\alpha)+\widetilde{\Psi}_1^+(\omega-\alpha)\Big]=R_1^+ +c_1^+S_1,$$

$$A\Big\{\widetilde{\Psi}_0^+[g(\omega)+\alpha]-\widetilde{\Psi}_0^+[g(\omega)-\alpha]\Big\}+\Big[\widetilde{\Psi}_1^+(\omega+\alpha)-\widetilde{\Psi}_1^+(\omega-\alpha)\Big]=R_2^+ +c_1^+\tan\alpha\, S_2,$$

(3.17)

and

$$A\Big\{\widetilde{\Psi}_0^-[g(\omega)+\alpha]+\widetilde{\Psi}_0^-[g(\omega)-\alpha]\Big\}+\Big[\widetilde{\Psi}_1^-(\omega+\alpha)+\widetilde{\Psi}_1^-(\omega-\alpha)\Big]=R_2^- +c_1^-S_2,$$

$$\Big\{\widetilde{\Psi}_0^-[g(\omega)+\alpha]-\widetilde{\Psi}_0^-[g(\omega)-\alpha]\Big\}+B\Big[\widetilde{\Psi}_1^-(\omega+\alpha)-\widetilde{\Psi}_1^-(\omega-\alpha)\Big]=R_1^- -c_1^-\tan\alpha\, S_1,$$

(3.18)

where we use the notation

$$A=\frac{t_{21}(\omega)}{t_{22}(\omega)}=-\tan 2\omega,\qquad B=\frac{t_{12}(\omega)}{t_{11}(\omega)}=\frac{2\cos\,\omega\sqrt{\gamma^2-\cos^2\omega}}{\cos\,2\omega},$$

$$R_1^\pm=-\Big\{\widehat{\Psi}_0^+[g(\omega)\pm\alpha]\pm\widehat{\Psi}_0^\pm[g(\omega)-\alpha]\Big\}-B\Big[\widehat{\Psi}_1^\pm(\omega+\alpha)\pm\widehat{\Psi}_1^\pm(\omega-\alpha)\Big],$$

$$R_2^\pm=-A\Big\{\widehat{\Psi}_0^\pm[g(\omega)+\alpha]\mp\widehat{\Psi}_0^\pm[g(\omega)-\alpha]\Big\}-\Big[\Psi_1^\pm(\omega+\alpha)\mp\widehat{\Psi}_1^\pm(\omega-\alpha)\Big],$$

$$S_1=\frac{\sqrt{\gamma^2-\cos^2\,\omega}}{\cos\,2\omega},\qquad S_2=\frac{\sin\,\omega}{\cos\,2\omega}. \tag{3.19}$$

To summarize, following Budaev and Bogy, the original problem can be reformulated as the following boundary value problem in the theory of analytic functions: Seek constants $c_1^\pm$ and functions $\widetilde{\Psi}_i^\pm(\omega)$ such that

1. $\widetilde{\Psi}_i^\pm(\omega)$ are analytic for Re $\omega\,\epsilon\, I$ and satisfy the asymptotic estimate (2.7);
2. the values that $\widetilde{\Psi}_i^\pm(\omega)$ take on the boundaries of the strip Re $\omega\,\epsilon\, I$ are linked by (3.17) and (3.18) (that is, we solve these equations for Re $\omega=\pi/2$).

The above considerations and the properties of the Sommerfeld transform, which are outlined in [15, Appendix A], show that such a pair exists if there exists a solution of the original problem. The uniqueness of $\Psi_i^\pm(\omega)$ is a more complicated issue, which we address in section 5.

3.2. Singular integral equations. Budaev [4] has suggested exploiting the fact that for all functions $F(\omega)$ satisfying the first of the above assumptions, the singular integral transform

$$(HF)(\omega)=\frac{1}{2\alpha\mathrm{i}}\mathrm{V.P.}\int_{\pi/2-\mathrm{i}\infty}^{\pi/2+\mathrm{i}\infty}\frac{F(\xi)\mathrm{d}\xi}{\sin\,[\frac{\pi}{2\alpha}(\xi-\omega)]},\quad \mathrm{Re}\ \omega=\frac{\pi}{2}, \tag{3.20}$$

has the property

$$H:\ F(\omega+\alpha)+F(\omega-\alpha)\to F(\omega+\alpha)-F(\omega-\alpha),\quad \mathrm{Re}\ \omega=\frac{\pi}{2}, \tag{3.21}$$

with V. P. standing for the Cauchy principal value. This means that on the vertical line Re $\omega=\pi/2$, the terms in the square brackets in (3.17) and (3.18) are related by

H. The terms in the curly brackets are linked by a similar explicit transform,

$$\overline{H}F(\omega) = \frac{1}{2\alpha \mathrm{i}}\, V.P. \int_{\pi/2-\mathrm{i}\infty}^{\pi/2+\mathrm{i}\infty} \frac{F(\xi)g'(\xi)\mathrm{d}\xi}{\sin\left\{\frac{\pi}{2\alpha}[g(\xi)-g(\omega)]\right\}}, \quad \mathrm{Re}\ \omega = \frac{\pi}{2}, \tag{3.22}$$

where $g'(\xi) = dg/d\xi$. This suggests introducing new unknown functions

$$X^{\pm}(\omega) = \widetilde{\Psi}_0^{\pm}[g(\omega)+\alpha] + \widetilde{\Psi}_0^{\pm}[g(\omega)-\alpha], \quad Y^{\pm}(\omega) = \widetilde{\Psi}_1^{\pm}(\omega+\alpha) + \widetilde{\Psi}_1^{\pm}(\omega-\alpha). \tag{3.23}$$

Note that the line Re $\omega = \pi/2$ is of special significance, because the function $g(\omega)$ maps it onto itself. We can now use (3.21) to transform (3.17) and (3.18) into the system comprising algebraic equations and singular integral equations which hold on the vertical line Re $\omega = \frac{\pi}{2}$. This is the crux of Budaev and Bogy's approach.

Changing to the new independent real variable η, such that $\omega = \pi/2 + \mathrm{i}\eta$, the symmetric problem becomes

$$x^+(\eta) + b(\eta)y^+(\eta) = r_1^+(\eta) - c_1^+ \frac{\sqrt{\gamma^2 + \sinh^2\eta}}{\cosh 2\eta}, \tag{3.24}$$

$$a(\eta)\overline{\mathcal{H}}x^+(\eta) + \mathcal{H}y^+(\eta) = r_2^+(\eta) - c_1^+ \tan\alpha \frac{\cosh\eta}{\cosh 2\eta}, \tag{3.25}$$

where we have

$$x^{\pm}(\eta) = X^{\pm}\left(\frac{\pi}{2} + \mathrm{i}\eta\right), \quad y^{\pm}(\eta) = Y^{\pm}\left(\frac{\pi}{2} + \mathrm{i}\eta\right). \tag{3.26}$$

Standardizing notations and substituting (3.24) into (3.25), the problem transforms to a final *singular integral equation* in two unknowns, a function $y^+(\eta)$ and a constant c_1^+,

$$M^+y^+(\eta) = q_0^+(\eta) + c_1^+ q_1^+(\eta), \quad \eta - \text{real}, \tag{3.27}$$

where $M^+ = \mathcal{H} - a\overline{\mathcal{H}}b$. Using the same approach, the antisymmetric problem transforms to

$$a(\eta)x^-(\eta) + y^-(\eta) = r_2^-(\eta) + c_1^- \frac{\cosh\eta}{\cosh 2\eta}, \tag{3.28}$$

$$M^-x^-(\eta) = q_0^-(\eta) + c_1^- q_1^-(\eta), \quad \eta - \text{real}, \tag{3.29}$$

where $M^- = \overline{\mathcal{H}} - b\mathcal{H}a$. The rest of the nomenclature can be found in Appendix A.

4. A new numerical schedule. Budaev and Bogy [5, 6, 9] have advanced various implementations of the numerical schedule for computing $\Psi_i^{\pm}(\omega)$, all of which involve the following three steps:

1. evaluating $y^+(\eta)$ and $x^-(\eta)$ on the line $\eta = 0$ (Re $\omega = \pi/2$) by solving the singular integral equations (3.27) and (3.29), and then evaluating $x^+(\eta)$ and $y^-(\eta)$ by solving the algebraic equations (3.24) and (3.28);
2. evaluating $\widetilde{\Psi}_i^{\pm}(\omega)$ in the strip Re $\omega\, \epsilon\, I$, using the convolution type transforms (4.13) and (4.14) below, with the kernels singular on the boundary of this strip;
3. continuing the computed Sommerfeld amplitudes $\Psi_i^{\pm}(\omega)$ analytically to the right of the strip Re $\omega\, \epsilon\, I$ by using the functional equations (3.14). Recasting these equations to effect the continuation to the left of Re $\omega\, \epsilon\, I$.

We have developed an alternative recipe for carrying out the first two steps.

4.1. Solving singular integral equations in two unknowns on the line $\eta = 0$ (Re $\omega = \pi/2$). Let us consider the symmetric case first. Operator M^+ is not analytically invertible, but [5] suggest that (3.27) can be rewritten as

$$(\mathcal{H}d + K)y^+(\eta) = q_0^+(\eta) + c_1^+ q_1^+(\eta), \tag{4.1}$$

where $\mathcal{H}$ is the singular operator introduced above, analytically invertible in the space of bounded functions; K is a regular operator; and $d(t)$ is an exponentially decreasing function. Importantly, $\mathcal{H}$ has the property

$$\int_{-\infty}^{\infty} \mathcal{H}f(\eta)d\eta = 0, \tag{4.2}$$

and therefore its range consists of all $L^2(\mathbb{R})$ functions, with the zero integral, where $L^2(\mathbb{R})$ is the space of all integrable functions of real variable. Budaev and Bogy [6] state that they regularize (4.1) by applying $\mathcal{H}^{-1}$ to both its sides. They carry out numerical evaluation of the resulting singular integral equation by using (4.2) as a constraint, and calculate c_1 and $y(\eta)$ both at once. By contrast, below we argue that the right-hand side of (4.1) belongs to the domain of $\mathcal{H}^{-1}$ for only one value of c_1, and we carry out the regularization by finding this value and thus arriving at a singular integral equation in one unknown, $y(\eta)$. At present, our schedule works only for $\alpha < \pi/2$.

We start by observing that (4.1) is solvable only if its right-hand side belongs to the range of $\mathcal{H}d + K$. We cannot describe this range explicitly. However, it is clear that $(-Ky^+ + q_0^+ + c_1^+ q_1^+)(\eta)$ should be in the range of $\mathcal{H}$. It follows that we must have

$$\int_{-\infty}^{\infty} \left[(Ky^+)(\eta) - q_0^+(\eta) - c_1^+ q_1^+(\eta)\right]\, d\eta = 0. \tag{4.3}$$

All our numerical experiments confirm that neither $q_0^+(\eta)$ nor $q_1^+(\eta)$ are in the range of $\mathcal{H}d + K$—by producing the nonzero "solution defects" λ_0^+ and λ_1^+ defined by (4.11) below. Therefore, (4.1) is solvable only if the right-hand side of (4.1) is in the range. This gives the following relationship between c_1^+ and $y^+(\eta)$:

$$c_1^+ = \int_{-\infty}^{\infty} \left[(Ky^+)(\eta) - q_0^+(\eta)\right]\, d\eta \left[\int_{-\infty}^{\infty} q_1^+(\eta)\, d\eta\right]^{-1}. \tag{4.4}$$

By substituting (4.4) into (4.1), c_1^+ is eliminated and we obtain

$$(\mathcal{H}d + P_{q_1^+}K)y^+(\eta) = P_{q_1^+}q_0^+, \tag{4.5}$$

where an unbounded projector

$$(P_{q_1^+}u)(\eta) = u(\eta) - q_1^+(\eta)\int_{-\infty}^{\infty} u(t)\, dt \left[\int_{-\infty}^{\infty} q_1^+(t)\, dt\right]^{-1} \tag{4.6}$$

maps any function in $L^2(\mathbb{R})$ with a finite integral into the range of $\mathcal{H}$ and has the property

$$P_q u(\eta) = \begin{cases} u(\eta) & \text{for all } u(\eta) \text{ such that } \int_{-\infty}^{\infty} u(t)\, dt = 0, \\ 0 & \text{for } u(\eta) = q(\eta). \end{cases} \tag{4.7}$$

We refer to the function $q(\eta)$ as the projector kernel.

Note that in (4.5), the integrals of both $P_{q_1^+}Ky^+(\eta)$ and $P_{q_1^+}q_0^+(\eta)$ are zero, and therefore the inverse operator $\mathcal{H}^{-1}$ can now be safely applied to both sides. Introducing on top of that a new unknown function $\widetilde{y}^+(\eta) = d^{1/2}(\eta)y^+(\eta)$, the final regularized integral equation is

$$\widetilde{y}^+(\eta) + \widetilde{L}^+\widetilde{y}^+(\eta) = \widetilde{q}^+(\eta), \tag{4.8}$$

where $\widetilde{L}^+ = d^{-1/2}\mathcal{H}^{-1}P_{q_1^+}Kd^{-1/2}$, an operator with a smooth kernel, and $\widetilde{q}^+(\eta) = d^{-1/2}(\eta)\mathcal{H}^{-1}P_{q_1^+}q_0^+(\eta)$. The equation involves one unknown, $\widetilde{y}^+(\eta)$, and can be solved using a standard quadrature method (see, e.g., [1]). Note that normalizing the original unknown by $d^{1/2}(\eta)$ rather than $d(\eta)$ leads to a new operator with a bounded kernel (cf. [5]). The normalization achieves symmetrization of the kernel, so that whether $\eta \to \infty$ or $t \to \infty$, it exhibits the same singular behavior.

Equations (4.4) and (4.5) imply (4.1). This means that the combination of c_1^+ and a solution of (4.5) gives us the solution of (4.1). However, in our code instead of solving (4.8), we implement a slightly different approach: Since $q_1^+(\eta)$ is rather complex, instead of $P_{q_1^+}$ we employ the projector P_{q_2}, with the kernel

$$q_2(\eta) = \frac{1}{2\alpha}\frac{1}{\cosh\frac{\pi}{2\alpha}\eta}. \tag{4.9}$$

Numerical experiments have shown that this kernel leads to a stable evaluation scheme. We then regularize and solve two equations

$$(\mathcal{H}d + P_{q_2}K)y_i^+(\eta) = P_{q_2}q_i^+(\eta), \quad i = 0, 1 \tag{4.10}$$

(see Appendix C). The "solution defects"

$$\lambda_i^+ = \int_{-\infty}^{\infty}\big[(Ky_i^+)(t) - q_i^+(t)\big]dt, \quad i = 0, 1, \tag{4.11}$$

turn out to be nonzero, indicating that neither $q_0^+(\eta)$ nor $q_1^+(\eta)$ is in the range of $\mathcal{H}d + K$. It follows that the solution $(y^+(\eta), c_1^+)$ of (3.27) can be obtained using

$$c_1^+ = -\frac{\lambda_0}{\lambda_1}, \qquad y^+(\eta) = y_0^+(\eta) + c_1^+y_1^+(\eta). \tag{4.12}$$

The antisymmetric problem can be treated in a similar manner (see Appendix C).

4.2. Evaluating $\widetilde{\Psi}_i^{\pm}(\omega)$ in the strip Re $\omega \in I$. As already mentioned above, according to [5], evaluation of $\widetilde{\Psi}_i^{\pm}(\omega)$ in the strip Re $\omega \in I$ can be carried out by using the singular convolution-type integrals,

$$\Psi_0^{\pm}(\omega) = \frac{1}{4\alpha\mathrm{i}}\mathrm{V.P.}\int_{-\infty}^{\infty}\frac{f_0^{\pm}(\eta)}{\cos\frac{\pi}{2\alpha}(\frac{\pi}{2} - \omega + \mathrm{i}\chi(\eta))}d\eta, \tag{4.13}$$

with $f_0^+(\eta) = y^+(\eta)\tanh 2\eta - \mathrm{i}c_1^+\cosh\eta/\cosh 2\eta + \mathrm{i}r_1^+(\eta)\chi'(\eta)$, $f_0^-(\eta) = \mathrm{i}x^-(\eta)\chi'(\eta)$, $\chi'(\eta) = d\chi/d\eta$, and

$$\widetilde{\Psi}_1^{\pm}(\omega) = \pm\frac{1}{4\alpha\mathrm{i}}\mathrm{V.P.}\int_{-\infty}^{\infty}\frac{f_1^{\pm}(\eta)\,d\eta}{\cos\frac{\pi}{2\alpha}(\frac{\pi}{2} - \omega + \mathrm{i}\eta)}, \tag{4.14}$$

with $f_1^+(\eta) = \mathrm{i}y^+(\eta)$ and $f_1^-(\eta) = x^-(\eta)\tanh 2\eta - \mathrm{i}c_1^-\cosh\eta/\cosh 2\eta - \mathrm{i}r_1^-(\eta)$.

If—as is the case for the Sommerfeld amplitudes of the solution of the original problem—as Im $\omega \to \pm\infty$, the leading terms in (B.7) are $O(\exp(\pm \mathrm{i} p\omega))$, with Re $p > 0$, then for $\alpha < \pi$, $-\operatorname{Re} p - \pi/2\alpha < 0$, and therefore the above integrals converge.

We start with the integral of the type (4.14). Its generic form is

$$\text{V.P.}\int_{-\infty}^{\infty} \frac{f(\eta)}{\cos\frac{\pi}{2\alpha}(\xi + \mathrm{i}\eta)} d\eta, \qquad |\operatorname{Re}\ \xi| \le \alpha, \tag{4.15}$$

where the new complex variable is $\xi = \pi/2 - \omega$. When $|\operatorname{Re}\ \xi| < \alpha$, the integral (4.15) can be approximated using the trapezoidal rule. The approximation error is of order $O\left[\exp\left(-2\pi\sigma/h\right)\right]$, with h—the distance between the nodes of a uniform mesh and $\sigma(\omega)$—the half width of the strip that is centered on the real line and inside which the integrand is regular. In our case, $\sigma \le \min\{\alpha - \operatorname{Re}\ \xi, \alpha + \operatorname{Re}\ \xi\}$, and as $\operatorname{Re}\xi \to \alpha$, the accuracy of the trapezoidal rule deteriorates. Therefore, a more robust quadrature formula is required, with accuracy depending on the function $f(\eta)$ and not on parameter ξ. One such formula may be obtained with a modified sinc function,

$$\omega_h(\eta) = \frac{h}{2\alpha} \frac{\sin\frac{\pi}{h}\eta}{\sinh\frac{\pi}{2\alpha}\eta}. \tag{4.16}$$

Note that we have

$$\omega_h(nh) = \begin{cases} 1, & n = 0, \\ 0, & n \ne 0, \end{cases} \tag{4.17}$$

and

$$f(\eta) \approx \sum_{n=-\infty}^{\infty} f(nh)\omega_h(\eta - nh), \tag{4.18}$$

where the sum on the right-hand side interpolates $f(\eta)$. Therefore, the integral (4.15) may be approximated by

$$\text{V.P.}\int_{-\infty}^{\infty} \frac{f(\eta)}{\cos\frac{\pi}{2\alpha}(\xi + \mathrm{i}\eta)} d\eta \approx \sum_{-\infty}^{\infty} \mathcal{A}_n(\xi) f(nh), \tag{4.19}$$

with the coefficients given by

$$\mathcal{A}_n(\xi) = \text{V.P.}\int_{-\infty}^{\infty} \frac{\omega_h(\eta - nh)}{\cos\frac{\pi}{2\alpha}(\xi + \mathrm{i}\eta)} d\eta. \tag{4.20}$$

These can be evaluated approximately by introducing new variables $\check{\eta} = \eta - nh$ and $\check{\xi} = \xi + \mathrm{i}nh$. Then (4.20) can be rewritten as

$$\begin{aligned} \mathcal{A}_n(\xi) &= \frac{h}{2\alpha}\text{V.P.}\int_{-\infty}^{\infty} \frac{\sin\frac{\pi}{h}\check{\eta}}{\sinh\frac{\pi}{2\alpha}\check{\eta}} \frac{d\check{\eta}}{\cos\frac{\pi}{2\alpha}(\check{\xi} + \mathrm{i}\check{\eta})} \\ &= \frac{h}{4\alpha\mathrm{i}}\text{V.P.}\int_{-\infty}^{\infty} \frac{\mathrm{e}^{\frac{\pi}{h}\check{\eta}\mathrm{i}}}{\sinh\frac{\pi}{2\alpha}\check{\eta}} \frac{d\check{\eta}}{\cos\frac{\pi}{2\alpha}(\check{\xi} + \mathrm{i}\check{\eta})} - \frac{h}{4\alpha\mathrm{i}}\text{V.P.}\int_{-\infty}^{\infty} \frac{\mathrm{e}^{-\frac{\pi}{h}\check{\eta}\mathrm{i}}}{\sinh\frac{\pi}{2\alpha}\check{\eta}} \frac{d\check{\eta}}{\cos\frac{\pi}{2\alpha}(\check{\xi} + \mathrm{i}\check{\eta})}. \end{aligned} \tag{4.21}$$

In the upper (lower) half plane, where we can utilize the Jordan lemma to evaluate the first (second) integral, each of the respective integrands possesses two sets of poles, zeros of $\sinh\pi\check{\eta}/2\alpha$, $\check{\eta} = \pm 2\alpha m\mathrm{i}$, $m = 0, 1, 2\ldots$, with the respective residues

$$\pm\frac{2\alpha}{\pi} \frac{\mathrm{e}^{-\frac{2\alpha\pi}{h}m}}{\cos\frac{\pi}{2\alpha}\check{\xi}}, \tag{4.22}$$

and zeros of $\cos\,(\pi(\check{\xi}+\mathrm{i}\check{\eta})/(2\alpha))$, $\check{\eta}=\mathrm{i}[\check{\xi}\pm\alpha(2m+1)]$, with the respective residues

$$\mp\frac{2\alpha}{\pi}\frac{\mathrm{e}^{\frac{\pi}{h}[\mp\check{\xi}-\alpha(2m+1)]}}{\cos\,\frac{\pi}{2\alpha}\check{\xi}}. \tag{4.23}$$

It is clear that a significant contribution to (4.21) is made only by the poles, $\check{\eta}=0$ and $\check{\eta}=\mathrm{i}[\check{\xi}\pm\alpha]$ (in the upper and lower half plane, respectively); other residues contain small exponential factors. Therefore, applying the Cauchy residue theorem and taking into account that the first pole lies on the contour of integration, we have

$$\frac{h}{4\alpha\mathrm{i}}\mathrm{V.P.}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{\pm\frac{\pi}{h}\check{\eta}\mathrm{i}}}{\sinh\,\frac{\pi}{2\alpha}\check{\eta}}\frac{d\check{\eta}}{\cos\,\frac{\pi}{2\alpha}(\check{\xi}+\mathrm{i}\check{\eta})}\approx\pm\frac{h}{\cos\,\frac{\pi}{2\alpha}\check{\xi}}\left(\frac{1}{2}-\mathrm{e}^{-\frac{\pi\alpha}{h}\mp\frac{\pi}{h}\check{\xi}}\right), \tag{4.24}$$

which—returning to the original variables ξ and η—gives us a new quadrature formula

$$\mathrm{V.P.}\int_{-\infty}^{\infty}\frac{f(\eta)}{\cos\,\frac{\pi}{2\alpha}(\xi+\mathrm{i}\eta)}d\eta\approx h\sum_{n=-\infty}^{\infty}\frac{1}{\cos\,\frac{\pi}{2\alpha}(\xi+\mathrm{i}nh)}\left[1-2(-1)^n\mathrm{e}^{-\frac{\pi}{h}\alpha}\cosh\frac{\pi}{h}\xi\right]f(nh),$$
$$|\,\mathrm{Re}\;\xi|<\alpha. \tag{4.25}$$

Let us now consider the integrals of the type (4.13). Their generic form is

$$\mathrm{V.P.}\int_{-\infty}^{\infty}\frac{f(\eta)}{\cos\,\frac{\pi}{2\alpha}[\xi+\mathrm{i}\chi(\eta)]}d\eta,\quad|\,\mathrm{Re}\;\xi|<\alpha, \tag{4.26}$$

where $\chi(\eta)$ is a smooth monotone function. We could change the integration variable η to $\chi(\eta)$, reduce (4.26) to the integral of type (4.15), and evaluate the result using a uniform mesh in χ. However, both integrals (4.15) and (4.26) involve the solution of (3.27), and therefore it is more reasonable to evaluate both integrals using the same mesh. Then following the same reasoning as above, (4.26) may be approximated by

$$\mathrm{V.P.}\int_{-\infty}^{\infty}\frac{f(\eta)}{\cos\,\frac{\pi}{2\alpha}[\xi+\mathrm{i}\chi(\eta)]}d\eta\approx\sum_{-\infty}^{\infty}\mathcal{B}_n(\xi)f(nh), \tag{4.27}$$

where the coefficients are given by

$$\mathcal{B}_n(\xi)=\mathrm{V.P.}\int_{-\infty}^{\infty}\frac{\omega_h(\eta-nh)}{\cos\,\frac{\pi}{2\alpha}[\xi+\mathrm{i}\chi(\eta)]}d\eta, \tag{4.28}$$

and the main contributions to (4.28) are made by the zero $\check{\eta}=0\,(\eta=nh)$ of the hyperbolic sine in ω_h and the zero $\check{\eta}=\mathrm{i}a_\pm-nh\,(\eta=\mathrm{i}a_\pm)$ of the cosine-function, where $\xi+\mathrm{i}\chi(\mathrm{i}a_\pm)=\mp\alpha$. The latter equation implies that $\mathrm{i}\chi(\mathrm{i}a_\pm)=-\sin^{-1}(\gamma^{-1}\sin\,a_\pm)=\mp\alpha-\xi$, and therefore we have

$$a_\pm=\sin^{-1}\big[\gamma\,\sin\,(\xi\pm\alpha)\big], \tag{4.29}$$

with $\mathrm{Re}\;a_+>0$ and $\mathrm{Re}\;a_-<0$. Applying to (4.28) the Cauchy residue theorem and noting that the pole $\check{\eta}=0$ lies on the contour of integration, we obtain

$$\frac{h}{4\alpha\mathrm{i}}\mathrm{V.P.}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{\pm\frac{\pi}{h}\check{\eta}\mathrm{i}}}{\sinh\,\frac{\pi}{2\alpha}\check{\eta}}\frac{d\check{\eta}}{\cos\,\frac{\pi}{2\alpha}[\xi+\mathrm{i}\chi(\check{\eta}+nh)]}$$
$$\approx h\left\{\pm\frac{1}{2\cos\,\frac{\pi}{2\alpha}[\xi+\mathrm{i}\chi(nh)]}\mp\frac{\mathrm{e}^{\mp\frac{\pi}{h}(a_\pm+\mathrm{i}nh)}}{\chi'(a_\pm)\sin\,\frac{\pi}{2\alpha}(a_\pm+\mathrm{i}nh)}\right\}, \tag{4.30}$$

where $\chi'(a_\pm) = \cos a_\pm / \sqrt{\gamma^2 - \sin^2 a_\pm}$. Returning to the original variable η, the resulting quadrature formula is

$$\text{V.P.}\int_{-\infty}^{\infty} \frac{f(\eta)}{\cos\,\frac{\pi}{2\alpha}[\xi + \mathrm{i}\chi(\eta)]}d\eta \approx h \sum_{n=-\infty}^{\infty} \left\{ \frac{1}{\cos\,\frac{\pi}{2\alpha}[\xi + \mathrm{i}\chi(nh)]} - (-1)^n \left[\frac{\mathrm{e}^{-\frac{\pi}{h}a_+}}{\chi'(a_+)\sin\,\frac{\pi}{2\alpha}(a_+ + inh)} + \frac{\mathrm{e}^{\frac{\pi}{h}a_-}}{\chi'(a_-)\sin\,\frac{\pi}{2\alpha}(a_- + inh)} \right] \right\} f(nh), \quad |\,\mathrm{Re}\;\xi| < \alpha. \tag{4.31}$$

The first terms on the right of both (4.25) and (4.31) effect the trapezoidal rule, and the second give a correction. When Re $\xi \le \alpha - 10^{-6}$, $\Psi_i^{\pm}(\omega)$ can be approximated as $[\Psi_i^{\pm}(\omega - 0.05) + \Psi_i^{\pm}(\omega + 0.05)]/2$.

5. Code testing. Using the above considerations, we have developed a new code for evaluating the Rayleigh reflection and transmission coefficients for elastic wedges (see Appendix E). The integral equations we solve have the form of the Fredholm equations of the second kind, but it can be shown that the operators involved are not Fredholm (cf. the statements in [5, p. 251]. We possess no analytical proof that these equations can be solved uniquely. Nevertheless, our code produces a solution, and below we describe verification tests that allow us to state with confidence that when transformed back to the physical space this solution satisfies the original diffraction problem. We also describe successful validation tests, comparing output of our code with published numerical and experimental data. Of course, the positive outcomes of these tests do not constitute a theoretical proof that the code is correct. Note that throughout this section we characterize materials by their Poisson ratios ν, where $\gamma = \sqrt{(1-2\nu)/[2(1-\nu)]}$.

5.1. Code verification. We have designed verification tests to establish that the computed functions $\Psi_i(\omega)$ are the solutions of the original physical problem; in particular, that they

(i) are bounded at imaginary infinity;
(ii) are analytic at the boundary of the strip Re $\omega \,\epsilon\, I$;
(iii) possess only physically meaningful singularities.

The property (i) is confirmed by direct examination of the computed functions $\widetilde{x}^-(\eta)$ and $\widetilde{y}^+(\eta)$ divided by $\exp(-|\eta|)$. At large $|\eta|$ the ratios appear to behave as $O(1)$. It follows that the amplitudes $\Psi_i(\omega)$ obtained by the analytical continuation must be bounded at the imaginary infinity, $\eta = \mathrm{Im}\;\omega \to \infty$.

Since the last step in the analytical continuation is carried out strip-by-strip, all 2α wide, there is no guarantee that any computed $\Psi_i(\omega)$ should be smooth at the boundaries of the initial strip Re $\omega \,\epsilon\, I$. However, examination of the numerical data used to plot Figures 5.1–5.3 confirm that our approximations *are* smooth. It follows that the property (ii) is satisfied. Interestingly, when attempting to solve an incorrect problem, with the constants $c_1^{\pm}$ put to zero, the computed $\Psi_i(\omega)$ themselves jump at the boundaries by about 10^{-2}.

Similarly to (ii), the property (iii) should be satisfied by the Sommerfeld amplitudes of the solutions of the original wedge diffraction problem, but it is not obvious that the computed solutions of the corresponding functional equations should satisfy it as well. Indeed, the way they are constructed assures that the computed amplitudes $\Psi_i(\omega)$ have physically meaningful poles, and since the functional equations that are used to effect the analytical continuation involve reflection coefficients $\mathrm{r}_{jk}(\omega)$, with

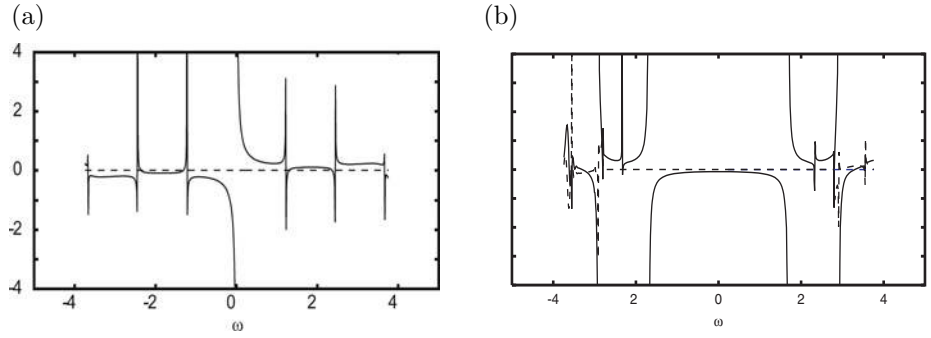


FIG. 5.1. *The computed Sommerfeld amplitudes:* (a) Re $\Psi_0(\omega)$—*dashed line and* Im $\Psi_0(\omega)$—*solid line,* (b) Re $\Psi_1(\omega)$—*dashed line and* Im $\Psi_1(\omega)$—*solid line. Wedge angle* $2\alpha = 70°$, $I = [0.96, 2.18]$, *Poisson's ratio* $\nu = 0.25$, *incident wave—compressional and* $\theta^{inc} = 0°$.

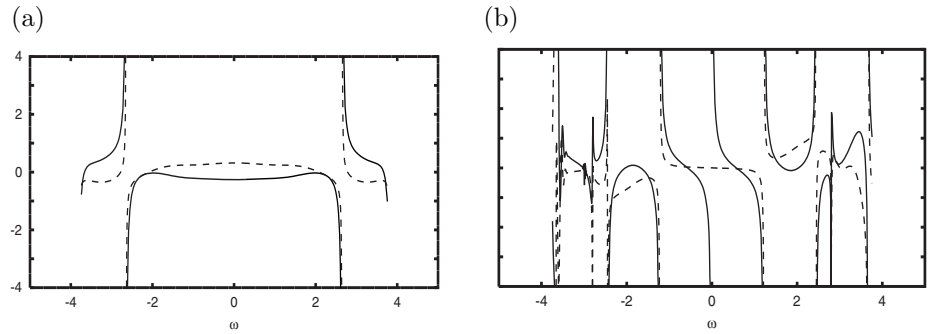


FIG. 5.2. *The computed Sommerfoeld amplitudes:* (a) Re $\Psi_0(\omega)$—*dashed line and* Im $\Psi_0(\omega)$—*solid line,* (b) Re $\Psi_1(\omega)$—*dashed line and* Im $\Psi_1(\omega)$—*solid line. Wedge angle* $2\alpha = 70°$, $I = [0.96, 2.18]$, *Poisson's ratio* $\nu = 0.25$, *incident wave—shear and* $\theta^{inc} = 0°$.

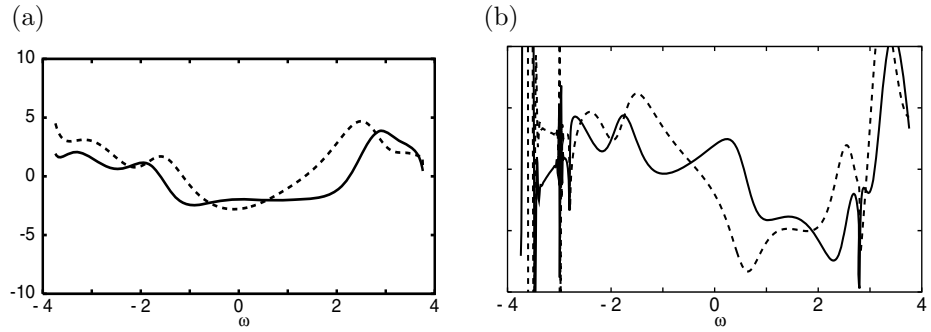


FIG. 5.3. *The computed Sommerfeld amplitudes:* (a) Re $\Psi_0(\omega)$—*dashed line and* Im $\Psi_0(\omega)$—*solid line,* (b) Re $\Psi_1(\omega)$—*dashed line and* Im $\Psi_1(\omega)$—*solid line. Wedge angle* $2\alpha = 70°$, $I = [0.96, 2.18]$, *Poisson's ratio* $\nu = 0.25$, *incident wave—Rayleigh and* Re $\theta^{inc} = -\alpha$.

the branch point at $\omega = \theta_h$ and poles at $\omega = \pm i\beta_R$ (see (3.14) and Appendix A), they possess physically meaningful branch points and Rayleigh poles too. However, by the same token, the analytical continuation scheme *might* endow these amplitudes with *extra, physically meaningless singularities.* Remarkably, when the incident wave is compressional or shear at $\omega = -\alpha \pm i\beta_R$ all our computed residues are of order

10^{-7}, i.e., are numerical zeros. Thus, the computed Sommerfeld amplitudes possess no Rayleigh poles corresponding to physically meaningless Rayleigh waves incoming from infinity. Furthermore, Figures 5.1–5.2, which respectively relate to a purely symmetric and purely asymmetric case, confirm that inside a neighborhood of zero which includes the strip Re $\omega \epsilon I$, the computed Sommerfeld amplitudes possess the symmetries described in (3.1); that is, $\Psi_0^+(\omega)$ and $\Psi_1^-(\omega)$ are odd, while $\Psi_1^+(\omega)$ and $\Psi_0^-(\omega)$ are even. (Outside this neighborhood, the symmetries are not apparent in Figures 5.1– 5.3 due to the accumulation of numerical errors.) As we show in Appendix D, such symmetries imply the absence of physically meaningless branch points.

The properties (i), (ii), and (iii) of the computed Sommerfeld amplitudes respectively imply that they possess all the properties expected of the Sommerfeld amplitudes of the solutions $\psi_i(kr, \theta)$ of the original diffraction problem, so that their corresponding Sommerfeld integrals satisfy (i) the Helmholtz equations and correct tip condition; (ii) zero stress boundary conditions; and (iii) radiation conditions (which exclude nonphysical Rayleigh or head waves incoming from infinity).

Figure 5.1 provides one more confirmation that the computed functions $\Psi_i(\omega)$ are the Sommerfeld amplitudes of solutions $\psi_i(kr, \theta)$ of the original wedge problem: It shows that for the symmetric compressional wave incidence, both $\Psi_i(\omega)$ are imaginary, and therefore the corresponding displacements are real. This is consistent with the physics of the problem, since unlike with the symmetric shear wave incidence, there is no total internal reflection, that is, no imaginary displacement component.

Finally, a numerical stability of the scheme is ascertained by the fact that different choices of adjustable function $\sigma(\omega)$ in (3.16) all give similar results (see Appendix F).

5.2. Code validation. Our first validation results are presented in Tables 5.1 and 5.2, where the approximate values of amplitudes and phases of reflection and transmission coefficients R^{ref} and R^{tran} as computed with our code are compared with numerical results of [11]. Each Fujii's column contains values corresponding to different choices of an adjustable parameter. The parameter allows one to evaluate singular integrals on the real axis by moving the poles away from the axis into the complex plane. This is equivalent to employing the radiation condition at infinity in the form of the limiting absorption principle. From the physical point of view, the singularities cannot be moved too far. However, when they are too close the evaluation algorithm becomes numerically unstable. The top rows in the tables are obtained with

TABLE 5.1
Rayleigh reflection coefficients computed with our code and Fujii's (see [11]*);* $\nu = 0.25$.

Wedge	$\lvert R^{\text{ref}}\rvert$		arg R^{ref}	
Angle	Fujii	This paper	Fujii	This paper
	0.50552		−169.87°	
50°	0.49924		−169.54°	
	0.49278	0.47427	−169.07°	−161.4°
	0.05257		170.53°	
	0.05252		170.14°	
150°	0.05236	0.05197	169.85°	170.5°
	0.05217		169.65°	
	0.05196		169.53°	
	0.05151		169.43°	

TABLE 5.2
Rayleigh transmission coefficients computed with our code and Fujii's (see [11]*);* $\nu = 0.25$.

Wedge	$\lvert R^{\text{tran}}\rvert$		arg R^{tran}	
Angle	Fujii	This paper	Fujii	This paper
	0.49123		$-32.59°$	
50°	0.48372		$-32.84°$	
	0.47552	0.55189	$-33.00°$	$-26.9°$
	0.78940		52.79°	
	0.78899		52.80°	
150°	0.78866	0.78942	52.81°	52.9°
	0.78842		52.82°	
	0.78825		52.84°	
	0.78800		52.86°	

the parameter values that correspond to a more physically meaningful situation, and the bottom ones, with the values that give better numerical stability. The tables demonstrate that for the larger wedge angles the agreement with our computations is quite good, but for the smaller ones our values lie outside Fujii's range. This is not surprising, because when the wedge angles are small there are many multiply reflected waves, and the residues of many resulting poles are large. For this reason, when the wedge angles are small, the present version of our code loses its numerical stability.

To continue, in Figure 5.4(a) and (b) we present our Rayleigh reflection and transmission coefficients as functions of the wedge angle, computed for $\nu = 0.234$. They fit Fujii's numerical and experimental data extremely well (see our Figure 5.5 or [11, Figure 7]. Note that on Fujii's plots the solid lines represent his numerical results, and discrete points, his experimental data.) Indeed, for the wedge angles between 45° and 150° we cannot put the results on the same graph—there is no visible difference. Note that the jumps in the phase of the reflection coefficient that take place at the wedge angles of about 45° and 145° are from 180° to −180° and −180° to 180°, respectively, and therefore no jumps in physical quantities take place. For the wedge angles between 150° and 180° the reflection coefficients are practically zero. This is understandable, because when the wedge angle is 180° there is no reflection. In this region, the phases of our reflection coefficients differ from Fujii's, but the limiting value of 90° agrees with the one obtained by [13]. It appears that in this region Fujii's scheme loses its stability.

The amplitude curves reported by Budaev and Bogy [8] are the same as ours (see their Figure 6 and our Figure 5.4(a)), but for the larger wedge angles, the phase of their reflection coefficient is somewhat different—see Figure 5.4(b). The discrepancy might not be crucial, because at these angles the amplitudes of the reflection coefficients are very small, but the problem is indicative of numerical instability. Note that the results on the wedge angles greater than 180° as presented by [6] are incorrect—see their errata [9]. Note too that even though Poisson's ratio used by Budaev and Bogy [8] is $\nu = 0.294$, the above comparison is valid: The coefficients should not be effected by a small difference in ν (see, e.g., Figure 5.6.)

We finish this section by comparing our computed Rayleigh reflection and transmission coefficients for the quarter space with Gautesen's [13]. On taking into account that Gautesen's coefficients are complex conjugates of ours and thus our phases must have the opposite sign, the agreement between the calculations is very good (see Figure 5.6).

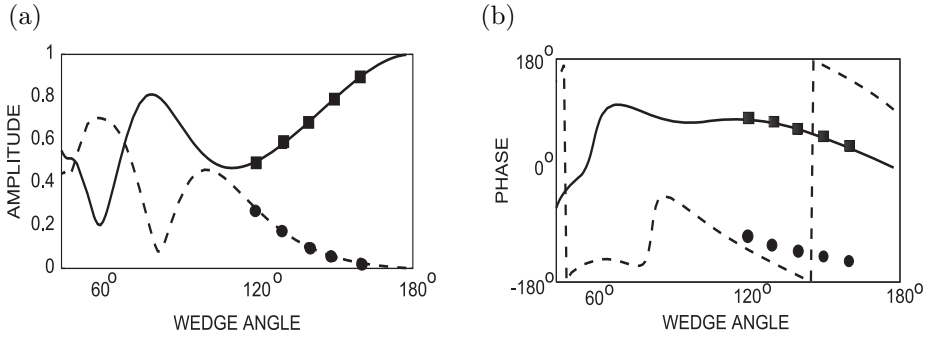


FIG. 5.4. *Rayleigh transmission coefficients (solid line) and reflection coefficients (dashed line) computed with our code versus the coefficients computed with Budaev and Bogy's code (squares and circles, respectively—see Budaev and Bogy* [8]*, Figure* 6*). Poisson's ratio* $\nu = 0.234$*, incident wave—Rayleigh and* Re $\theta^{inc} = -\alpha$.

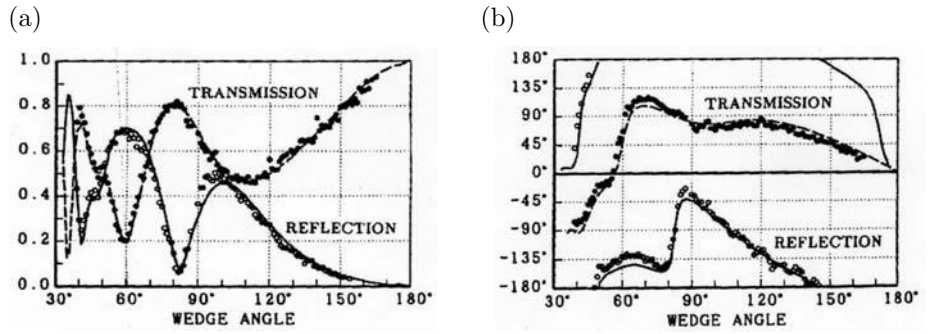


FIG. 5.5. *The Fujii' computed (solid line) and experimental (dots) transmission and reflection coefficients:* (a) *amplitudes,* (b) *phases. Poisson's ratio* $\nu = 0.234$*, incident wave–Rayleigh and* Re $\theta^{inc} = -\alpha$*. Reproduced from Figure* 7 *in* [11].

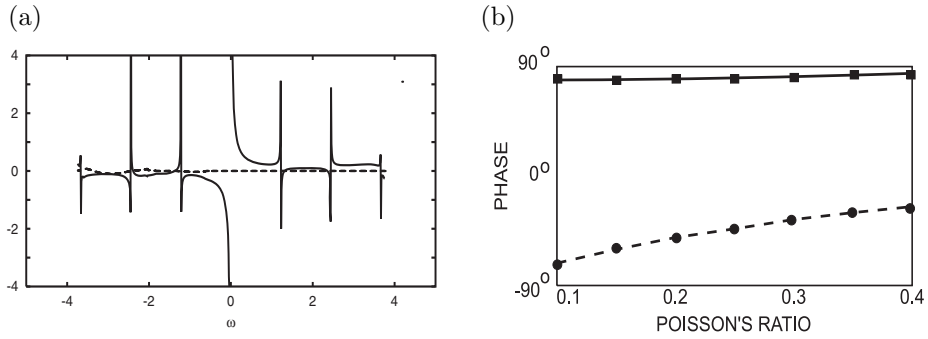


FIG. 5.6. *Rayleigh transmission coefficients (solid line and squares) and reflection coefficients (dashed line and circles) computed respectively with our code and Gautesen's code (see* [13, *Figures* 3 *and* 4]*):* (a) *amplitudes,* (b) *phases. Wedge angle* 90°*, incident wave—Rayleigh and* Re $\theta^{inc} = -\alpha$.

6. Conclusions. We have studied the properties of the underlying integral operators and developed a new numerical schedule for solving the singular integral problem that arises in Budaev and Bogy's approach to diffraction by 2D traction-free isotropic elastic wedges. We have also developed new quadrature formulas for evaluating the

singular convolution-type integrals that are utilized in this approach. Although the analytical justification of the method is not entirely rigorous, the code has undergone a series of stringent internal verification tests directed at establishing that it solves the original physical problem, as well as validation tests against numerical and experimental results reported by other authors. It appears to be successful when simulating diffraction by wedges of angles between 40° and 178° of plane incident waves, compressional or shear. When the incident wave is a Rayleigh, the lower limit of applicability goes up to 45°.

Appendix A. Nomenclature.

$$a(\eta) = -\mathrm{i}\tanh 2\eta,$$

$$b(\eta) = \frac{2\mathrm{i}\sinh\eta\sqrt{\gamma^2+\sinh^2\eta}}{\cosh 2\eta},$$

$$d(\eta) = 1 - \frac{\tanh^2 2\eta}{\chi'(\eta)},$$

$$e_1^+(\omega) = -\tan\alpha\sin\,2\omega + \cos\,2\omega,$$

$$e_1^-(\omega) = \tan\alpha\cos\,2\omega + \sin\,2\omega,$$

$$e_2^+(\omega) = \tan\alpha\cos\,2\omega\frac{\sin\,\omega}{\sqrt{\gamma^2-\cos^2\omega}} + \sin\,2\omega,$$

$$e_2^-(\omega) = \tan\alpha\sin\,2\omega - \cos\,2\omega\frac{\sin\,\omega}{\sqrt{\gamma^2-\cos^2\omega}},$$

$$g(\omega) = \cos^{-1}(\gamma^{-1}\cos\,\omega),$$

$$\mathcal{H}f(\eta) = \frac{1}{2\alpha\mathrm{i}}\mathrm{V.P.}\int_{-\infty}^{\infty}\frac{f(t)\mathrm{d}t}{\sinh[\frac{\pi}{2\alpha}(t-\eta)]},$$

$$(\mathcal{H}^{-1}f)(\eta) = \frac{1}{2\alpha\mathrm{i}}\mathrm{V.P.}\int_{-\infty}^{\infty}\coth\frac{\pi}{2\alpha}(t-\eta)f(t)\,dt,$$

$$\overline{\mathcal{H}}f(\eta) = \frac{1}{2\alpha\mathrm{i}}\mathrm{V.P.}\int_{-\infty}^{\infty}\frac{f(t)\chi'(t)\mathrm{d}t}{\sinh\frac{\pi}{2\alpha}[\chi(t)-\chi(\eta)]},$$

$$Kf(\eta) = \frac{1}{2\alpha\mathrm{i}}\int_{-\infty}^{\infty}\left\{\frac{\tanh^2 2t}{\chi'(t)\sinh\frac{\pi}{2\alpha}(t-\eta)} - \frac{\tanh 2t\tanh 2\eta}{\sinh\frac{\pi}{2\alpha}[\chi(t)-\chi(\eta)]}\right\}f(t)\,dt,$$

$$q_0^+(\eta) = r_2^+(\eta) - a(\eta)\overline{\mathcal{H}}r_1^+(\eta),$$

$$q_1^+(\eta) = -\tan\alpha\frac{\cosh\eta}{\cosh 2\eta} - \frac{\tanh 2\eta}{2\alpha}\mathrm{V.P.}\int_{-\infty}^{\infty}\frac{\gamma\cosh\tau d\tau}{(1+2\gamma^2\sinh^2\tau)\sinh\frac{\pi}{2\alpha}[\tau-\chi(\eta)]},$$

$$q_0^-(\eta) = r_1^-(\eta) - b(\eta)\mathcal{H}r_2^-(\eta),$$

$$q_1^-(\eta) = -\tan\alpha\frac{\cosh\eta}{\chi'(\eta)\cosh 2\eta} - \frac{\tanh 2\eta}{2\alpha\chi'(\eta)}\mathrm{V.P.}\int_{-\infty}^{\infty}\frac{\cosh\tau\,d\tau}{\cosh 2\tau\,\sinh\frac{\pi}{2\alpha}(\tau-\eta)},$$

$$\begin{aligned} r_1^\pm(\eta) = &-\left[\widehat{\Psi}_0^\pm\left(g\left(\frac{\pi}{2}+\mathrm{i}\eta\right)+\alpha\right) \pm \widehat{\Psi}_0^\pm\left(g\left(\frac{\pi}{2}+\mathrm{i}\eta\right)-\alpha\right)\right] \\ &- b(\eta)\left[\widehat{\Psi}_1^\pm\left(\frac{\pi}{2}+\alpha+\mathrm{i}\eta\right) \pm \widehat{\Psi}_1^\pm\left(\frac{\pi}{2}-\alpha+\mathrm{i}\eta\right)\right], \\ r_2^\pm(\eta) = &-a(\eta)\left[\widehat{\Psi}_0^\pm\left(g\left(\frac{\pi}{2}+\mathrm{i}\eta\right)+\alpha\right) \mp \widehat{\Psi}_0^\pm\left(g\left(\frac{\pi}{2}+\mathrm{i}\eta\right)-\alpha\right)\right] \\ &- \left[\widehat{\Psi}_1^\pm\left(\frac{\pi}{2}+\alpha+\mathrm{i}\eta\right) \mp \widehat{\Psi}_1^\pm\left(\frac{\pi}{2}-\alpha+\mathrm{i}\eta\right)\right], \end{aligned}$$

$$\mathrm{r}_{11}(\omega) = -\mathrm{r}_{22}(\omega) = \frac{2\sin\,2\omega\,\cos\,\omega\sqrt{\gamma^2-\cos^2\,\omega}-\cos^2\,2\omega}{\Delta(\omega)},$$
$$\mathrm{r}_{12}(\omega) = -\frac{4\cos\,2\omega\,\cos\,\omega\sqrt{\gamma^2-\cos^2\,\omega}}{\Delta(\omega)},$$
$$\mathrm{r}_{21}(\omega) = -\frac{2\sin\,2\omega\,\cos\,2\omega}{\Delta(\omega)},$$
$$\Delta(\omega) = \cos^2\,2\omega + 2\,\sin\,2\omega\,\cos\,\omega\sqrt{\gamma^2-\cos^2\,\omega},$$
$$\chi(\eta) = \sinh^{-1}(\gamma^{-1}\sinh(\eta)),$$
$$\chi'(\eta) = \frac{\cosh\eta}{\sqrt{\gamma^2+\sinh^2\eta}}.$$

Note that the Rayleigh function $\Delta(\omega)$ has the purely imaginary root $\mathrm{i}\beta_{\mathrm{R}}$, with $\beta_{\mathrm{R}} > 0$.

Appendix B. The tip asymptotics of the elastic potentials and asymptotics of the Sommerfeld amplitudes at infinity. The behavior of solutions of the elliptic problems in regions with piecewise smooth boundaries has been studied by many authors (see [21] and references therein). A rigorous theory has been developed after a breakthrough by Kondrat'ev [13], who constructed and justified the field asymptotics in the vicinity of conical and edge points. The theory implies that the solution of the underlying Lamé problem must have the asymptotic expansion

$$\mathbf{u}(kr,\theta) \sim \sum_{\ell,m=0}^{\infty} (kr)^{q_m+2\ell} \sum_{n=0}^{N_m-1} \mathbf{u}_{\ell,m,n}(\theta)\,(\ln kr)^n, \quad kr \to 0, \tag{B.1}$$

where for one m, $q_m = 0$ and $N_m = 1$ (otherwise, the tip conditions are violated); for any other m, Re $q_m > 0$ and q_m is a root of a transcendental equation, with a natural number N_m being its multiplicity.

The asymptotic expansions (B.1) can be differentiated and substituted into the boundary conditions. Therefore, using (2.4), similar expansions may be written for the elastodynamic potentials $\psi_i^{\pm}(kr,\theta)$ as

$$\psi_i^{\pm}(kr,\theta) \sim \sum_{\ell,m=0}^{\infty} (kr)^{p_m^{\pm}+2\ell} \sum_{n=0}^{N_m^{\pm}-1} \psi_{i,\ell,m,n}^{\pm}(\theta)(\ln kr)^n, \quad kr \to 0. \tag{B.2}$$

Let us arrange the sets of exponents $\{p_m^+,\ m = 0,1,\dots\}$ and $\{p_m^-,\ m = 0,1,\dots\}$, each in order of the increasing real part. Following [19], these sets may be described as follows: Each contains 1, while any other element is a root of the transcendental equation

$$(p^{\pm}+1)\,\sin\,2\alpha \pm \sin\,2\alpha(p^{\pm}+1) = 0 \tag{B.3}$$

and satisfies condition

$$\mathrm{Re}\ p^{\pm} > -1 \tag{B.4}$$

(otherwise, the tip conditions are violated). Note that if in (B.1) a $q_m \neq 0$, then in (B.2) the corresponding p_m^+ or p_m^- equals $q_m - 1$, but applying the nabla operator to the term with $q_m = 0$ and $\ell = 0$ always gives us zero—because the corresponding $N_m = 1$. Thus, for $q_m = 0$, only the next, r^2, term in (B.1) gives rise to a nonzero term

in (B.2). The corresponding $p_m^{\pm} = 1$. Note that while for any wedge angle there exists an m such that $p_m^- = 0$ is a solution of (B.3), in our range of wedge angles $2\alpha \in (0, \pi)$, all p_m^+ differ from zero. The full set of solutions of the transcendental equations (B.3) is described in [19]. The main facts can be summarized in plots representing roots of the transcendental equations as functions of the wedge angle 2α (see, e.g., [22]). The analysis of these plots shows that in our range of wedge angles, the tip conditions are assured for those nonnegative exponents with minimal real part that are either 1 or else are solutions of the corresponding transcendental equations, with real part less than or equal to 1. In other words, all leading exponents in (B.1), that is, the exponents with the minimal real part, lie in the strip

$$0 \le \operatorname{Re} p_0^{\pm} \le 1. \tag{B.5}$$

The root loci in [22] also show that at one wedge angle, $2\alpha_* \approx 0.8\pi$, we have a degeneracy: In the corresponding symmetric problem, the exponent with the minimal real part, $p_*^+ \approx 0.76$, is a multiple root of the corresponding transcendental equation (B.3). The corresponding multiplicity $N_*^+ = 2$. There are no multiple roots p_0^- which have the minimal real part and simultaneously satisfy (B.5). It follows that for α_*, the leading terms in (B.1) are

$$\psi_{i,0,1,1}^+ (kr)^{p_*^+} \ln kr = O\big((kr)^p\big), \quad 0 < p < p_*^+. \tag{B.6}$$

The behavior of potentials $\psi_i^{\pm}(kr, \theta)$ in the vicinity of the wedge tip dictates the asymptotic behavior of the Sommerfeld amplitudes $\Psi_i^{\pm}(\omega)$ at infinity: For example, it is easy to check that for any small $\varepsilon > 0$, as $\operatorname{Im}\ \omega \to \infty$, $\Psi_i^{\pm}(\omega)$ have expansions

$$\Psi_i^{\pm}(\omega) \sim \sum_{0 \le \operatorname{Re}\ p_m^{\pm} \le 1} \Psi_{im}^{\pm} \mathrm{e}^{\mathrm{i}p_m^{\pm}\omega} + O(\mathrm{e}^{\mathrm{i}(1+\varepsilon)\omega}), \qquad \alpha \ne \alpha_*,\ 0 < \alpha < \frac{\pi}{2},$$

$$\Psi_i^+(\omega) \sim \overline{\Psi}_{i*} \omega \mathrm{e}^{\mathrm{i}p_*^+\omega} + \Psi_{i*} \mathrm{e}^{\mathrm{i}p_*^+\omega} + \Psi_{i1} \mathrm{e}^{\mathrm{i}\omega} + O(\mathrm{e}^{\mathrm{i}(2+\varepsilon)\omega}), \qquad \alpha = \alpha_*. \tag{B.7}$$

Note that in the symmetric case, the expansions contain no constant terms (so that the leading exponents are 1 and possibly a solution of (B.3), with the real part in $(0, 1)$), but these may be present in the antisymmetric case (so that the leading exponents there are 0 and 1).

Appendix C. The integral equations for one unknown.

Symmetric problem. The two final regularized integral equations to solve are

$$\widetilde{y}_i^+(\eta) + \widetilde{L}^+ \widetilde{y}_i^+(\eta) = \widetilde{q}_i^+(\eta), \quad i = 0, 1, \tag{C.1}$$

where $\widetilde{L}^+$ is an operator with a smooth kernel

$$(\widetilde{L}^+ u)(\eta) = -\frac{1}{4\alpha^2} \int_{-\infty}^{\infty} \frac{l^+(\eta, t) \tanh 2t}{\sqrt{d(t)}\sqrt{d(\eta)} \cosh \frac{\pi}{2\alpha}\eta} u(t)\, dt, \tag{C.2}$$

with

$$l^+(\eta, t) = \mathrm{V.P.} \int_{-\infty}^{\infty} \frac{\cosh \frac{\pi}{2\alpha}\tau}{\sinh \frac{\pi}{2\alpha}(\tau - \eta)} \left\{ \frac{\tanh 2t}{\chi'(t) \sinh \frac{\pi}{2\alpha}(t - \tau)} - \frac{\tanh 2\tau}{\sinh \frac{\pi}{2\alpha}[\chi(t) - \chi(\tau)]} \right\} d\tau.$$

The respective right-hand sides of (C.1) are given by

$$\widetilde{q}_i^+(\eta) = \frac{1}{2\alpha \mathrm{i} \sqrt{d(\eta)} \cosh \frac{\pi}{2\alpha}\eta} \mathrm{V.P.} \int_{-\infty}^{\infty} \frac{\cosh \frac{\pi}{2\alpha}t}{\sinh \frac{\pi}{2\alpha}(t - \eta)} q_i^+(t)\, dt, \tag{C.3}$$

with $q_i^+(\eta)$ given in Appendix A. On solving (C.1), $y^+(\eta)$ is obtained using

$$y^+(\eta) = d^{-1/2}(\eta)[\widetilde{y_0^+}(\eta) + c_1^+ \widetilde{y_1^+}(\eta)], \tag{C.4}$$

where $c_1^+ = -\lambda_0^+/\lambda_1^+$ and we have

$$\lambda_i^+ = \int_{-\infty}^{\infty} \big[\widetilde{y_i^+}(t) d^{-1/2}(t)(B^+\mathbb{I})(t) - q_i^+(t)\big]\,dt, \quad , i = 0, 1,$$
$$(B^+\mathbb{I})(t) = \frac{\gamma^2 \tanh 2t}{2\alpha \mathrm{i}} \int_{-\infty}^{\infty} \frac{\sinh 2(\chi(t)+\tau)}{1 + 2\gamma^2 \sinh^2(\tau + \chi(t))} \cdot \frac{d\tau}{\sinh \frac{\pi}{2\alpha}\tau}.$$

Antisymmetric problem. Analogously to the symmetric case, the two final regularized integral equations to solve are

$$\widetilde{x_i^-}(\eta) + \widetilde{L}^- \widetilde{x_i^-}(\eta) = \widetilde{q_i^-}(\eta), \quad i = 0, 1, \tag{C.5}$$

where the integral operator is

$$(\widetilde{L}^- u)(\eta) = -\frac{1}{4\alpha^2} \int_{-\infty}^{\infty} \frac{l^-(\eta, t) \tanh 2t}{\sqrt{d(t)}\sqrt{d(\eta)} \cosh \frac{\pi}{2\alpha}\chi(\eta)} u(t)\,dt, \tag{C.6}$$

with

$$l^-(\eta, t) = -\mathrm{V.P.} \int_{-\infty}^{\infty} \frac{\cosh \frac{\pi}{2\alpha}\tau}{\sinh \frac{\pi}{2\alpha}[\chi(\eta) - \tau]} \left\{ \frac{\tanh 2t}{\sinh \frac{\pi}{2\alpha}(\chi(t) - \tau)} + \frac{\tanh 2\chi^{-1}(\tau)(\chi^{-1})'(\tau)}{\sinh \frac{\pi}{2\alpha}[\chi^{-1}(\tau) - t]} \right\} d\tau.$$

The respective right-hand sides of (C.5) are given by

$$\widetilde{q_i^-}(\eta) = \frac{1}{2\alpha \mathrm{i}\sqrt{d(\eta)} \cosh \frac{\pi}{2\alpha}\chi(\eta)} \mathrm{V.P.} \int_{-\infty}^{\infty} \frac{\cosh \frac{\pi}{2\alpha}t}{\sinh \frac{\pi}{2\alpha}[t - \chi(\eta)]} q_i^-(\chi^{-1}(t))\,dt, \tag{C.7}$$

with $q_i^-(\eta)$ given in Appendix A. As above, x^- is obtained on solving (C.5) using

$$x^-(\eta) = d^{-1/2}(\eta)[\widetilde{x_0^-}(\eta) + c_1^- \widetilde{x_1^-}(\eta)], \tag{C.8}$$

where $c_1^- = -\lambda_0^-/\lambda_1^-$ and we have

$$\lambda_i^- = \int_{-\infty}^{\infty} \big[\widetilde{x_i^-}(t) d^{-1/2}(t)(B^-\chi')(t) - q_i^-(t)\chi'(t)\big]\,dt, \quad i = 0, 1,$$
$$(B^-\chi')(t) = \frac{\tanh 2t}{2\alpha \mathrm{i}} \int_{-\infty}^{\infty} \frac{\tanh 2(t+\tau)\,d\tau}{\sinh \frac{\pi}{2\alpha}\tau}.$$

Appendix D. The branch points of $\Psi_i(\omega)$. Let us show that the Sommerfeld amplitudes $\Psi_i^+(\omega)$ that satisfy the functional equations (3.5) and conditions (3.1) can have only physically meaningful branch points. For simplicity of presentation, let us assume that $\theta_h < 2\alpha$ (in the opposite case, a slightly more involved argument still goes through.) Using the branch points of $g(\omega)$, the only branch points that $\Psi_1^+(\omega)$ can have inside (2.8) are $\pm(\pi + \alpha - \theta_h)$. The corresponding branch cuts run along the segments $[-\pi - \alpha - \theta_h,\ -\pi - \alpha + \theta_h]$ and $[\pi + \alpha - \theta_h,\ \pi + \alpha + \theta_h]$. Note that the branch points $\pm(\pi + \alpha + \theta_h)$ lie outside the physical region (2.8). The analogous branch points of $\Psi_0^+(\omega)$ can be only $\pm(\pi + \alpha \pm \mathrm{i}\cosh^{-1}(1/\theta_h))$, with the

cuts along the segments $[-\pi-\alpha+\mathrm{icosh}^{-1}(1/\theta_h),\ -\pi-\alpha-\mathrm{icosh}^{-1}(1/\theta_h)]$ and $[\pi+\alpha+\mathrm{icosh}^{-1}(1/\theta_h),\ \pi+\alpha-\mathrm{icosh}^{-1}(1/\theta_h)]$.

Indeed, all possible branch points of $g(\omega)$ are $\pi n\pm\theta_h$ (see (3.7)). In principle, applying (3.5), they could generate many branch points in $\Psi_1^+(\omega)$ which have no physical interpretation. Let us start by showing that $-\alpha+\theta_h$ is not a branch point: Let us use the fact that $\Psi_0^+(\omega)$ is odd to rewrite the functional equation (3.5) as

(D.1)
$$t_{11}(\omega+\alpha)\{\Psi_0^+[\alpha+g(\omega+\alpha)]-\Psi_0^+[\alpha-g(\omega+\alpha)]\}+t_{12}(\omega+\alpha)[\Psi_1^+(\omega+2\alpha)+\Psi_1^+(\omega)]=Q_1^+,$$

(D.2)
$$t_{21}(\omega+\alpha)\{\Psi_0^+[\alpha+g(\omega+\alpha)]+\Psi_1^+[\alpha-g(\omega+\alpha)]\}+t_{22}(\omega+\alpha)[\Psi_1^+(\omega+2\alpha)-\Psi_1^+(\omega)]=Q_1^+.$$

In the vicinity of θ_h, there exist the constants a_n such that we have

$$g(\omega+\alpha)=\sum_{n=0}^{\infty}a_n(\omega+\alpha-\theta_h)^{n+1/2}. \tag{D.3}$$

Since $\Psi_0^+[\alpha+g(\omega+\alpha)]-\Psi_0^+[\alpha-g(\omega+\alpha)]$ is odd in g, there also exist constants A_n such that this function has the expansion

$$\Psi_0^+[\alpha+g(\omega+\alpha)]-\Psi_0^+[\alpha-g(\omega+\alpha)]=\sum_{n=0}^{\infty}A_n(\omega+\alpha-\theta_h)^{n+1/2}. \tag{D.4}$$

On the other hand, there exist constants b_n such that we can write

$$t_{11}(\omega+\alpha)=\sum_{n=0}^{\infty}b_n(\omega+\alpha-\theta_h)^{n-1/2}. \tag{D.5}$$

Thus, the first term on the left-hand side of (D.1) contains no branch points. The coefficient t_{12} has no branch points either. It follows that there are no branch points in (D.1) at all.

Let us move on to (D.2). The function $\Psi_0^+[\alpha+g(\omega+\alpha)]+\Psi_0^+[\alpha-g(\omega+\alpha)]$ is an even function of g, and therefore there exist constants B_n such that we can write

$$\Psi_0^+[\alpha+g(\omega+\alpha)]+\Psi_0^+[\alpha-g(\omega+\alpha)]=\sum_{n=0}^{\infty}B_n(\omega+\alpha-\theta_h)^{n}. \tag{D.6}$$

Since the coefficients $t_{21}(\omega)$ and $t_{22}(\omega)$ have no branch points, there are no branch points in (D.2). It follows that the point $-\alpha+\theta_h$, which does not have any physical interpretation, is not a branch point of the function $\Psi_1^+(\omega)$. Similarly, it can be shown that the points $-\alpha-\theta_h$ and $\alpha\pm\theta_h$ are not branch points of $\Psi_1^+(\omega)$. Analogous considerations apply in the antisymmetric case.

We conclude that the branch points of the Sommerfeld amplitudes of the solution of the original problem that lie in the physical region (2.8), and therefore give rise to physical waves, lie outside the strip Re $\omega\,\epsilon\, I$. This means that they do not have to be taken into account in the functional equations for $\widetilde{\Psi}_i^{\pm}(\omega)$ or, by the same token, in the resulting singular integral problem. On the other hand, we have no theoretical proof that the $\Psi_i^{\pm}(\omega)$ that we eventually compute have only physical branch points in the physical region (2.8). We can confirm this fact only by carrying out numerical tests.

Appendix E. The Rayleigh reflection and transmission coefficients. When evaluating (2.5), a pole $\theta_{1\mathrm{R}}^{\mathrm{inc}} = \alpha - \mathrm{i}\beta_{\mathrm{R}}$ of $\Psi_1(\omega)$, with $\beta_{\mathrm{R}} > 0$, corresponds to a plane wave with the phase factor

$$\mathrm{e}^{\mathrm{i}kr\cos\,(\theta-\theta_{1\mathrm{R}}^{\mathrm{inc}})} = \mathrm{e}^{\mathrm{i}kr\cos\,(\theta-\alpha)\cosh\beta_{\mathrm{R}}}\mathrm{e}^{-kr\sin\,(\alpha-\theta)\sinh\beta_{\mathrm{R}}}, \tag{E.1}$$

so that its amplitude is exponentially small everywhere except for a small neighborhood of the wedge face $\theta = \alpha$. Thus, we describe a Rayleigh wave incident from infinity along the upper face $\theta = \alpha$ by two potentials

$$\psi_i^{\mathrm{inc}}(kr,\theta) = 4\pi\mathrm{i}\psi_{i0}\mathrm{e}^{\mathrm{i}\gamma kr\cos(\theta-\theta_{i\mathrm{R}}^{\mathrm{inc}})}, \tag{E.2}$$

where $\psi_{00} = 1$ and $\psi_{10} = -2\mathrm{i}\gamma_{\mathrm{R}}\sqrt{\gamma_{\mathrm{R}}^2-\gamma^2}/(2\gamma_{\mathrm{R}}^2-1)$; $\gamma_{\mathrm{R}} = c_{\mathrm{S}}/c_{\mathrm{R}}$ with c_{S} being the Rayleigh wave speed and $\theta_{0\mathrm{R}}^{\mathrm{inc}} = \alpha - g(\mathrm{i}\beta_{\mathrm{R}})$. The reflected wave propagates along the same wedge face as the incident but from the tip to infinity, and the transmitted propagates along the other face, again away from the tip. They are described respectively by

$$\begin{aligned}\psi_i^{\mathrm{ref}}(kr,\theta) &= 4\pi\mathrm{i}R^{\mathrm{ref}}\psi_{i0}\mathrm{e}^{-\mathrm{i}\gamma kr\cos(\theta-\theta_{i\mathrm{R}}^{\mathrm{sc}})},\\ \psi_i^{\mathrm{tran}}(kr,\theta) &= 4\pi\mathrm{i}R^{\mathrm{tran}}\psi_{i0}\mathrm{e}^{-\mathrm{i}\gamma kr\cos(\theta+\theta_{i\mathrm{R}}^{\mathrm{sc}})},\end{aligned} \tag{E.3}$$

with "scattering angles" $\theta_{i\mathrm{R}}^{\mathrm{sc}}$ being the complex conjugates of $\theta_{i\mathrm{R}}^{\mathrm{inc}}$, so that $\theta_{0\mathrm{R}}^{\mathrm{sc}} = \alpha + g(\mathrm{i}\beta_{\mathrm{R}})$ and $\theta_{1\mathrm{R}}^{\mathrm{sc}} = \alpha + \mathrm{i}\beta_{\mathrm{R}}$. Above, R^{ref} and R^{tran} are the Rayleigh reflection and transmission coefficients

$$R^{\mathrm{ref}} = \frac{1}{2}[R^{+\mathrm{ref}} + R^{-\mathrm{ref}}], \qquad R^{\mathrm{tran}} = \frac{1}{2}[R^{+\mathrm{ref}} - R^{-\mathrm{ref}}], \tag{E.4}$$

with the symmetric and antisymmetric parts given respectively by

$$R^{\pm\mathrm{ref}} = \mathrm{Res}[\Psi_0^{\pm};\, g(\omega_{\mathrm{R}}) + \alpha], \;\text{ with } \omega_{\mathrm{R}} = \pi + \mathrm{i}\beta_{\mathrm{R}}, \tag{E.5}$$

so that, using the additional angles $\theta_{0,\mathrm{R}} = -\alpha + g(\omega_{\mathrm{R}})$ and $\theta_{1,\mathrm{R}} = -\alpha + \omega_{\mathrm{R}}$, we have

$$R^{\pm\mathrm{ref}} = \pm\sum_{k=1}^{2}\mathrm{r}_{1k}(\omega_{\mathrm{R}})\Psi_{k-1}^{\pm}(\theta_{k-1,\mathrm{R}}) + c_1^{\pm}e_1^{\pm}(\omega_{\mathrm{R}})\frac{g'(\omega_{\mathrm{R}})\Delta(\omega_{\mathrm{R}})}{\Delta'(\omega_{\mathrm{R}})}\sqrt{\gamma^2-\cos^2\omega_{\mathrm{R}}}. \tag{E.6}$$

As before, the dash denotes the derivative with respect to the argument.

Appendix F. Adjustable functions and parameters in singular terms. As with any other numerical code, ours relies on a choice of certain options which effect a tradeoff between numerical accuracy and either running time or else numerical stability. Apart from the relevant grids, these options are the following

(i) *The adjustable function $\sigma(\omega)$ in* (3.15). In most cases, $\sigma(\omega)$ is chosen to be

$$\sigma(\omega) = \frac{\frac{\pi}{2\alpha}}{\sin\frac{\pi}{2\alpha}\omega} \tag{F.1}$$

(cf. [5, (15)]). The choice is convenient, because it simplifies the right-hand sides of our functional, and therefore integral, equations. Also, $\sigma(\omega)$ in (F.1) decays at infinity reasonably fast. However, for any wedge angle 2α, there

exists a critical incident angle θ_0^{inc} such that one of the poles of $\Psi_i^{\pm}(\omega)$ lies on the boundary of the strip Re $\omega \epsilon I$. For illustration purposes, let it be $\widehat{\Psi}_1^{\pm}(\omega)$, and let the pole be $\theta_0 = \pi/2 - \alpha$. Then the corresponding term Res $(\Psi_1^{\pm}; \theta_0)\sigma(\omega - \theta_0)$ has one more nonphysical pole, $\theta_0 + 2\alpha$. In situations like these, another choice of $\sigma(\omega)$ is called for, with poles further apart. We have tested

$$\sigma(\omega) = \frac{\beta}{\sin \beta\omega}, \tag{F.2}$$

with various values $\beta < \pi/(2\alpha)$. However, any $\sigma(\omega)$ different from (F.1) leads to more cumbersome right-hand sides of the integral equations and exhibits a slower decay. As a result, the function (F.2), while increasing the stability of the solution, increases the code run time roughly tenfold. For this reason, we abandon (F.1) only when θ^{inc} is near critical angle. In this region we use (F.2), with $\beta = \pi/(6\alpha)$.

(ii) *Number of poles.* When evaluating the poles of the Sommerfeld amplitudes we do not have to restrict ourselves to the strip Re $\omega \epsilon I$. The more poles that are utilized in evaluation, the wider the domain of analyticity of the corresponding unknowns $\widetilde{\Psi}_i(\omega)$, and therefore the higher the accuracy. On the other hand, some nonphysical poles possess residues with large amplitudes and cause numerical instability. In the present version of the code, when θ is away from the critical angle we take into account all poles in the strip Re $\omega \epsilon I$, and when θ is near the critical angle we take into account all poles in the wider strip Re $\omega \epsilon [\pi/2 - 2\alpha, \pi/2 + 2\alpha]$.

Acknowledgments. We are grateful to Drs. R.K. Chapman, D. Gridin, and J. Hudson for many useful and insightful comments, and to Profs. V. P. Smyshlyaev and A. Gautesen for fruitful discussions and suggestions. We would also like to thank Profs. S. A. Nazarov and B. A. Plamenevskij for invaluable advice and relevant references.

REFERENCES

[1] K. E. Atkinson, *The Numerical Solutions of Integral Equations of the Second Kind*, Cambridge University Press, Cambridge, UK, 1997.

[2] V. M Babich, V. A. Borovikov, L. Ju. Fradkin, D. Gridin, V. Kamotski, and V. P. Smyshlyaev, *Diffraction coefficients for surface breaking cracks*, in Proceedings of the IUTAM Symposium, Manchester, UK, 2000, I. D. Abrahams, P. A. Martin, and M. J. Simon, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 209–216.

[3] V. M Babich, V. A. Borovikov, L. Ju. Fradkin, V. Kamotski, and B. A. Samokish, *Ultrasonic modelling of tilted surface breaking cracks*, Non-Destructive Testing and Evaluation Internationl, 37 (2003), pp. 105–110.

[4] B. V. Budaev, *Diffraction by Wedges*, Longman Scientific and Technical, Harlow, UK, 1995.

[5] B. V. Budaev and D. B. Bogy, *Rayleigh wave scattering by a wedge*, Wave Motion, 22 (1995), pp. 239–257.

[6] B. V. Budaev and D. B. Bogy, *Rayleigh wave scattering by a wedge* II, Wave Motion, 24 (1996), pp. 307–314.

[7] B. V. Budaev and D. B. Bogy, *Rayleigh wave scattering by two adhering elastic wedges*, Proc. Roy. Soc. London A, 454 (1998), pp. 2949–2996.

[8] B. V. Budaev and D. B. Bogy, *Scattering of Rayleigh and Stoneley waves by two adhering elastic wedges*, Wave Motion, 33 (2001), pp. 321–337.

[9] B. V. Budaev and D. B. Bogy, *Errata to the paper "Rayleigh wave scattering by a wedge* II," Wave Motion, 35 (2002), p. 275.

[10] J.-P. Croisille and G. Lebeau, *Diffraction by an Immersed Elastic Wedge*, Lecture Notes in Math. 1723, Springer-Verlag, Berlin, 1999.

[11] K. Fujii, *Rayleigh-wave scattering of various wedge corners: Investigation in the wider range of wedge angles*, Bull. Seismol. Soc. Am., 84 (1994), pp. 1916–1924.

[12] A. K. Gautesen, *Scattering of Rayleigh wave by an elastic quarter space*, J. Appl. Mech., 52 (1985), pp. 664–668.

[13] A. K. Gautesen, *Scattering of a Rayleigh wave by an elastic quarter space—Revisited*, Wave Motion, 35 (2002), pp. 91–98.

[14] A. K. Gautesen, *Scattering of a Rayleigh wave by an elastic wedge whose angle is less then* 180°, Wave Motion, 36 (2002), pp. 417–424.

[15] V. V. Kamotski, L. Ju. Fradkin, B. A. Samokish, V. A. Borovikov, and V. M. Babich, *On Budaev and Bogy's Approach to Diffraction by the* 2*D Traction-free Elastic Wedge*, preprint, 2006; available online at http://arxiv.org/abs/math-ph/0604011.

[16] V. V. Kamotski, *On the incidence of the plane wave on an elastic wedge at a critical angle*, Algebra and Analysis, 15 (2003), pp. 145–169. (English translation of the Russian original in St. Petersburg Math. J., 15 (2003), pp. 419–436.)

[17] V. V. Kamotski and G. Lebeau, *Diffraction by an elastic wedge with stress free boundary: Existence and uniqueness*, Proc. Roy. Soc. A, 462 (2006), pp. 289–317.

[18] V. A. Kondrat'ev, *Boundary value problems for elliptic equations in domains with conical and angular points*, Trudy Moskovskogo Matematicheskogo Obschestva, 16 (1963), pp. 219–292.

[19] V. G. Kozlov, V. A. Maz'ya, and J. Rossmann, *Spectral Problems Associated with Corner Singularities of Solutions to Elliptic Equations*, Math. Surveys Monogr. 85, American Mathematical Society, Providence, RI, 2001.

[20] G. D. Malyuzhinets, *Radiation of sound by oscillating faces of an arbitrary wedge* II, Acoust. J., 1 (1955), pp. 240–248 (in Russian).

[21] S. A. Nazarov and B. A. Plamenevskij, *Elliptic Problems in Domains with Piecewise Smooth Boundaries*, Walter de Gruyter, Berlin, 1994.

[22] T. C. T. Ting, *The wedge subjected to tractions: A paradox re-examined.*, J. Elasticity, 14 (1984), pp. 235–247.

SIAM J. APPL. MATH.
Vol. 67, No. 1, pp. 260–278

GLOBAL ANALYSIS OF NEW MALARIA INTRAHOST MODELS WITH A COMPETITIVE EXCLUSION PRINCIPLE*

ABDERRHAMAN IGGIDR†, JEAN-CLAUDE KAMGANG‡, GAUTHIER SALLET†, AND JEAN-JULES TEWA§

Abstract. In this paper we propose a malaria within-host model with k classes of age for the parasitized red blood cells and n strains for the parasite. We provide a global analysis for this model. A competitive exclusion principle holds. If $\mathcal{R}_0$, the basic reproduction number, satisfies $\mathcal{R}_0 \leq 1$, then the disease-free equilibrium is globally asymptotically stable. On the contrary if $\mathcal{R}_0 > 1$, then generically there is a unique endemic equilibrium which corresponds to the endemic stabilization of the most virulent parasite strain and to the extinction of all the other parasites strains. We prove that this equilibrium is globally asymptotically stable on the positive orthant if a mild sufficient condition is satisfied.

Key words. nonlinear dynamical systems, intrahost models, global stability, *Plasmodium falciparum*, competitive exclusion principle

AMS subject classifications. 34A34, 34D23, 34D40, 92D30

DOI. 10.1137/050643271

1. Introduction.

In this paper we consider intrahost models for malaria. These models describe the interaction of a parasite, namely a protozoa *Plasmodium falciparum*, with its target cells, the red blood cells (RBC). During the past decade there has been considerable work on the mathematical modeling of *Plasmodium falciparum* infection [2, 14, 21, 22, 24, 23, 25, 28, 30, 52, 55, 56, 58, 64]. A review has been done by Molineaux and Dietz in [59].

We give a brief review of the features of malaria. Malaria in a human begins with an inoculum of *Plasmodium* parasites (sporozoites) from a female *Anopheles* mosquito. The sporozoites enter the liver within minutes. After a period of asexual reproduction in the liver the parasites (merozoites) are released in the bloodstream where the asexual erythrocyte cycle begins. The merozoites enter RBC, grow, and reproduce over a period of approximately 48 hours after which the erythrocyte ruptures releasing 8–32 "merozoites" daughter parasites that quickly invade a fresh erythrocyte to renew the cycle. This blood cycle can be repeated many times, in the course of which some of the merozoites instead develop in the sexual form of the parasites: gametocytes. Gametocytes are benign for the host and are waiting for the mosquitoes.

The first mathematical model of the erythrocyte cycle was proposed by Anderson, May, and Gupta [3]. This original model has been extended in different directions [2, 3, 21, 25, 28, 30, 64].

The original model [3] is given by the following system:

$$\begin{cases} \dot{x} = \Lambda - \mu_x x - \beta x\, m, \\ \dot{y} = \beta x\, m - \mu_y\, y, \\ \dot{m} = r\, \mu_y\, y - \mu_m\, m - \beta\, x\, m. \end{cases} \tag{1.1}$$

*Received by the editors October 21, 2005; accepted for publication (in revised form) July 11, 2006; published electronically December 11, 2006.

http://www.siam.org/journals/siap/67-1/64327.html

†INRIA-Lorraine and Laboratoire de Mathématiques et Applications de Metz UMR CNRS 7122, University of Metz, 57045 Metz Cedex 01, France (iggidr@math.univ-metz.fr, sallet@loria.fr).

‡Department of Mathematics, ENSAI, University of Ngaoundéré, P.O. Box 455, Ngaoundéré, Cameroon (kamgang@loria.fr).

§Department of Mathematics, University of Yaoundé I, Yaoundé, Cameroon (tewajules@yahoo.fr).

The state variables are denoted by x, y, and m. The variable x denotes the concentration of uninfected RBC, y the concentration of parasitized red blood cells (PRBC), and m the concentration of the free merozoites in the blood.

We briefly sketch the interpretation of the parameters. Parameters μ_x, μ_y, and μ_m are the death rates of the RBC, PRBC, and free merozoites, respectively. The parameter β is the contact rate between RBC and merozoites. Uninfected blood cells are recruited at a constant rate Λ from the bone marrow and have a natural life-expectancy of $\frac{1}{\mu_x}$ days. Death of a PRBC results in the release of an average number of r merozoites. Free merozoites die or successfully invade a RBC.

This system is isomorphic to numerous systems considered in the mathematical modeling of virus dynamics; see [60, 61, 62] and the references therein. Some authors ignore the loss term $-\beta\, x\, m$ that should appear in the m equation. Indeed without this loss term, merozoites can infect RBC without themselves being absorbed, and this allows one merozoite *to infect more than one RBC*.

The original and the derived malaria models were intended to explain observations, namely parasitaemia, i.e., the concentration y of PRBC and also the decrease of the healthy RBC leading to anaemia. An important characteristic of *Plasmodium falciparum*, the most virulent malaria parasite, is sequestration. At the halfway point of parasite development, the infected erythrocyte leaves the circulating peripheral blood and binds to the endothelium in the microvasculature of various organs where the cycle is completed. A measurement of *Plasmodium falciparum* parasitaemia taken from a blood smear therefore samples young parasites only. Physician treating malaria use the number of parasites in peripheral blood smears as a measure of infection, and this does not give the total parasite burden of the patient. In some respects this is a weak point of the model (1.1). Moreover antimalarial drugs are known to act preferentially on different stages of parasite development. These facts lead some authors to give a general approach to modeling the age structure of *Plasmodium* parasites [22, 23, 24, 57]. Their model is a linear catenary compartmental model. This model is based on a finite number of compartments, each representing a stage of development of the parasite inside the PRBC. The models describe only the dynamics of the morphological stage evolution of the parasites and make no allowance for the dynamics of the healthy RBC.

In this paper we propose a model which combines the advantages of the two approaches. We also consider this model with different strains for the parasites. To encompass the different models of the literature we allow, in this model, to ignore or not the loss term in the m equation. To begin we consider the model with one strain:

$$\left\{\begin{array}{l} \dot{x} = f(x) - \mu_x x - \beta x\, m, \\ \dot{y}_1 = \beta x\, m - \alpha_1\, y_1, \\ \dot{y}_2 = \gamma_1\, y_1 - \alpha_2\, y_2, \\ \cdots \\ \dot{y}_k = \gamma_{k-1}\, y_{k-1} - \alpha_k\, y_k, \\ \dot{m} = r\, \gamma_k\, y_k - \mu_m\, m - u\, \beta\, x\, m. \end{array}\right. \tag{1.2}$$

In this system $f(x) - \mu_x\, x$ is the density-dependent growth rate of RBC. The other parameters are positive. In the model of Gravenor et al. [21] $\alpha_i = \gamma_i + \mu_i$, and hence $\alpha_i > \gamma_i$. We do not need this requirement, which implies that our model is not necessarily a catenary compartmental model. In the literature the parameter u takes the values $u = 0$ when the loss of the merozoite when it enters a RBC is ignored or takes $u = 1$ when this loss is not ignored. In our analysis u is simply a nonnegative

parameter. Except for these generalizations this system has already been suggested by Gravenor and Lloyd [21] in their reply to the criticism of Saul [64]. We provide a global analysis of this system related to the basic reproduction ratio $\mathcal{R}_0$ of the considered model.

One problem is how to decide upon the number of parasite compartments in the model. A starting point can be the morphological appearance of the parasite. But if the objective is to reflect the distribution of cycle lengths, the number of compartment can be increased to obtain a gamma distribution. Finally the two approaches can be combined: some compartments are for morphological reasons and others are for behavioral reasons. Then this model can also be interpreted as the application of the method of stages (or the linear chain trick) to the life cycle of PRBC [3, 31, 47, 49, 48, 51]. In other words a chain of compartments is included to generate a distribution of lags. It is also possible to add a class y_{k+1} in order to allow for the production of gametocytes. Different numbers of stages, ranging from 5 to 48, are used in [20, 22, 23, 24].

It is well grounded that a *falciparum* infection consists of distinct parasite genotypes. The model of Anderson, May, and Gupta has been extended in this direction [25, 66]. With regard to such features we propose a model with k stages for the infected RBC, production of gametocytes, and n genotypes, in the population of parasites.

One of the important principles of theoretical ecology is the competitive exclusion principle which states that no two species can indefinitely occupy the same ecological niche [7, 8, 11, 17, 25, 39, 53, 54]. We provide a global analysis of this model and obtain a generic competitive exclusion result within one host individual. This confirms the simulation results obtained in [25]. We compute the basic reproduction ratio $\mathcal{R}_0$ of the model. For this model there is always a disease-free equilibrium (DFE). To put it more precisely this equilibrium corresponds to the extinction of all the parasites, including the free parasites and the intraerythrocyte parasites. We prove that if $\mathcal{R}_0 \leq 1$, then the DFE is globally asymptotically stable (GAS); in other words the parasites are cleared. If $\mathcal{R}_0 > 1$, then, generically, a unique endemic equilibrium exists corresponding to the extinction of all the strains of parasites but one. We prove that this equilibrium is GAS on the positive orthant under a mild condition. For example this condition is automatically satisfied when $u = 0$ and $f(x) = \Lambda - \mu_x\, x$. When $u \neq 0$ the criteria, obtained for deciding the winning strain, differs from other results in the literature. To each i-strain can be associated a basic reproduction number $\mathcal{R}_0^i$ and a threshold $\mathcal{T}_0^i$. It turns out, when $u \neq 0$, that this is precisely this threshold $\mathcal{T}_0^i$ which distinguishes the fate of the strain and not $\mathcal{R}_0^i$ at the difference of [7, 11].

The paper is organized as follows. In section 2 we introduce the model with k stages for the infected RBC and one parasite strain, with and without gametocyte production. We compute the basic reproduction number and provide a stability analysis.

In section 3 we consider the model of Anderson, May, and Gupta with n distinct genotypes and production of gametocytes. This model with a constant recruitment function for the erythrocytes, two strains, and one class of age has been proposed in [25]. We have studied this model in [1]. Here using the computation of section 2, we prove for the general n strain k class of age model that if $\mathcal{R}_0 \leq 1$, then the parasites are cleared and if $\mathcal{R}_0 > 1$, then generically the different genotypes cannot coexist. Namely a unique equilibrium exists, for which only one genotype is positive, and which is GAS on a dense subset of the nonnegative orthant. This result confirms the simulations given in [25].

Global results of stability for the DFE as well for the endemic equilibrium for epidemic models are not so common [26, 27, 33, 43, 65, 67, 68]. Global stability

results for the endemic equilibrium have often been obtained by using monotone system techniques [29, 36]. Usually the Poincaré–Bendixson property of monotone systems in dimension 3 is used [40, 41, 42, 43, 44, 45]. Our results generalize the results of [13].

2. Stability analysis of a one strain model with *k* stages. We consider a general class of systems. The haemopoiesis is a complex system. In the cited references the recruitment of RBC is given by $\Lambda - \mu_x\, x$. In this paper we will use a more general function $\varphi(x)$. In a more complex system the haemopoiesis could be an input coming from another system:

$$\left\{\begin{array}{l} \dot{x} = f(x) - \mu_x x - \beta x\, m = \varphi(x) - \beta\, x\, m, \\ \dot{y}_1 = \beta x\, m - \alpha_1\, y_1, \\ \dot{y}_2 = \gamma_1\, y_1 - \alpha_2\, y_2, \\ \dots \\ \dot{y}_k = \gamma_{k-1}\, y_{k-1} - \alpha_k\, y_k, \\ \dot{m} = r\, \gamma_k\, y_k - \mu_m\, m - u\, \beta\, x\, m. \end{array}\right. \tag{2.1}$$

We denote by y the column vector $(y_1, \dots, y_k)^T$. The parameter u is nonnegative. The reason for this parameter is to encompass some malaria models in which the term $-\beta\, x\, m$ can appear or not. In [2] Anderson has considered a system without the $-\beta\, x\, m$ in the $\dot{m}$ equation. In [60] all the basic models of virus dynamics are also without this term. One feature of *Plasmodium falciparum*, responsible for the deadly case of malaria, is that more than one parasite can invade RBC. In this case u is the mean number of parasites invading RBC and thus disappearing from the circulating blood.

Some authors [25, 56] have included in the model production of gametocytes. In the course of the production of merozoites from bursting erythrocytes, some invading merozoites develop into the sexual, nonreplicating transmission stages known as gametocytes. The gametocytes are benign and transmissible to mosquitoes. We can also, following these authors, include a production of gametocytes in our model. If we denote by y_{k+1} the "concentration of gametocytes," the model becomes

$$\left\{\begin{array}{l} \dot{x} = f(x) - \mu_x x - \beta x\, m = \varphi(x) - \beta\, x\, m, \\ \dot{y}_1 = \beta x\, m - \alpha_1\, y_1, \\ \dot{y}_2 = \gamma_1\, y_1 - \alpha_2\, y_2, \\ \dots \\ \dot{y}_k = \gamma_{k-1}\, y_{k-1} - \alpha_k\, y_k, \\ \dot{y}_{k+1} = \rho\, \gamma_k\, y_k - \alpha_{k+1}\, y_{k+1}, \\ \dot{m} = r\, \gamma_k\, y_k - \mu_m\, m - u\, \beta\, x\, m. \end{array}\right. \tag{2.2}$$

We start to analyze the system with minimal hypothesis on f but nevertheless plausible from the biological point of view. The function f gives the production of erythrocytes from the bone marrow. The function $\varphi(x) = f(x) - \mu_x x$ models the population dynamic of RBC in the absence of parasites. The RBC have a finite lifetime, and then μ_x represents the average per capita death rate of RBC. The function f models in some way homeostasis. In this paper we suppose that f depends only on x. It could be assumed that the recruitment function depends on x and the total population of erythrocytes $x + \sum_i y_i$. In this paper we will analyze the simplified case which is the model considered in all the referenced literature. The rationale behind this simplification is that in a malaria primo-infection typically y is in the order of 10^{-1}

to 10^{-4} of the concentration of healthy erythrocytes x. This can be confirmed from the data of malaria therapy. In the last century neurosyphilitic patients were given malaria therapy, which was routine care at that time. Some of them were infected with *Plasmodium falciparum*. Data were collected at the National Institutes of Health laboratories in Columbia, SC and Milledgeville, GA during the period 1940 to 1963 [12].

We assume that f is a $\mathcal{C}^1$. Since homeostasis is maintained we assume that the dynamic without parasites is asymptotically stable. In other words, for the system

$$\dot{x} = f(x) - \mu_x\, x = \varphi(x)$$

there exists a unique $x^* > 0$ such that

$$\varphi(x^*) = 0, \quad \text{and} \ \ \varphi(x) > 0 \quad \text{for} \ \ 0 \le x < x^*, \ \ \text{and} \ \ \varphi(x) < 0 \ \ \text{for} \ x > x^*. \tag{2.3}$$

2.1. Notation. We will rewrite systems (2.1) and (2.2) in a condensed simpler form.

Before we introduce some classical notation.

We identify vectors of $\mathbb{R}^n$ with $n \times 1$ column vectors. $\langle\, |\, \rangle$ denotes the euclidean inner product. $\|z\|_2^2 = \langle\, z \mid z\, \rangle$ is the usual euclidean norm.

The family $\{e_1, \ldots, e_n\}$ denotes the canonical basis of the vector space $\mathbb{R}^n$. For example $e_1 = (1, 0, \ldots, 0)^T$. We denote by e_ω the last vector of the canonical basis, $e_\omega = (0, \ldots, 0, 1)^T$.

If $z \in \mathbb{R}^n$, we denote by z_i the ith component of z. Equivalently $z_i = \langle\, z \mid e_i\, \rangle$.

For a matrix A we denote by $A(i, j)$ the entry at the row i, column j. For matrices A, B we write $A \le B$ if $A(i, j) \le B(i, j)$ for all i and j, $A < B$ if $A \le B$ and $A \ne B$, and $A \ll B$ if $A(i, j) < B(i, j)$ for all i and j.

A^T denotes the transpose of A. Then $\langle\, z_1 \mid z_2\, \rangle = z_1^T\, z_2$. The notation A^{-T} will denote the transpose of the inverse of A.

For this section we rewrite the systems (2.1) and (2.2) under a unique form:

$$\begin{cases} \dot{x} = \varphi(x) - \beta\, x\, \langle\, e_\omega \mid z\, \rangle, \\ \dot{z} = \beta\, x\, \langle\, e_\omega \mid z\, \rangle\, e_1 + A_0\, z - u\, \beta\, x\, \langle\, e_\omega \mid z\, \rangle\, e_\omega. \end{cases} \tag{2.4}$$

In the case of the system (2.1) we have for A_0

$$A_0 = \begin{bmatrix} -\alpha_1 & 0 & 0 & \cdots & 0 & 0 \\ \gamma_1 & -\alpha_2 & 0 & \cdots & 0 & 0 \\ 0 & \gamma_2 & -\alpha_3 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \gamma_{k-1} & -\alpha_k & 0 \\ 0 & \cdots & 0 & 0 & r\gamma_k & -\mu_m \end{bmatrix} \tag{2.5}$$

and an analogous formula for (2.2).

We define the matrix $A(x) = A_0 - \beta\, x\, e_\omega\, e_\omega^T$. This a Metzler stable matrix. (A Metzler matrix is a matrix with nonnegative off-diagonal entries [5, 32, 50].)

It is not difficult to check that the nonnegative orthant is positively invariant by (2.4) and that there exists a compact absorbing set K for this system. An absorbing set D is a neighborhood such that a trajectory of the system starting from any initial condition enters and remains in D for a sufficiently large time T.

2.2. Global stability results. We can now give the main result of this section.

THEOREM 2.1. *We consider the system* (2.4) *with the hypothesis* (2.3) *on* φ *satisfied. We define the basic reproduction ratio of the system* (2.1) *and* (2.2) *by*

$$\mathcal{R}_0 = \frac{r\beta x^*}{\mu_m + u\,\beta\,x^*}\,\frac{\gamma_1\cdots\gamma_k}{\alpha_1\cdots\alpha_k}. \tag{2.6}$$

1. *The system* (2.1) *is GAS on* $\mathbb{R}^{k+2}_+$ *(respectively,* (2.2) *on* $\mathbb{R}^{k+3}_+$*) at the DFE* $(x^*, 0, \ldots, 0)$ *if and only if* $\mathcal{R}_0 \le 1$.

2. *If* $\mathcal{R}_0 > 1$*, then the DFE is unstable and there exists a unique endemic equilibrium (EE) in the positive orthant,* $(\bar{x}, \bar{z}) \gg 0$*, given by*

$$\begin{cases} \bar{x} = \dfrac{\mu_m}{\beta\left[r\dfrac{\gamma_1\cdots\gamma_k}{\alpha_1\cdots\alpha_k} - u\right]}, \\ \bar{z} = \varphi(\bar{x})\,(-A_0)^{-1}\,(e_1 - u\,e_\omega). \end{cases} \tag{2.7}$$

Denoting $\alpha^* = -\max_{x\in[0,x^*]}(\varphi'(x))$*, if*

$$u\,\beta\,\varphi(\bar{x}) \le \alpha^*\,\mu_m, \tag{2.8}$$

then the EE is GAS on the nonnegative orthant, except for initial conditions on the x-axis.

Proof of Theorem 2.1. To begin we will consider the system (2.1) without gametocytes, i.e., the system (2.4) with A_0 as defined in (2.5). The stability analysis for (2.2) follows easily from the stability analysis of (2.1).

In a first step we will compute $\mathcal{R}_0 \le 1$. We use our preceding notation and define $A^* = A(x^*)$, i.e., the matrix computed at the equilibrium x^* of φ, which is a stable Metzler matrix. We will use, repeatedly in what follows, the property that if M is a stable Metzler matrix, then $-M^{-1} \ge 0$ [5]. The expression of $\mathcal{R}_0$ is obtained easily by using the next generation matrix of the system (2.1) [9, 15, 16]. We have for the basic reproduction number

$$\mathcal{R}_0 = \beta\,x^*\left\langle -(A^*)^{-1}\,e_1 \mid e_\omega \right\rangle.$$

If we remark that the matrix A^* is the matrix A_0 modified by a rank-one matrix, namely $A^* = A_0 - u\,\beta\,x^*\,e_\omega\,e_\omega^T$, we can use the Sherman–Morrison–Woodbury formula

$$-(A^*)^{-1} = -A_0^{-1} - \frac{u\,\beta\,x^*}{1 + u\,\beta\,x^*\,e_\omega^T\,(-A_0)^{-1}\,e_\omega}\,(-A_0)^{-1}\,e_\omega\,e_\omega^T\,(-A_0)^{-1}$$

or equivalently

$$-(A^*)^{-1} = -A_0^{-1} - \frac{u\,\beta\,x^*}{\mu_m + \beta\,x^*}\,e_\omega\,e_\omega^T\,(-A_0)^{-1}.$$

This shows that $-(A^*)^{-1}$ is obtained from $-A_0^{-1}$ by multiplying the last line of $-A_0^{-1}$ by $\frac{\mu_m}{\mu_m + u\,\beta\,x^*}$. Then we get

$$\mathcal{R}_0 = \beta\,x^*\,\frac{\mu_m}{\mu_m + u\,\beta\,x^*}\left\langle -(A_0)^{-1}\,e_1 \mid e_\omega \right\rangle,$$

and then in computing the last entry of the first column of A_0 we obtain (2.6).

We remark that $\mathcal{R}_0 > 1$ is equivalent to the following threshold condition:

(2.9)
$$\mathcal{T}_0 = \frac{\beta\, x^*}{\mu_m}\left[\mu_m \left\langle -(A_0)^{-1}\, e_1 \mid e_\omega \right\rangle - u\right] = \beta\, x^* \left\langle -(A_0)^{-1}\, (e_1 - u\, e_\omega) \mid e_\omega \right\rangle > 1.$$

We are now ready to analyze the stability of the DFE.

It is well known that if $\mathcal{R}_0 > 1$, then the DFE is unstable [15], which implies that the condition $\mathcal{R}_0 \leq 1$ is necessary for stability.

To prove the sufficiency, in a second step, we consider the following function defined on the nonnegative orthant:

$$V_{DFE}(z) = \beta\, x^* \langle\, e_\omega \mid (-A_0^{-1})z \,\rangle. \tag{2.10}$$

Its time derivative along the trajectories of system (2.4) is

$$\dot{V}_{DFE} = \beta\, x \langle\, e_\omega \mid z \,\rangle\, \beta\, x^* \langle\, e_\omega \mid (-A_0)^{-1}\, (e_1 - u\, e_\omega) \,\rangle - \beta\, x^* \langle\, e_\omega \mid z \,\rangle$$

or equivalently, using the expression of $\mathcal{T}_0$ given in (2.9),

$$\dot{V}_{DFE} = \beta \langle\, e_\omega \mid z \,\rangle\, (\mathcal{T}_0\, x - x^*)\,. \tag{2.11}$$

Now we take as a candidate Liapunov function, defined on the nonnegative orthant minus the hyperplane face $x = 0$,

$$V = (x - x^* \ln x) - x^*(1 - \ln x^*) + V_{DFE}(z).$$

This function is positive definite (relatively to the DFE) on $\mathbb{R}^{k+2}_{+,x>0} = \{(x, y, m) \in \mathbb{R}^{k+2}_{+} \;:\; x > 0\}$. Its time derivative is given by

$$\dot{V} = \frac{x - x^*}{x}\, \varphi(x) - (x - x^*)\, \beta \langle\, e_\omega \mid z \,\rangle + \beta \langle\, e_\omega \mid z \,\rangle\, (\mathcal{T}_0\, x - x^*)$$

or assuming $\mathcal{R}_0 \leq 1$

$$\dot{V} = \frac{x - x^*}{x}\, \varphi(x) + \beta\, x \langle\, e_\omega \mid z \,\rangle\, (\mathcal{T}_0 - 1) \;\leq\; 0.$$

By assumption (2.3) we have $(x - x^*)\varphi(x) \leq 0$ for all $x \geq 0$. Therefore $\dot{V} \leq 0$ for all $(x, z) \in \mathbb{R}^{k+2}_{+,x>0}$, which proves the stability of the DFE. Its attractivity follows from LaSalle's invariance principle [6, 37, 38], since the largest invariant set contained in $\{(x, z) \in \mathbb{R}^{k+2}_{+,x>0} \;:\; \dot{V} = 0\}$ is reduced to the DFE. On the other hand the vector field is strictly entrant on the face $x = 0$. Hence the whole orthant $\mathbb{R}^{k+2}_{+}$ belongs to the region of attraction of the DFE.

Now we assume that $\mathcal{R}_0 > 1$. The equilibria $(\bar{x}, \bar{z})$ of the system, different from the DFE, are determined by the relations

$$\bar{z} = \beta\, \bar{x} \langle\, \bar{z} \mid e_\omega \,\rangle\, (-A_0)^{-1}\, (e_1 - u\, e_\omega).$$

Replacing $\bar{z}$ in $\langle\, \bar{z} \mid e_\omega \,\rangle$ we obtain

$$\langle\, \bar{z} \mid e_\omega \,\rangle = \beta\, \bar{x} \langle\, \bar{z} \mid e_\omega \,\rangle \langle\, (-A_0)^{-1}\, (e_1 - u\, e_\omega) \mid e_\omega \,\rangle. \tag{2.12}$$

If $\langle\, \bar{z} \mid e_\omega \,\rangle = 0$, then $\varphi(\bar{x}) = 0$, we obtain $\bar{x} = x^*$, and hence $\bar{z} = 0$; i.e., the corresponding equilibrium is the DFE. In the other case, i.e., $\langle\, \bar{z} \mid e_\omega \,\rangle \neq 0$, the relation (2.12) gives

$$\beta\, \bar{x} \left\langle\, (-A_0)^{-1}\, (e_1 - u\, e_\omega) \mid e_\omega \,\right\rangle = 1. \tag{2.13}$$

Using $\langle\, (-A_0)^{-1}\, e_\omega \mid e_\omega \,\rangle = \frac{1}{\mu_m}$ we finally have

$$\bar{x} = \frac{\mu_m}{\beta\, \left[\mu_m \,\langle\, (-A_0)^{-1}\, e_1 \mid e_\omega \,\rangle - u\right]} = \frac{x^*}{\mathcal{T}_0}.$$

We deduce that if $\mathcal{R}_0 > 1$, then $0 < \bar{x} < x^*$, and hence $\varphi(\bar{x}) > 0$. Therefore

$$\bar{z} = \varphi(\bar{x})\ (-A_0)^{-1}\, (e_1 - u\, e_\omega).$$

The last component of $\bar{z}$, $\langle\, \bar{z} \mid e_\omega \,\rangle = \bar{m}$, is given by

$$\bar{m} = \frac{\varphi(\bar{x})}{\beta\, \bar{x}} > 0.$$

The k first components of $\bar{z}$ are given by the k first components of $\varphi(\bar{x})\ (-A_0)^{-1}\, e_1$. It is straightforward to check that the first column of $(-A_0)^{-1}$ namely $(-A_0)^{-1}\, e_1 \gg 0$, which proves that $\bar{z} \gg 0$. We have then proved that there is a unique EE in the positive orthant if and only if $\mathcal{R}_0 > 1$.

Finally we will prove a sufficient condition for the global asymptotic stability of the EE. To this end we define the following candidate Liapunov function on the positive orthant minus the face corresponding to $x = 0$:

$$V_{EE}(x, y, m) = a(x - \bar{x} \ln x) + \sum_{i=1}^{k} b_i\, (y_i - \bar{y}_i \ln y_i)\ + b_{k+1}\, (m - \bar{m}\ \ln m). \tag{2.14}$$

This function has a unique global minimum in $(\bar{x}, \bar{y}, \bar{m})$. We will choose the coefficients a, b_i, b_{k+1} such that in the computation of $\dot{V}$, the linear terms in y_i and m and the bilinear terms in $x\, m$ cancel. Let us show that it is possible with positive coefficients. To this end we rewrite the function V_{EE} using the notation $z = (y, m)^T$, $\ln z = (\ln z_1, \ln z_2, \ldots, \ln z_{k+1})^T$, and $b = (b_1, \ldots, b_k, b_{k+1})^T$:

$$V_{EE}(x, z) = a(x - \bar{x} \ln x) + \langle\, b \mid z - diag(\bar{z}) \ln z \,\rangle.$$

Consider the block matrix

$$M = \begin{bmatrix} -1 & (e_1 - u\, e_\omega)^T \\ \beta\, \bar{x}\, e_\omega & A_0^T \end{bmatrix}.$$

Using classical Schur complement techniques and the relation (2.13) on $\bar{x}$, we have

$$\begin{aligned} \det(M) &= \det(A_0)[-1 + \beta\, \bar{x} (e_1 - u\, e_\omega)^T\, (-A_0^{-T})\, e_\omega] \\ &= \det(A_0)[-1 + \beta\, \bar{x}\, \langle\, -A_0^{-1}\, (e_1 - u\, e_\omega) \mid e_\omega \,\rangle] = 0. \end{aligned}$$

Since the matrix M is obviously of codimension 1 (A_0 is nonsingular) the kernel of M is of dimension 1. Then there exists $a \in \mathbb{R}$ and $b \in \mathbb{R}^{k+1}$ such that

$$a = (e_1 - u\, e_\omega)^T\, b = \langle\, b \mid e_1 - u\, e_\omega \,\rangle \tag{2.15a}$$

and

$$b = a\,\beta\,\bar{x}\,(-A_0^{-T})\,e_\omega. \tag{2.15b}$$

Since the kernel is one dimensional, a can be chosen arbitrarily. Thanks to the structure of A_0, if $a > 0$, then $b \gg 0$.

The derivative of V along the trajectories of (2.4) is given by

$$\begin{aligned}\dot{V}_{EE} &= a\,\frac{x-\bar{x}}{x}\,\varphi(x) - a\,\beta x\,\langle\, e_\omega \mid z\,\rangle + a\,\beta\,\bar{x}\,\langle\, e_\omega \mid z\,\rangle + \beta\,x\,\langle\, e_\omega \mid z\,\rangle\,\langle\, b \mid e_1 - u\,e_\omega\,\rangle \\ &\quad + \langle\, b \mid A_0 z\,\rangle + \langle\, b \mid \operatorname{diag}(\bar{z})\operatorname{diag}(z)^{-1}\,\dot{z}\,\rangle \\ &= a\,\frac{x-\bar{x}}{x}\,\varphi(x) + \langle\, b \mid \operatorname{diag}(\bar{z})\operatorname{diag}(z)^{-1}\,\dot{z}\,\rangle \\ &\quad + a\,\beta\,\bar{x}\,\langle\, e_\omega \mid z\,\rangle + \langle\, b \mid A_0 z\,\rangle + \beta x\,\langle\, e_\omega \mid z\,\rangle\,\big(\langle\, b \mid e_1 - u\,e_\omega\,\rangle - a\big).\end{aligned}$$

Using the relation (2.15b) we see that

$$\langle\, b \mid A_0\,z\,\rangle = -a\,\beta\,\bar{x}\,\langle\,(A_0^{-T})\,e_\omega \mid A_0 z\,\rangle = -a\,\beta\,\bar{x}\,\langle\, e_\omega \mid z\,\rangle.$$

Therefore the linear terms in z cancel. The same is true for the bilinear terms thanks to the relation (2.15a). Finally we get

$$\dot{V}_{EE} = a\,\frac{x-\bar{x}}{x}\,\varphi(x) + \langle\, b \mid \operatorname{diag}(\bar{z})\operatorname{diag}(z)^{-1}\,\dot{z}\,\rangle.$$

We choose $b_{k+1} = 1 = \langle\, b \mid e_\omega\,\rangle = a\,\beta\,\bar{x}\,\langle\, -A^{-T}\,e_\omega \mid e_\omega\,\rangle = a\,\beta\,\bar{x}\frac{1}{\mu_m}$. In other words $a = \frac{\mu_m}{\beta\,\bar{x}}$. With the hypothesis $\mathcal{R}_0 > 1$ we have $a > 0$, and hence $b \gg 0$ as wanted. With this choice developing $\dot{V}$ gives

$$\begin{aligned}\dot{V}_{EE} &= a\,f(x) - a\mu_x\,x - af(x)\frac{\bar{x}}{x} + a\,\mu_x\,\bar{x} - b_1\,\beta\bar{y}_1\,\frac{xm}{y_1} - \sum_{i=2}^{k} b_i\gamma_{i-1}y_{i-1}\frac{\bar{y}_i}{y_i} \\ &\quad + \sum_{i=1}^{k} b_i\,\alpha_i\,\bar{y}_i - r\,\gamma_k\,y_k\,\frac{\bar{m}}{m} + u\,\beta\bar{m}x + \mu_m\bar{m}.\end{aligned}$$

We collect some useful relations between our coefficients at the EE. We have from the definitions of a and b, since $b_{k+1} = 1$,

$$\begin{cases} a + u = b_1, \\ b_1\,\alpha_1 = \gamma_1\,b_2, \\ b_2\,\alpha_2 = \gamma_2\,b_3, \\ \cdots \\ b_{k-1}\,\alpha_{k-1} = \gamma_{k-1}\,b_k, \\ b_k\,\alpha_k = r\,\gamma_k. \end{cases} \tag{2.16}$$

From these relations and the properties of the EE $\bar{z}$ we have

$$b_1\,\beta\,\bar{x}\,\bar{m} = b_i\,\alpha_i\,\bar{y}_i = b_i\,\gamma_{i-1}\,\bar{y}_{i-1} = r\,\gamma_k\,\bar{y}_k \tag{2.17}$$

and

$$a\alpha_1\bar{y}_1 = \mu_m\,\bar{m}. \tag{2.18}$$

Replacing, in the expression of $\dot{V}$, $a\mu_x\bar{x}$ by $a\,f(\bar{x}) - a\beta\bar{x}\bar{m} = af(\bar{x}) - a\alpha_1\,\bar{y}_1$ we obtain

$$\dot{V}_{EE} = kr\gamma_k\bar{y}_k + af(\bar{x}) + af(x) + (u\,\beta\bar{x}\bar{m} - a\mu_x\bar{x})\frac{x}{\bar{x}} - af(x)\frac{\bar{x}}{x} - b_1\,\beta\,\bar{x}\,\bar{m}\,\frac{x}{\bar{x}}\,\frac{m}{\bar{m}}\,\frac{\bar{y}_1}{y_1} - \sum_{i=2}^{k} b_i\gamma_{i-1}\bar{y}_{i-1}\frac{y_{i-1}}{\bar{y}_{i-1}}\frac{\bar{y}_i}{y_i} - r\gamma_k\bar{y}_k\frac{y_k}{\bar{y}_k}\frac{\bar{m}}{m}.$$

Using again the relations between the coefficients we get

$$\dot{V}_{EE} = kr\gamma_k\bar{y}_k + af(\bar{x}) + af(x) + (r\gamma_k\bar{y}_k - af(\bar{x}))\frac{x}{\bar{x}} - af(x)\frac{\bar{x}}{x} - r\,\gamma_k\,y_k\,\frac{x}{\bar{x}}\,\frac{m}{\bar{m}}\,\frac{\bar{y}_1}{y_1} - \sum_{i=2}^{k} r\gamma_k\bar{y}_k\frac{y_{i-1}}{\bar{y}_{i-1}}\frac{\bar{y}_i}{y_i} - r\gamma_k\bar{y}_k\frac{y_k}{\bar{y}_k}\frac{\bar{m}}{m}$$

and finally

$$\dot{V}_{EE} = a\left[f(x) + f(\bar{x}) - f(\bar{x})\frac{x}{\bar{x}} - f(x)\frac{\bar{x}}{x}\right] + r\,\gamma_k\,\bar{y}_k\left[k + \frac{x}{\bar{x}} - \frac{x}{\bar{x}}\,\frac{m}{\bar{m}}\,\frac{\bar{y}_1}{y_1} - \sum_{i=2}^{k}\frac{y_{i-1}}{\bar{y}_{i-1}}\frac{\bar{y}_i}{y_i} - \frac{y_k}{\bar{y}_k}\frac{\bar{m}}{m}\right].$$

Now we will use the fact that there exists ξ in the open interval $\xi \in]x, \bar{x}[$ such that $f(x) = f(\bar{x}) + (x - \bar{x})\,f'(\xi)$. Replacing in the preceding expression gives

$$\dot{V}_{EE} = af(\bar{x})\left[2 - \frac{x}{\bar{x}} - \frac{\bar{x}}{x}\right] + a\,f'(\xi)\frac{(x-\bar{x})^2}{x} + r\,\gamma_k\,\bar{y}_k\left[k + \frac{x}{\bar{x}} - \frac{x}{\bar{x}}\,\frac{m}{\bar{m}}\,\frac{\bar{y}_1}{y_1} - \sum_{i=2}^{k}\frac{y_{i-1}}{\bar{y}_{i-1}}\frac{\bar{y}_i}{y_i} - \frac{y_k}{\bar{y}_k}\frac{\bar{m}}{m}\right].$$

Using the relations (2.16)–(2.17) we have

$$af(\bar{x}) = (b_1 - u)f(\bar{x}) = b_1(\mu_x\bar{x} + \beta\bar{x}\bar{m}) - u\,f(\bar{x}) = b_1\,\mu_x\,\bar{x} + r\gamma_k\bar{y}_k - u\,f(\bar{x}).$$

Replacing in the preceding expression of $\dot{V}$ gives

$$\dot{V}_{EE} = (b_1\,\mu_x\,\bar{x} - u\,f(\bar{x}))\,\left[2 - \frac{x}{\bar{x}} - \frac{\bar{x}}{x}\right] + a\,f'(\xi)\frac{(x-\bar{x})^2}{x} + r\,\gamma_k\,\bar{y}_k\left[k + 2 - \frac{\bar{x}}{x} - \frac{x}{\bar{x}}\,\frac{m}{\bar{m}}\,\frac{\bar{y}_1}{y_1} - \sum_{i=2}^{k}\frac{y_{i-1}}{\bar{y}_{i-1}}\frac{\bar{y}_i}{y_i} - \frac{y_k}{\bar{y}_k}\frac{\bar{m}}{m}\right].$$

This can also be written

$$\dot{V}_{EE} = \Phi(x, y, m) = -\left[b_1\,\mu_x\,\bar{x} - u\,f(\bar{x}) - a\,\bar{x}\,f'(\xi)\right]\frac{(x-\bar{x})^2}{x\bar{x}} + r\,\gamma_k\,\bar{y}_k\left[k + 2 - \frac{\bar{x}}{x} - \frac{x}{\bar{x}}\,\frac{m}{\bar{m}}\,\frac{\bar{y}_1}{y_1} - \sum_{i=2}^{k}\frac{y_{i-1}}{\bar{y}_{i-1}}\frac{\bar{y}_i}{y_i} - \frac{y_k}{\bar{y}_k}\frac{\bar{m}}{m}\right]. \tag{2.19}$$

The term between brackets in the last expression of $\dot{V}$ is nonpositive by the inequality between the arithmetical mean and the geometrical mean. Then a sufficient condition for $\dot{V} \leq 0$ is

$$b_1\,\mu_x\,\bar{x} - u\,f(\bar{x}) - a\,\bar{x}\,f'(\xi) \geq 0.$$

Moreover with this condition $\dot{V}$ is negative, except at the EE for the system (2.1). This proves the global asymptotic stability of the EE on the positive orthant for the system (2.1).

The vector field associated with the system is strictly entrant on the faces of the orthant, except the x-axis, where it is tangent. The basin of attraction of the EE is then the orthant, except the x-axis, which is the stable manifold of the DFE.

Using the function $\varphi(x) = f(x) - \mu_x\,x$ the preceding condition is equivalent to

$$u\,\varphi(\bar{x}) \leq -a\,\bar{x}\,\varphi'(\xi),$$

or equivalently, replacing a by its value $a = \frac{\mu_m}{\beta\,\bar{x}}$, the condition becomes

$$u\,\beta\,\varphi(\bar{x}) \leq -\mu_m\,\varphi'(\xi).$$

Setting $\alpha^* = -\max_{x\in[0,x^*]}\varphi'(x)$ a sufficient condition for global asymptotic stability of the EE is

$$\mathcal{R}_0 > 1 \quad \text{and} \quad u\,\beta\,\varphi(\bar{x}) \leq \mu_m\alpha^*.$$

We have proved the theorem for the system without gametocytes. We have seen that $\mathcal{R}_0$ does not depend on the production of gametocytes. If $\mathcal{R}_0 \leq 1$, it is easy, integrating the linear stable y_{k+1} equations of (2.2) from the solutions of (2.1), to see that the DFE is asymptotically stable and that all the trajectories converge to the equilibrium. The same argument is used when $\mathcal{R}_0 > 1$. This ends the proof of Theorem 2.1. □

Remark 1. If this model is a model for a within-host model of malaria, each coefficient α_i is made of the mortality of the i-class and the rate of transmission in the $i+1$-class: $\alpha_i = \mu_i + \gamma_i$. This implies that $\gamma_i \leq \alpha_i$. We do not need this assumption, and our conclusions are valid for our more general model. The only hypothesis is that the parameters of the system are positive.

Remark 2. In the proof of Theorem 2.1 the quantity

$$\beta\,x^*\left\langle -(A_0)^{-1}\,(e_1 - u\,e_\omega) \mid e_\omega \right\rangle,$$

which we have called $\mathcal{T}_0$ when $\mathcal{R}_0 > 1$, plays a prominent role. When $\mathcal{R}_0 \leq 1$ and $u \neq 0$ three cases occur: $0 < \mathcal{T}_0 \leq 1$ or $\mathcal{T}_0 < 0$ or $\mathcal{T}_0 = 0$.

In the two first cases we can define $\bar{x} = \frac{x^*}{\mathcal{T}_0}$, and we obtain an equilibrium $(\bar{x}, \bar{z})$ of the system which is not in the nonnegative orthant (either $\bar{x} < 0$ or $\bar{z} < 0$).

In the third case, the computations, done in the proof of Theorem 2.1, for the research of an equilibrium show that $\langle\, z \mid e_\omega \,\rangle = 0$, and hence $z = 0$, and finally the equilibrium is the DFE $(x^*, 0)$.

We introduce a definition of $\mathcal{T}_0$ that will simplify future computations. The case $\mathcal{T}_0 = 0$ is special, since $\mathcal{T}_0 = \frac{x^*}{\bar{x}}$ is no longer true. However this case can be thought, by convention and misuse of language, as $\bar{x} = +\infty$.

DEFINITION 2.2. *We define for the system* (2.1) *the threshold*

$$\mathcal{T}_0 = \frac{x^*}{\dfrac{\mu_m}{\beta\left[r\dfrac{\gamma_1\cdots\gamma_k}{\alpha_1\cdots\alpha_k} - u\right]}} = \beta\, x^* \left\langle -(A_0)^{-1}\,(e_1 - u\, e_\omega) \mid e_\omega \right\rangle. \tag{2.20}$$

When $\mathcal{T}_0 \neq 0$ *we have also* $\mathcal{T}_0 = \frac{x^*}{\bar{x}}$.

Remark 3. It should be pointed out that the kind of Liapunov function defined by (2.14) has a long history of application to Lotka–Volterra models [18, 19] and was originally discovered by Volterra himself, although he did not use the vocabulary and the theory of Liapunov functions. Since epidemic models are "Lotka–Volterra" like models, the pertinence of this function is not surprising. Similar Liapunov functions have been used in epidemiology [4, 34, 35, 46, 63], although with different parameters. We have already used this kind of function in a simplified version of this paper in [1].

2.3. Comparison with known results. Our stability result improves the one of De Leenheer and Smith [13] in two directions:

1. We introduce n stages for latent classes.
2. Our sufficient condition for the global asymptotic stability of the endemic equilibrium is weaker than the one provided in [13]; for instance the sufficient condition given in Theorem 2.1 is satisfied for malaria parameters given in [3], while the condition of [13] is not satisfied.

2.4. Application to the original AMG model [3]. The original Anderson–May–Guptka model is a three dimensional system (1.1) which has the same form as system (2.1) with $f(x) = \Lambda$. The sufficient condition (2.8) applied to the AMG model (1.1) can be written

$$\beta\Lambda \leq \frac{r}{r-1}\,\mu_x\,\mu_m. \tag{2.21}$$

For the system (1.1), it is possible to give a weaker sufficient stability condition.

PROPOSITION 2.3. *If* $\mathcal{R}_0 > 1$ *and* $\beta\Lambda \leq (\sqrt{r} + \sqrt{r-1})^2\,\mu_x\,\mu_m$, *then the EE is a GAS steady state for system* (1.1) *with respect to initial states not on the* x*-axis.*

Since in general the parameter r is larger than 2 (see, for instance, [28]), we have $(\sqrt{r} + \sqrt{r-1})^2 > \frac{r}{r-1}$.

Proof. Thanks to the computations done before, we have for system (1.1)

$$\dot{V}_{EE} = (r-1)\Lambda\left[2 - \frac{x}{\bar{x}} - \frac{\bar{x}}{x}\right] + r\,\mu_y\,\bar{y}\left[1 + \frac{x}{\bar{x}} - \frac{y}{\bar{y}}\frac{\bar{m}}{m} - \frac{x}{\bar{x}}\frac{m}{\bar{m}}\frac{\bar{y}}{y}\right].$$

Define $X = \frac{x}{\bar{x}}$ and $S = \frac{y}{\bar{y}}\frac{\bar{m}}{m}$. Then one can write

$$\begin{aligned}\dot{V}_{EE} &= -(r-1)\Lambda\tfrac{(X-1)^2}{X} + r\,\mu_y\,\bar{y}\left(1 + X - S - \tfrac{X}{S}\right)\\ &= -(r-1)\Lambda\tfrac{(X-1)^2}{X} + r\,\mu_y\,\bar{y}\;\Psi(X,S).\end{aligned}$$

We have $\Psi(X,S) \geq 0 \Leftrightarrow X \leq S \leq 1$ or $X \geq S \geq 1$. On the other hand $\Psi(X,S) \leq \Psi(X,\sqrt{X}) = (\sqrt{X}-1)^2$. Therefore

$$\begin{aligned}&\dot{V}_{EE} \leq (r-1)\Lambda(\sqrt{X}-1)^2\left(\tfrac{r\,\mu_y\,\bar{y}}{(r-1)\Lambda} - \left(1 + \tfrac{1}{\sqrt{X}}\right)^2\right),\\ &\dot{V}_{EE} \leq (r-1)\Lambda(\sqrt{X}-1)^2\left(\sqrt{\tfrac{r\,\mu_y\,\bar{y}}{(r-1)\Lambda}} + 1 + \tfrac{1}{\sqrt{X}}\right)\left(\sqrt{\tfrac{r\,\mu_y\,\bar{y}}{(r-1)\Lambda}} - 1 - \tfrac{1}{\sqrt{X}}\right).\end{aligned} \tag{2.22}$$

We have $\frac{\mu_y\,\bar{y}}{\Lambda} = \frac{\Lambda-\mu_x\bar{x}}{\Lambda} < 1$. Hence for $X \le X^* = \frac{x^*}{\bar{x}} = \frac{(r-1)\beta}{\mu_m}x^*$ we have the following: $\sqrt{\frac{r\,\mu_y\,\bar{y}}{(r-1)\Lambda}} - 1 - \frac{1}{\sqrt{X}} < \sqrt{\frac{r}{(r-1)}} - \frac{\sqrt{\mu_m}}{\sqrt{(r-1)\beta x^*}} - 1 \le 0$, since by assumption $\frac{\beta x^*}{\mu_m} = \frac{\beta\Lambda}{\mu_x\mu_m} \le (\sqrt{r}+\sqrt{r-1})^2$. Therefore, the derivative of V_{EE} along the trajectories of system (1.1) is negative definite on the set $\mathcal{D}_0 = \{(x,y,m) \in \mathbb{R}^3_+ \,:\, 0 < x \le x^*,\ y > 0,\ m > 0\}$. By continuity, there exists $\epsilon > 0$ such that $\dot{V}_{EE}$ is negative definite on the set $\mathcal{D}_\epsilon = \{(x,y,m) \in \mathbb{R}^3_+ \,:\, 0 < x < x^* + \epsilon,\ y > 0,\ m > 0\}$. The global asymptotic stability of the EE follows from the fact that $\mathcal{D}_\epsilon$ is an absorbing set for system (1.1). □

3. The general case: n strains with k classes of parasitized erythrocytes. We define the following system with k classes and n parasite strains:

$$
\left\{
\begin{array}{l}
\dot{x} = f(x) - \mu_x x - \; x \displaystyle\sum_{i=1}^{n} \beta_i\, m_i = \varphi(x) - \; x \sum_{i=1}^{n} \beta_i\; m_i \\
\text{and for } i = 1, \ldots, n, \\
\dot{y}_{1,i} = \beta_i x\, m_i - \alpha_{1i}\, y_{1,i}, \\
\dot{y}_{2,i} = \gamma_{1,i}\, y_{1,i} - \alpha_{2,i}\, y_{2,i}, \\
\cdots \\
\dot{y}_{k,i} = \gamma_{k-1,i}\, y_{k-1,i} - \alpha_{k,i}\, y_{k,i}, \\
\dot{g}_i = \delta_i\, y_{k,i} - \mu_{g_i}\, g_i, \\
\dot{m}_i = r_i\, \gamma_{k,i}\, y_{k,i} - \mu_{m_i}\, m_i - u\, \beta_i\, x\, m_i.
\end{array}
\right. \tag{3.1}
$$

As in preceding sections we rewrite the system as

$$
\left\{
\begin{array}{l}
\dot{x} = \varphi(x) - \; x \displaystyle\sum_{i=1}^{n} \beta_i \,\langle\, z_i \mid e_{i,\omega} \,\rangle \\
\text{and for } i = 1, \ldots, n, \\
\dot{z}_i = x\, \beta_i \,\langle\, z_i \mid e_{i,\omega} \,\rangle\, e_{i,1} + A_i\, z_i - u\, x\, \beta_i \,\langle\, z_i \mid e_{i,\omega} \,\rangle\, e_{i,\omega},
\end{array}
\right. \tag{3.2}
$$

where the matrix A_i is the analogous of the matrix A_0 defined in section 2.2, but corresponding to the genotype i, and the vectors $e_{i,1}$ and $e_{i,\omega}$ are defined accordingly. We drop the index 0 in A for readability.

THEOREM 3.1. *We consider the system* (3.1) *with the hypotheses* (2.3) *satisfied. We define the basic reproduction ratio* $\mathcal{R}_0$ *of the system* (3.1) *by*

$$\mathcal{R}_0^i = \frac{r_i \beta_i x^*}{\mu_{m_i} \; + u\, \beta_i\, x^*} \, \frac{\gamma_{1,i} \cdots \gamma_{k,i}}{\alpha_{1,i} \cdots \alpha_{k,i}}$$

and

$$\mathcal{R}_0 = \max_{i=1,\ldots,n} \mathcal{R}_0^i.$$

1. *The system* (3.1) *is GAS on* $\mathbb{R}_+$ *at the DFE* $(x^*, 0, \ldots, 0)$ *if and only if* $\mathcal{R}_0 \le 1$.

2. *If* $\mathcal{R}_0 > 1$, *then the DFE is unstable. If* $R_0^i > 1$, *there exists an EE in the nonnegative orthant corresponding to the genotype* i, *the value for the other indexes*

$j \neq i$ are $y_j = m_j = 0$, and

$$\text{(3.3)} \qquad \begin{cases} \bar{x}_i = \dfrac{\mu_{m_i}}{\beta_i \left[r_i \dfrac{\gamma_{1,i} \cdots \gamma_{k,i}}{\alpha_{1,i} \cdots \alpha_{k,i}} - u \right]}, \\ \bar{z}_i = \varphi(\bar{x}_i)\ (-A_i)^{-1} (e_{i,1} - u\, e_{i,\omega}), \\ \bar{g}_i = \dfrac{\delta_i}{\mu_{g_i}} \bar{z}_{i,k}, \end{cases}$$

where we denote by $\bar{z}_{i,k}$ the kth *component of $\bar{z}_i$.*

3. *We assume $\mathcal{R}_0 > 1$. We define $\mathcal{T}_0^i$ as in Definition* 2.2. *We assume that the generic conditions $\mathcal{T}_0^i \neq \mathcal{T}_0^j$ are satisfied for $i \neq j$. We suppose that the genotypes have been indexed such that*

$$\mathcal{T}_0^1 > \mathcal{T}_0^2 \geq \cdots \geq \mathcal{T}_0^n.$$

Then the EE corresponding to $\bar{x}_1$ is asymptotically stable and the EEs corresponding to $\bar{x}_j$ for $j \neq 1$ (for those which are in the nonnegative orthant) are unstable.

4. *We assume that the preceding hypothesis $\mathcal{T}_0^1 > \mathcal{T}_0^j$ is satisfied with $\mathcal{R}_0 > 1$. We denote it by $\alpha^* = -\max_{x \in [0, x^*]} (\varphi'(x))$. Then if*

$$u\, \beta_1\, \varphi(\bar{x}_1) \leq \mu_{m_1}\, \alpha^*,$$

the equilibrium $(\bar{x}_1, \bar{y}_1, \bar{m}_1, \bar{g}_1, 0, \ldots, 0)$ is GAS on the orthant minus the x-axis and the faces of the orthant defined by $y_1 = m_1 = g_1 = 0$. In other words the most virulent strain is the winner and the other strains go extinct.

Proof. As in Theorem 2.1 there exists a forward invariant compact absorbing set in the nonnegative orthant for the system (3.1), and hence all the forward trajectories are bounded. The variables g_i do not affect the dynamical evolution of the variables $x,\ y_{i,j},\ m_i$, and so we can consider the system without the production of gametocytes. We use the Liapunov function

$$V_{DFE}(z) = \sum_{i=1}^{n} V_{DFE}(z_i) = \sum_{i=1}^{n} \beta_i\, x^* \langle\, e_{i,\omega} \mid (-A_i^{-1}) z_i \,\rangle.$$

Using the system written as (3.2) and the computation (2.11) we easily obtain

$$\dot{V}_{DFE} = \sum_{i=1}^{n} \beta_i \langle\, e_{i,\omega} \mid z_i \,\rangle \left(\mathcal{T}_0^i\, x - x^* \right).$$

Now we define the Liapunov function on the nonnegative orthant minus the hyperplane face $x = 0$

$$V(x, z) = (x - x^*\, \ln x) - x^*(1 - \ln x^*) + \sum_{i=1}^{n} V_{DFE}(z_i)$$

which gives

$$\begin{aligned} \dot{V} &= \frac{x - x^*}{x}\, \varphi(x) + \sum_{i=1}^{n} x^* \beta_i \langle\, z_i \mid e_{i,\omega} \,\rangle - \sum_{i=1}^{n} x \beta_i \langle\, z_i \mid e_{i,\omega} \,\rangle \\ &\quad + \sum_{i=1}^{n} \beta_i \langle\, e_{i,\omega} \mid z_i \,\rangle \left(\mathcal{T}_0^i\, x - x^* \right) \\ &= \frac{x - x^*}{x}\, \varphi(x) + \sum_{i=1}^{n} \beta_i \langle\, e_{i,\omega} \mid z_i \,\rangle\, x \left(\mathcal{T}_0^i\ - 1 \right). \end{aligned}$$

Since $\mathcal{R}_0^i \leq 1$ for all index i, we have $\mathcal{T}_0^i \leq 1$, and hence $\dot{V} \leq 0$. The conclusion follows by Lasalle's invariance principle and consideration of the boundary of the positive orthant.

Now we assume $\mathcal{R}_0 > 1$. The instability of the DFE follows from the properties of $\mathcal{R}_0$ [15]. We assume that the genotypes are indexed such that their corresponding threshold are in decreasing order $\mathcal{T}_0^1 > \mathcal{T}_0^2 \geq \cdots \geq \mathcal{T}_0^n$.

We will define a Liapunov function on the nonnegative orthant minus the manifold defined by the equations $x = y_1 = m_1 = 0$. For this we need to recall the definition of the function $V_{EE}(x, y_1, m_1)$ defined in (2.14):

$$V_{EE}(x, y, m) = a(x - \bar{x} \ln x) + \sum_{i=1}^{k} b_{1,i} \left(y_{1,i} - \bar{y}_{1,i} \ln y_{1,i}\right) + b_{1,k+1} \left(m_1 - \bar{m}_1 \ln m_1\right).$$

The coefficients $(a, b_{1,i})$ are positive and defined from A_1 as in the proof of Theorem 2.1 from section 2.2. We also use the function V_{EE} defined in (2.10) to consider

$$V(x, z) = \mathcal{T}_0^1 V_{EE}(x, z_1) + a \sum_{i=2}^{n} V_{DFE}(z_i)$$

or equivalently

$$V(x, z) = \mathcal{T}_0^1 V_{EE}(x, z_1) + a \sum_{i=2}^{n} \beta_i x^* \langle e_{i,\omega} \mid (-A_i^{-1}) z_i \rangle.$$

Using the relation (2.19) and (2.11), we can compute the derivative of V along the trajectories of (3.2):

$$\dot{V} = \mathcal{T}_0^1 \Phi(x, z_1) + a \mathcal{T}_0^1 \sum_{i=2}^{n} \beta_i \bar{x}_1 \langle e_{i\omega} \mid z_i \rangle - a \mathcal{T}_0^1 \sum_{i=2}^{n} \beta_i x \langle e_{i\omega} \mid z_i \rangle + a \sum_{i=2}^{n} \beta_i \langle z_i \mid e_{i,\omega} \rangle \left(\mathcal{T}_0^i x - x^*\right).$$

Using $\mathcal{T}_0^1 \bar{x}_1 = x^*$ from the Definition 2.2 for the threshold we get

$$\dot{V} = \mathcal{T}_0^1 \Phi(x, z_1) + a \sum_{i=2}^{n} \beta_i \langle z_i \mid e_{i,\omega} \rangle x \left(\mathcal{T}_0^i - \mathcal{T}_0^1\right) \leq 0.$$

By Liapunov theorem this ends the proof for the stability. The global asymptotic stability is obtained by a straightforward use of LaSalle's invariance principle, which ends the proof of Theorem 3.1. □

Remark 4. In the nongeneric case it can be shown, with the help of the Liapunov functions used in the theorem, that there exists a continuum of stable EE. We omit the proof.

In the generic case, the dynamics of the system are completely determined. The nonnegative orthant is stratified in the union of stable manifolds corresponding to the different equilibria. Only the equilibrium corresponding to the winning strain has a basin of attraction with a nonempty interior.

Remark 5. We have proved that the most virulent strain, that is, the strain which maximizes its respective threshold $\mathcal{T}_0^i$, eliminates the other. We obtain the

same kind of result as in [7], where the authors consider a SIR model with n strains of parasite. They consider that infection by one parasite strain excludes superinfection by other strains (this is also our case) and induces permanent immunity against all strains in case of recovery. They also guarantee limited population by considering a recruitment depending on the density in a monotone decreasing way. They find that the strain which maximizes the basic reproduction ratio eliminates the others. In the case considered by the authors, actually, using our notation, $\mathcal{R}_0 = \frac{x^*}{\bar{x}}$. In fact in this model $\mathcal{T}_0$ and $\mathcal{R}_0$ coincide. This is also the case in our model when $u = 0$. Hence our result compares with the result of [7]. However in the case $u \neq 0$ this is $\mathcal{T}_0^i$, and not $\mathcal{R}_0^i$, which distinguishes the fate of the strain. Our result is then different from [7], where this role is devoted to $\mathcal{R}_0$. The same kind of remarks apply to [10] and [11].

Remark 6. In our model the chains are of equal length for each strain. If the chains are of unequal length, the proof is unchanged. We use equal length for notational convenience. A reason to have unequal length could be to model different behavior for two different strains of the parasite.

4. Conclusion. In this article we have given a parasitic within-host model and have provided a stability analysis of this model.

This model incorporates a number k of compartments for the parasitized target cells and considers n strains for the parasite. The rationale for including multicompartments can be multiple. One reason is to take into account biological reasons, e.g., consideration of morphological or age classes. The second is for behavioral modeling reasons, e.g., to model delays described by gamma distribution functions.

This model has been conceived from malaria infection, since it is well grounded that malaria is a multistrain infection. However other parasitic infections can be considered by this model.

We prove that if the basic reproduction number satisfies $\mathcal{R}_0 \leq 1$, then the DFE is GAS; i.e., the parasite is cleared from the host. Our stability result when $\mathcal{R}_0 > 1$ can be summarized as a competitive exclusion principle. To each i-strain we associate an individual threshold condition $\mathcal{T}_0^i$ as in Definition 2.2. If $\mathcal{R}_0 > 1$, if one strain has its individual threshold strictly larger than the thresholds of the other strains and if a mild sufficient condition is satisfied (for a constant recruitment, i.e., $f(x) = \Lambda$, this condition is simply $u\beta\Lambda \leq \frac{r}{r-u}\,\mu_x\mu_m$), then there exists a GAS equilibrium on the positive orthant. This equilibrium corresponds to the extinction of all strains, except the strain with the largest threshold. This winning strain maximizes the threshold and not its individual basic reproduction number, which is different from previous analogous results of the literature.

REFERENCES

[1] P. Adda, J. L. Dimi, A. Iggidr, J. C. Kamgang, G. Sallet, and J. J. Tewa, *General models of host-parasite systems. Global analysis*, Discrete Contin. Dyn. Syst. Ser. B, to appear.

[2] R. M. Anderson, *Complex dynamic behaviours in the interaction between parasite population and the host's immune system*, Int. J. Parasitol, 28 (1998), pp. 551–566.

[3] R. M. Anderson, R. M. May, and S. Gupta, *Non-linear phenomena in host-parasite interactions*, Parasitology, 99 (1989), pp. 59–79.

[4] A. Beretta and V. Capasso, *On the general structure of epidemic systems. Global asymptotic stability*, Comput. Math. Appl. Ser. A, 12 (1986), pp. 677–694.

[5] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994.

[6] N. P. Bhatia and G. P. Szegö, *Dynamical Systems: Stability Theory and Applications*, Springer-Verlag, Berlin, New York, 1967.

[7] H. J. Bremermann and H. R. Thieme, *A competitive exclusion principle for pathogen virulence*, J. Math. Biol., 27 (1989), pp. 179–190.
[8] G. J. Butler, H. S. B. Hsu, and P. Waltman, *Coexistence of competing predator in a chemostat*, J. Math. Biol., 17 (1983), pp. 133–151.
[9] C. Castillo-Chavez, Z. Feng, and W. Huang, *On the computation of R_0 and its role on global stability*, in Mathematical Approaches for Emerging and Reemerging Infectious Diseases: An Introduction (Minneapolis, MN, 1999), IMA Vol. Math. Appl. 125, Springer-Verlag, New York, 2002, pp. 229–250.
[10] C. Castillo-Chavez, W. Huang, and J. Li, *Competitive exclusion in gonorrhea models and other sexually transmitted diseases*, SIAM J. Appl. Math., 56 (1996), pp. 494–508.
[11] C. Castillo-Chavez, W. Huang, and J. Li, *Competitive exclusion and coexistence of multiple strains in an SIS STD model*, SIAM J. Appl. Math., 59 (1999), pp. 1790–1811.
[12] W. E. Collins and G. M. Jeffery, *A retrospective examination of the patterns of recrudescence in patients infected with plasmodium falciparum*, Am. J. Trop. Med. Hyg., 61 (1999), pp. 44–48.
[13] P. De Leenheer and H. L. Smith, *Virus dynamics: A global analysis*, SIAM J. Appl. Math., 63 (2003), pp. 1313–1327.
[14] H. H. Diebner, M. Eichner, L. Molineaux, W. E. Collins, G. M. Jeffery, and K. Dietz, *Modelling the transition of asexual blood stages of Plasmodium falciparum to gametocytes*, J. Theoret. Biol., 202 (2000), pp. 113–127.
[15] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz, *On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations*, J. Math. Biol., 28 (1990), pp. 365–382.
[16] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz, *Mathematical Epidemiology of Infectious Diseases Model Building, Analysis and Interpretation*, Wiley Ser. Math. Comput. Biol., John Wiley and Sons, Chichester, 2000.
[17] K. Dietz, *Epidemiologic interference of virus population*, J. Math. Biol., 8 (1979), pp. 291–300.
[18] B. S. Goh, *Global stability in two species interactions*, J. Math. Biol., 3 (1976), pp. 313–318.
[19] B. S. Goh, *Global stability in many-species systems*, Amer. Natur., (1977), pp. 135–143.
[20] M. B. Gravenor and D. Kwiatkowski, *An analysis of the temperature effects of fever on the intra-host population dynamics of plasmodium falciparum*, Parasitology, 117 (1998), pp. 97–105.
[21] M. B. Gravenor and A. L. Lloyd, *Reply to: Models for the in-host dynamics of malaria revisited: Errors in some basic models lead to large over-estimates of growth rates*, Parasitology, 117 (1998), pp. 409–410.
[22] M. B. Gravenor, A. L. Lloyd, P. G. Kremsner, M. A. Missinou, M. English, K. Marsh, and D. Kwiatkowski, *A model for estimating total parasite load in falciparum malaria patients*, J. Theoret. Biol., 217 (2002), pp. 137–48.
[23] M. B. Gravenor, A. R. McLean, and D. Kwiatkowski, *The regulation of malaria parasitaemia: Parameter estimates for a population model*, Parasitology, 110 (1995), pp. 115–122.
[24] M. B. Gravenor, M. B. van Hensbroek, and D. Kwiatkowski, *Estimating sequestered parasite population dynamics in cerebral malaria*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 7620–7624.
[25] B. Hellriegel, *Modelling the immune response to malaria with ecological concepts: Short-term behaviour against long-term equilibrium*, Proc. R. Soc. Lond. Ser. B Biol. Sci., 250 (1992), pp. 249–256.
[26] H. W. Hethcote and H. R. Thieme, *Stability of the endemic equilibrium in epidemic models with subpopulations*, Math. Biosci., 75 (1985), pp. 205–227.
[27] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Rev., 42 (2000), pp. 599–653.
[28] C. Hetzel and R. M. Anderson, *The within-host cellular dynamics of bloodstage malaria: Theoretical and experimental studies*, Parasitology, 113 (1996), pp. 25–38.
[29] M. W. Hirsch, *The dynamical systems approach to differential equations*, Bull Amer. Math. Soc. (N.S.), 11 (1984), pp. 1–64.
[30] M. B. Hoshen, R. Heinrich, W. D. Stein, and H. Ginsburg, *Mathematical modelling of the within-host dynamics of Plasmodium falciparum*, Parasitology, 121 (2001), pp. 227–235.
[31] J. A. Jacquez, *Compartmental Analysis in Biology and Medicine*, Biomedware, Ann Arbor, MI, 1996.
[32] J. A. Jacquez and C. P. Simon, *Qualitative theory of compartmental systems*, SIAM Rev., 35 (1993), pp. 43–79.
[33] J. A. Jacquez, C. P. Simon, and J. Koopman, *Core groups and the R_0s for subgroups in heterogeneous SIS and SI models*, in Epidemics Models: Their Structure and Relation to Data, D. Mollison, ed., Cambridge University Press, Cambridge, UK, 1996, pp. 279–301.

[34] A. Korobeinikov and G. C. Wake, *Lyapunov functions and global stability for SIR and SIRS and SIS epidemiolgical models*, Appl. Math. Lett., 15 (2002), pp. 955–961.
[35] A. Korobeinikov and P. K. Maini, *A Lyapunov function and global properties for SIR and SEIR epidemiological models with nonlinear incidence*, Math. Biosci. Eng., 1 (2004), pp. 57–60.
[36] A. Lajmanovich and J. A. Yorke, *A deterministic model for gonorrhea in a nonhomogeneous population*, Math. Biosci., 28 (1976), pp. 221–236.
[37] J. P. LaSalle and S. Lefschetz, *Stability by Liapunov's Direct Method with Applications*, Academic Press, New York, 1961.
[38] J. P. LaSalle, *The Stability of Dynamical Systems*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 25, SIAM, Philadelphia, 1976.
[39] S. A. Levin and D. Pimentel, *Selection of intermediate rates increase in parasite-host systems*, Amer. Natur., (1981), pp. 308–315.
[40] M. Y. Li, J. R. Graef, L. Wang, and J. Karsai, *Global dynamics of a SEIR model with varying total population size*, Math. Biosci., 160 (1999), pp. 191–213.
[41] M. Y. Li, J. S. Muldowney, and P. van den Driessche, *Global stability for the SEIR model in epidemiology*, Math. Biosci., 125 (1995), pp. 155–164.
[42] M. Y. Li, J. S. Muldowney, and P. van den Driessche, *Global stability of SEIRS models in epidemiology*, Can. Appl. Math. Q., 7 (1999), pp. 409–425.
[43] M. Y. Li and J. S. Muldowney, *Global stability for the SEIR model in epidemiology*, Math. Biosci., 125 (1995), pp. 155–164.
[44] M. Y. Li and J. S. Muldowney, *A geometric approach to global-stability problems*, SIAM J. Math. Anal., 27 (1996), pp. 1070–1083.
[45] M. Y. Li, H. L. Smith, and L. Wang, *Global dynamics of an SEIR epidemic model with vertical transmission*, SIAM J. Appl. Math., 62 (2001), pp. 58–69.
[46] X. Lin and J. W.-H. So, *Global stability of the endemic equilibrium and uniform persistence in epidemic models with subpopulations*, J. Aust. Math. Soc. Ser. B, 34 (1993), pp. 282–295.
[47] A. L. Lloyd, *The dependence of viral parameter estimates on the assumed viral life cycle: Limitations of studies of viral load data*, Proc. R. Soc. Lond. Ser. B Biol. Sci., (2001), pp. 847–854.
[48] A. L. Lloyd, *Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods*, Proc. R. Soc. Lond. Ser. B Biol. Sci., 268 (2001), pp. 985–993.
[49] A. L. Lloyd, *Realistic distributions of infectious periods in epidemic models: Changing patterns of persistence and dynamics*, Theor. Popul. Biol., 60 (2001), pp. 59–71.
[50] D. G. Luenberger, *Introduction to dynamic systems. Theory, models, and applications*, John Wiley and Sons, New York, 1979.
[51] N. MacDonald, *Time Lags in Biological Models*, Springer-Verlag, Berlin, New York, 1978.
[52] D. P. Mason, F. E. McKenzie, and W. H. Bossert, *The blood-stage dynamics of mixed plasmodium malariae-plasmodium falciparum infections*, J. Theoret. Biol., 198 (1999), pp. 549–566.
[53] R. M. May and R. M. Anderson, *Epidemiology and genetics in the coevolution of parasites and hosts*, Proc. R. Soc. Lond. Ser. B. Biol. Sci., 219 (1983), pp. 281–313.
[54] J. Maynard Smith, *Models in Ecology*, Cambridge University Press, Cambridge, UK, 1974.
[55] F. E. McKenzie and W. H. Bossert, *The dynamics of plasmodium falciparum blood-stage infection*, J. Theoret. Biol., 188 (1997), pp. 127–140.
[56] F. E. McKenzie and W. H. Bossert, *The optimal production of gametocytes by plasmodium falciparum*, J. Theoret. Biol., 193 (1998), pp. 419–428.
[57] P. G. McQueen and F. E. McKenzie, *Age-structured red blood cell susceptibility and the dynamics of malaria infections*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 9161–9166.
[58] L. Molineaux, H. H. Diebner, M. Eichner, W. E. Collins, G. M. Jeffery, and K. Dietz, *Plasmodium falciparum parasitaemia described by a new mathematical model*, Parasitology, 122 (2001), pp. 379–391.
[59] L. Molineaux and K. Dietz, *Review of intra-host models of malaria*, Parassitologia, 41 (2000), pp. 221–231.
[60] M. A. Nowak and R. M. May, *Virus Dynamics. Mathematical Principles of Immunology and Virology*, Oxford University Press, Oxford, UK, 2000.
[61] A. S. Perelson, D. E. Kirschner, and R. De Boer, *Dynamics of HIV infection of CD4+ T cells*, Math. Biosci., 114 (93), pp. 81–125.
[62] A. S. Perelson and P. W. Nelson, *Mathematical analysis of HIV-1 dynamics in vivo*, SIAM Rev., 41 (1999), pp. 3–44.
[63] J. Prüss, L. Pujo-Mejouet, G. F. Webb, and R. Zacher, *Analysis of a model for the dynamics of prions*, Discrete Contin. Dyn. Syst. Ser. B, 6 (2006), pp. 225–235.

[64] A. Saul, *Models for the in-host dynamics of malaria revisited: errors in some basic models lead to large over-estimates of growth rates*, Parasitology, 117 (1998), pp. 405–407, 409–410.

[65] C. P. Simon, J. A. Jacquez, and J. S. Koopman, *A Lyapunov function approach to computing* R_0, in Models for Infectious Human Diseases: Their Structure and Relation to Data, V. Isham and G. Medley, eds., Cambridge University Press, Cambridge, UK, 1996, pp. 311–314.

[66] J. Swinton, *The dynamics of blood-stage malaria: Modelling strain specific and strain transcending immunity*, in Models for Infectious Human Diseases: Their Structure and Relation to Data, V. Isham and G. Medley, eds., Cambridge University Press, Cambridge, UK, 1996, pp. 210–212.

[67] H. R. Thieme, *Global asymptotic stability in epidemic models*, in Equadiff 82, Lecture Notes in Math 1017, W. Knobloch and K. Schmitt, eds., Springer-Verlag, Berlin, 1983, pp. 608–615.

[68] H. R. Thieme, *Mathematics in Population Biology*, Princeton Ser. Theor. Comput. Biol., Princeton University Press, Princeton, NJ, 2003.

SIAM J. APPL. MATH.
Vol. 67, No. 1, pp. 279–304

A SPHERICAL PARTICLE MOVING SLOWLY IN A FLUID WITH A RADIALLY VARYING VISCOSITY*

SHIMON HABER†

Abstract. The Stokesian flow field induced by a spherical particle that undergoes a slow rotational and translational motion in an unbounded quiescent fluid with radially varying viscosity is investigated. For a rotating particle, it is demonstrated that only the near viscosity field contributes effectively to the hydrodynamic torque exerted on the particle. A powerful screening effect exists in which the contribution of the distant viscosity field is weighted with the inverse of the distance from the particle to the fourth power. Two specific cases are investigated in which the viscosity field varies either exponentially or periodically. The latter is of particular interest, since it may serve to model the torque exerted on a particle rotating in a suspension. It is shown that a small test particle rotating in a suspension consisting of larger particles is almost unaffected by the large particles and mainly "senses" the fluid viscosity, whereas a large test particle "senses" the suspension viscosity. For a translating particle, a general expression is obtained for the induced velocity, pressure, and drag force exerted on the particle. An approximate result is obtained for the case in which the viscosity field varies slowly with the distance from the particle. An exact solution is obtained for a case in which the viscosity field varies algebraically with the distance from the particle center. An explicit numerical scheme is also suggested, which may assist in obtaining the drag force exerted on a particle translating in a flow field with an arbitrary radially varying viscosity distribution. Based on this numerical scheme and on the approximate solution obtained for a slowly varying viscosity, the drag force exerted on a particle translating inside a fluid with periodically varying viscosity is calculated. We hypothesize that such a periodic distribution can be viewed as a suspension under low Péclet number conditions. Based on this assumption, we obtain that if a test particle is much smaller than the suspended particles, the initial drag force exerted on the test particle is insensitive to the composition of the suspended particles or droplets, and senses the viscosity of only the continuous liquid, provided that the test particle is far from the suspended particles. However, up to first order in suspension concentration, the apparent viscosity of a dilute suspension is indifferent to whether the test particle is forced to move with a constant velocity or is subjected to a constant external force, provided that the test particle is arbitrarily located between the suspended particles.

Key words. Stokes flows, variable viscosity, particle translation, particle rotation, suspensions

AMS subject classification. 76D07

DOI. 10.1137/S0036139903429610

1. Introduction.

The fluid-dynamical problem addressed in this paper is motivated by the search for insight into the transport of particles in suspensions or in fluids that may experience strong temperature gradients. Though the backbone of the paper consists of an analytical investigation of a flow model for a spherical particle motion in a fluid with variable viscosity, the introduction also provides a brief elucidation of the foregoing physical problems.

Suspension hydrodynamics was the focus of extensive past investigation. Past models attempted to explain such effects as hindered settling velocity of a suspension under gravity (e.g., Batchelor (1972)), increase in the suspension effective viscosity (e.g., the Einstein viscosity of dilute suspensions), and the effect of shear induced diffusion (e.g., Leighton and Acrivos (1987a,b) and Gadala-Maria and Acrivos (1980))

*Received by the editors June 12, 2003; accepted for publication (in revised form) August 7, 2006; published electronically December 15, 2006. This research was supported by the Fund for the Promotion of Research at the Technion.

http://www.siam.org/journals/siap/67-1/42961.html

†Department of Mechanical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel (mersh01@tx.technion.ac.il).

in which particles in a suspension cross streamlines, a phenomenon that a single freely suspended particle in shear flow does not experience (Happel and Brenner (1983)).

It is well accepted that the fundamental mechanism governing all of the foregoing results stems from multiparticle hydrodynamic interactions. However, an analytical solution of a multiparticle system is not tractable, and basically three important approaches were used to circumvent this difficulty. The first, numerical, approach termed Stokesian dynamics was applied by Brady (1988) and by others (e.g., Hassonjee, Ganatos, and Pfeffer (1988), Brenner et al. (1990), Hassonjee, Pfeffer, and Ganatos (1992), Chang and Powell (1993), Nott and Brady (1994)), in which many suspended particles were tracked simultaneously using either collocation or boundary integral methods.

A second approach was to apply simplified models for the complex microstructure of suspensions. For instance, the microscopic conformation of a suspension was viewed as a spatially periodic array. In essence, a unit cell approach was addressed that requires the hydrodynamic solution of a single particle in a bounded field. Yet another simplified approach was to solve the flow field generated by a generic two/three body subsystem. In many cases, the foregoing models were sufficiently simple to be handled analytically and to encapsulate the main effects stemming from hydrodynamic interactions between the particles. For example, Zuzovsky, Adler, and Brenner (1983) and Adler, Zuzovsky, and Brenner (1985) used a spatially periodic model to obtain the rheological properties of a suspension. Batchelor (1972) used two-body interactions to calculate the sedimentation velocity of a dilute suspension of rigid spherical particles under gravity; Batchelor and Green (1972b) used two-body interaction to calculate the rheology of a dilute suspension; Haber, Brenner, and Shapira (1990) used two-body interactions to derive the dispersion coefficient of a dilute suspension containing flexible dumbbells; and Wang, Mauri, and Acrivos (1998) utilized a three-body interaction to calculate the transverse shear induced gradient diffusion of dilute suspensions, etc.

A third phenomenological approach was used by Leighton and Acrivos (1987b), Philips et al. (1992), and others to explain shear induced migration. Thus, for instance, a self diffusion shear induced coefficient was defined that was based on scaling considerations and the available experimental data, circumventing the microscopic details of the problem and providing a direct macroscopic view. It proved quite successful for the case of a narrow gap Couette device (Acrivos, Mauri, and Fan (1993)).

The mathematical problem addressed in this paper deals with the flow field generated by a single particle rotating and translating in an unbounded single phase fluid with variable viscosity. We suggest that the solution of the foregoing problem may possibly be utilized to explore the mechanisms that govern the motion of a single test particle immersed in suspensions of various concentrations. Such an interpretation of the results stems from the observation that a suspension (or an emulsion) of particles (or droplets) is a two phase fluid with two different viscosities, which may be approximated by a single phase fluid with a *continuous* variable viscosity field. One can further assume that the suspension is at equilibrium where the concentration distribution ϕ_∞ satisfies the equation (e.g., Brenner (1979))

$$\nabla \cdot [De^{-E}\nabla(e^{E}\phi_\infty)] = 0. \tag{1a}$$

Here E is the energy potential function, and D stands for diffusion coefficient of the particles comprising the suspension. Henceforth, we assume that E is radially symmetric (i.e., the external force is centrally symmetric), and thereby the solution of (1a) yields a radially symmetric function ϕ_∞.

Introduction of a moving "test particle" into the suspension induces a velocity *disturbance* $\mathbf{v}'$, which in turn may cause a concentration *disturbance* ϕ' that destroys the assumed radial symmetry. A first order approximation for the differential equation governing this concentration disturbance is

$$\frac{\partial \phi'}{\partial t} + \mathbf{v}' \cdot \nabla \phi_\infty = \nabla \cdot [De^{-E} \nabla (e^E \phi')], \tag{1b}$$

where $\mathbf{v}'$ scales with the test particle velocity. Thus, the significance of the symmetry-destroying convective term vis-à-vis the restoring diffusion term is determined by the Péclet number based on the test particle velocity and diameter and the diffusion coefficient of the suspended particles. If this Péclet number is much smaller than unity, the diffusion mechanism will rapidly restore the concentration to its initial equilibrium concentration. Consequently, the initial equilibrium concentration distribution will practically prevail for all times. As shall be shown later, the small Péclet number assumption is not required for the case of a rotating sphere in a flow field with a radially symmetric viscosity (the convective term vanishes identically). For the case of a translating sphere, however, this assumption must be made so that local equilibrium is restored rapidly and the solution is meaningful for all times.

The paper is divided into the following main sections. In section 2 a simple closed analytical solution is obtained for the flow field induced by a sphere rotating in an unbounded fluid with radially varying viscosity. In section 3 the flow field generated by a sphere translating in an unbounded flow field is addressed.

In section 4 several examples are investigated exploiting the general expressions obtained in sections 2 and 3. For a rotating sphere, an exact solution is obtained for the case in which the viscosity field increases or decreases exponentially with the distance from the test particle. For a translating sphere, three cases are addressed. In case A, a weak radial variation of the viscosity field is assumed, and an approximate analytical solution is derived. In case B, the viscosity field increases or decreases algebraically with the distance from the test particle, and an exact solution is obtained. In case C, based on a semianalytical approach, an efficient numerical algorithm is suggested, by which the drag force can be obtained for general viscosity distributions.

In section 5 a solution is obtained for a test particle rotating and translating in a radially periodic viscosity field. The results are interpreted vis-à-vis particle motion in a homogeneous suspension. In section 6 a summary of the results is provided.

2. A sphere rotating in a field with radially varying viscosity.

2.1. Statement of problem. A spherical rigid particle of radius a rotates slowly with angular velocity ω inside an unbounded incompressible flow field with initial radially varying viscosity

$$\mu = \mu_0 \lambda(r), \tag{2}$$

where r is the radial distance measured from the center of the particle and μ_0 is a characteristic viscosity of the fluid. The Reynolds number based on a, ω, μ_0 and the fluid density is assumed to be smaller than unity, so that the quasi-steady creeping flow equations apply (Gurbebeck and Sprossig (1993)), namely,

$$\mu \nabla^2 \mathbf{v} + 2\nabla \mu \cdot \mathbf{S} = \nabla p, \qquad \nabla \cdot \mathbf{v} = 0, \tag{3a}$$

where $\mathbf{v}$ is the fluid velocity, p is the pressure, and $\mathbf{S}$ is the rate of strain dyadic,

$$\mathbf{S} = 0.5[\nabla \mathbf{v} + (\nabla \mathbf{v})^T]. \tag{3b}$$

The velocity field satisfies the no-slip condition over the sphere boundary, namely,

$$\mathbf{v} = \boldsymbol{\omega} \times \mathbf{r} \quad \text{at} \quad |\mathbf{r}| = a, \tag{4}$$

and decays to zero at infinity. Here, $\mathbf{r}$ is the radius vector measured from the sphere center.

If we assume that the viscosity depends on the volumetric concentration $\mu = \mu(\phi)$, the viscosity field at equilibrium, $\mu_\infty \equiv \mu(\phi_\infty) \equiv \mu_0 \lambda(r)$, is also radially symmetric. However, introduction of the test sphere may introduce a viscosity *disturbance* μ' that undergoes convection and diffusion and is governed by the following first order approximation of the convection-diffusion differential equation:

$$\frac{\partial \mu'}{\partial t} + \mathbf{v} \cdot \nabla \mu_\infty = \left[\frac{d\mu}{d\phi}\right]_{\phi_\infty} \nabla \cdot \left[De^{-E} \nabla \left(e^E \left[\frac{d\phi}{d\mu}\right]_{\phi_\infty} \mu' \right) \right] \tag{5a}$$

juxtaposed with the initial condition at $t = 0$,

$$\mu' = 0. \tag{5b}$$

2.2. Method of solution. Generally, the velocity, pressure, and viscosity fields are time-dependent, and (3) to (5) must be solved simultaneously. Equation (5) is nonlinear and couples the velocity and viscosity fields. However, it will be shown that the solution for the viscosity is time-independent, and the initial viscosity field remains unchanged for all times. In such a case, the problem is linear, and the differential equation (3) and boundary condition (4) imply that the velocity and pressure fields must linearly depend upon $\boldsymbol{\omega}$. Thus, their general form in Cartesian tensor notation is

$$v_i = V_{ij}\omega_j, \qquad p = \mu_0 P_j \omega_j + p_\infty, \tag{6}$$

where V_{ij} and P_j are the velocity second rank tensor and the pressure vector, respectively. Substituting (6) into (3), (4) and using (2) yields

$$\lambda \frac{\partial^2 V_{ij}}{\partial x_l \partial x_l} + \frac{\partial \lambda}{\partial x_l}\left(\frac{\partial V_{ij}}{\partial x_l} + \frac{\partial V_{lj}}{\partial x_i}\right) = \frac{\partial P_j}{\partial x_i}, \quad \frac{\partial V_{ij}}{\partial x_i} = 0, \tag{7}$$

$$V_{ij} = \varepsilon_{ijk} x_k \quad \text{at} \quad |\mathbf{r}| = a, \tag{8}$$

where V_{ij} vanishes at infinity and ε_{ijk} is the third rank permutation pseudotensor.

It is clear from (7) and (8) that V_{ij} and P_j must be pseudotensors that depend only on the permutation tensor ε_{ijk}, the particle radius a, the radius vector x_i, and its magnitude r. The latter is due to the isotropy of the sphere and the radial symmetry of the viscosity field. Equations (7) and (8) also prove that neither V_{ij} nor P_j depends on the angular velocity $\boldsymbol{\omega}$.

In this case, the velocity tensor must possess the tensorial form,

$$V_{ij} = \varepsilon_{ijk} x_k f(r), \tag{9a}$$

and the pressure vector field P_i must vanish (the only possible pseudovector that combines ε_{ijk} and x_j is $\varepsilon_{ijk} x_j x_k$, which is zero identically):

$$P_i = 0. \tag{9b}$$

Here, $f(r)$ is a scalar function of r (and a) to be determined.

Introducing the flow field $\mathbf{v}$ (see (9a)) into (5) yields that the second term in the LHS of (5) vanishes identically, and consequently the viscosity disturbance vanishes identically. Thus, our basic assumption is validated, and the viscosity field remains unaltered during particle rotation.

Substituting (9) into (7) yields an ordinary differential equation for f,

$$\lambda\left(\frac{d^2f}{dr^2}+\frac{4}{r}\frac{df}{dr}\right)+\frac{d\lambda}{dr}\frac{df}{dr}=0. \tag{10}$$

From (8) and the condition at infinity we obtain that

$$f(a)=1 \quad \text{and} \quad f(\infty)=0. \tag{11}$$

The general solution of (10) subjected to boundary conditions (11) is

$$f(r)=\frac{\int_r^\infty \frac{dr}{\lambda(r)r^4}}{\int_a^\infty \frac{dr}{\lambda(r)r^4}}. \tag{12}$$

Obviously, for (12) to represent a valid solution, (9a) requires that the rf product vanish at infinity. Consequently, a formal solution exists even if the viscosity λ vanishes at infinity (approaches asymptotically to zero no faster than $1/r^3$). Clearly, this remarkable result is quite hypothetical, since in this case the Reynolds number at infinity would grow without bound, and the creeping flow equations are no longer valid.

If λ is fixed, the well-known creeping flow solution for a rotating sphere in a field with uniform viscosity is recovered. Equation (12) also manifests that the sphere "senses" the viscosity field near the sphere boundary, whereas the far field contribution is insignificant due to the r^4 factor in the integrand denominator.

If a suspension of small particles can be perceived as a single phase fluid with variable viscosity (say, λ is infinite inside the particles and uniform inside the suspending fluid), then (12) offers an approximate solution for a rotating sphere in a suspension. It clearly manifests the "screening" effect of the adjacent particles and that the contribution of particles far from the rotating sphere is negligible. A more detailed analysis of this example will be provided in section 5.

2.3. The torque exerted on the particle. The torque about the sphere center that the fluid exerts on a rotating particle is

$$(T_o)_i=\varepsilon_{ijk}\int_S x_j\sigma_{kl}n_l dS, \tag{13}$$

where "O" denotes the sphere center, n_l is the unit vector normal to the sphere boundary S, and σ_{kl} stands for the symmetric hydrodynamic stress,

$$\sigma_{kl}=\mu_0\omega_m\left[-P_m\delta_{kl}+\lambda\left(\frac{\partial V_{km}}{\partial x_l}+\frac{\partial V_{lm}}{\partial x_k}\right)\right], \tag{14}$$

where δ_{kl} is the second rank idem-tensor. Substitution of (9) and (12) into (14) and (13) and utilization of the tensorial identities

$$\int_S x_ix_jdS=\frac{4\pi a^4}{3}\delta_{ij}, \quad \int_S x_ix_jx_kx_ldS=\frac{4\pi a^6}{15}(\delta_{ij}\delta_{kl}+\delta_{ik}\delta_{lj}+\delta_{il}\delta_{kj}) \tag{15}$$

yields the following simple formula:

$$(T_o)_i = -8\pi\mu_0 a^3 \omega_i C_T, \tag{16}$$

where the dimensionless correction factor C_T is

$$C_T = \frac{1}{3}\left(a^3 \int_a^\infty \frac{dr}{\lambda(r) r^4}\right)^{-1}. \tag{17}$$

Thus, as expected, the torque is colinear with ω and resists the rotation of the particle. For a fluid with uniform viscosity, the correction factor $C_T = 1$, and the well-known Kirchoff law for a slowly rotating sphere is recovered. Equation (17) also manifests that the torque is significantly affected by the viscosity field close to the rotating sphere and that the effect of the far-field viscosity is normally negligible.

It is also convenient to view $\mu_0 C_T$ as the mean viscosity of the flow field affecting a rotating sphere. One would expect a result that is related to a weighted mean of the viscosity reciprocal. (E.g., the mean viscosity $\bar{\mu}$ of two phase fluids with viscosities μ_A and μ_B and respective volumetric concentrations ϕ_A and ϕ_B is commonly calculated by the formula $(\bar{\mu})^{-1} = \phi_A(\mu_A)^{-1} + \phi_B(\mu_B)^{-1}$.) However, the "screening-factor" of r^4 in (17) is not obvious.

3. A sphere translating in a field with radially varying viscosity (for $P_e \ll 1$).

3.1. Statement of problem. A sphere of radius a translates slowly with velocity $\mathbf{U}$ inside an unbounded incompressible flow field $\mathbf{v}$ that possesses an initial radially varying viscosity $\mu = \mu_0\lambda(r)$. The Reynolds number based on a, U, μ_0, and the fluid density is assumed to be smaller than unity, so that the quasi-steady creeping flow equations (3) can be applied. The velocity field satisfies the no-slip condition over the sphere boundary, namely,

$$\mathbf{v} = \mathbf{U} \quad \text{at} \quad |\mathbf{r}| = a, \tag{18}$$

and decays to zero at infinity.

3.2. Method of solution. The general quasi-steady solution for the velocity and pressure fields strongly depends on the time-dependent viscosity field. As time evolves, the radially symmetric structure of the viscosity field is no longer preserved. Since (5) is nonlinear and couples the velocity and viscosity fields, a general *analytic* time-dependent solution is probably hopeless. Notwithstanding, as we have shown in the introduction, if the Péclet number $P_e = Ua/D$ is much smaller than unity, the second convection term in (5) can be neglected, and the system rapidly regains its initial equilibrium radial distribution. Henceforth, we shall limit our discussion to Péclet numbers much smaller than unity, so that the initial radial viscosity distribution $\mu_\infty(r)$ remains practically unaltered.

Boundary condition (18) implies that the velocity and pressure fields must linearly depend upon $\mathbf{U}$. Thus, the general form of the velocity and pressure field (in Cartesian tensor notation) is

$$v_i = V_{ij}U_j, \qquad p = \frac{\mu_0 P_j U_j}{a} + p_\infty, \tag{19}$$

where V_{ij} and P_j are the dimensionless velocity second rank tensor and the pressure vector, respectively. It is convenient to scale distances with a and define the

dimensionless Cartesian coordinates

$$y_i = \frac{x_i}{a}, \qquad y = \frac{r}{a} = \frac{(x_i x_i)^{1/2}}{a}. \tag{20}$$

Substituting (19) and (20) into (3) and using (2) yields the differential equations

$$\lambda \frac{\partial^2 V_{ij}}{\partial y_l \partial y_l} + \frac{\partial \lambda}{\partial y_l}\left(\frac{\partial V_{ij}}{\partial y_l} + \frac{\partial V_{lj}}{\partial y_i}\right) = \frac{\partial P_j}{\partial y_i}, \qquad \frac{\partial V_{ij}}{\partial y_i} = 0, \tag{21a}$$

which are subjected to the boundary conditions

$$V_{ij} = \delta_{ij} \quad \text{at} \quad |\mathbf{y}| = 1, \tag{21b}$$

and the condition that V_{ij}, P_j vanish at infinity.

From (21), V_{ij} and P_j must be tensors that depend only upon the second rank tensor δ_{ij} and the dimensionless radius vector y. The latter is due to the isotropy of the sphere and the radial symmetry of the viscosity field. Equations (21a, b) also prove that neither V_{ij} nor P_j depends on the particle velocity U.

Consequently, the velocity and pressure tensors must possess the tensorial form

$$V_{ij} = y_i y_j g(y) + \delta_{ij} h(y), \qquad P_j = y_j q(y), \tag{22}$$

where $g(y)$, $h(y)$, and $q(y)$ are scalar functions of y to be determined.

Substitution of (22) into (21) yields three coupled ordinary differential equations for g, h, and q,

$$-q + \lambda\left(2g + \frac{2h_y}{y} + h_{yy}\right) + \lambda_y(yg + h_y) = 0, \tag{23a}$$

$$-\frac{q_y}{y} + \lambda\left(\frac{6g_y}{y} + g_{yy}\right) + \lambda_y\left(\frac{3g}{y} + 2g_y + \frac{h_y}{y^2}\right) = 0, \tag{23b}$$

$$4yg + y^2 g_y + h_y = 0, \tag{23c}$$

and boundary conditions

$$\begin{aligned} &g(1) = 0, && h(1) = 1, \\ &\left[y^2 g\right]_{y\to\infty} \to 0, && [h]_{y\to\infty} \to 0, \quad [yq]_{y\to\infty} \to 0, \end{aligned} \tag{24}$$

where the subscript y denotes differentiation with respect to y.

Elimination of h and q from (23) yields a third order equation in g:

$$g_{yyy} + \left(\frac{11}{y} + 2\frac{\lambda_y}{\lambda}\right) g_{yy} + \left(\frac{24}{y^2} + \frac{14}{y}\frac{\lambda_y}{\lambda} + \frac{\lambda_{yy}}{\lambda}\right) g_y + \left(\frac{12}{y^2}\frac{\lambda_y}{\lambda} + \frac{3}{y}\frac{\lambda_{yy}}{\lambda}\right) g = 0. \tag{25}$$

We define a new independent variable z,

$$z = \ln(y), \tag{26}$$

to transform (25) into

$$g_{zzz} + \left(8 + 2\frac{\lambda_z}{\lambda}\right) g_{zz} + \left(15 + 11\frac{\lambda_z}{\lambda} + \frac{\lambda_{zz}}{\lambda}\right) g_z + \left(9\frac{\lambda_z}{\lambda} + 3\frac{\lambda_{zz}}{\lambda}\right) g = 0, \tag{27}$$

where the subscript z denotes differentiation with respect to z.

A remarkable, exact first integral of (27) exists (notice that (36) can serve as a simple clue to its existence):

$$g_{zz} + \left(5 + \frac{\lambda_z}{\lambda}\right) g_z + 3\frac{\lambda_z}{\lambda} g = C_D \frac{e^{-3z}}{\lambda}, \tag{28}$$

where C_D is a yet undetermined constant of integration, subsequently shown to relate closely to the drag force exerted on the translating particle. Utilizing a new dependent variable G,

$$G(z) = e^{3z} g(z), \tag{29}$$

equation (28) is transformed into a simplified second order differential equation:

$$G_{zz} - \left(1 - \frac{\lambda_z}{\lambda}\right) G_z - 6G = \frac{C_D}{\lambda}. \tag{30}$$

Notice that the original Stokes equation and (30) possess only first order derivatives of the viscosity field. Hence, a second order derivative appearing in (27) is redundant and is not truly required.

The differential equation (23c) in terms of G is

$$h_z = -e^{-z}(G + G_z). \tag{31}$$

Upon substitution of (26) and (29) into (24), the transformed boundary conditions are

$$\begin{aligned} &G(z=0) = 0, \qquad h(z=0) = 1, \\ &[e^{-z}G]_{z\to\infty} \to 0, \quad [h]_{z\to\infty} \to 0 \quad [e^z q]_{z\to\infty} \to 0. \end{aligned} \tag{32}$$

Thus, a workable solution scheme is as follows: A general solution of (30) for G is obtained first; subsequently, (31) is solved for h so that boundary conditions (32) can be applied; a solution for q is then easily obtained upon direct substitution of the solutions for G and h into (23a).

3.3. The general expression for the drag force exerted on a translating particle. The drag force that the fluid exerts on a translating particle can be obtained from either of the following integrals:

$$F_i = \int_{S_B} \sigma_{il} n_l dS = \int_{S_\infty} \sigma_{il} n_l dS, \tag{33}$$

where S_B stands for the sphere boundary and S_∞ denotes any surface enclosing the particle. The second equality in (33) can easily be proven using the divergence theorem and applying the momentum equation $\partial\sigma_{il}/\partial x_i = 0$. Here, σ_{il} stands for the hydrodynamic stress:

$$\sigma_{il} = \mu_0 U_m \left[-P_m \delta_{il} + \lambda\left(\frac{\partial V_{im}}{\partial x_l} + \frac{\partial V_{lm}}{\partial x_i}\right)\right]. \tag{34}$$

Substitution of (22) into (33) and (34) and utilizing identities (15) yields

$$F_i = -\frac{4\pi}{3}\mu_0 a U_i [y^3 q(y) + 10y^3\lambda(y)g(y) + 2y^4\lambda(y)g_y(y)], \tag{35}$$

where $y > 1$ is an arbitrary radius of a sphere that encompasses the particle and whose center coincides with that of the particle. Substitution of (23a) and (26) into (35) yields

$$\begin{aligned} F_i &= \frac{4\pi}{3}\mu_0 a U_i[\lambda(g_{zz}+5g_z)+\lambda_z(g_z+3g)]e^{3z} \\ &= \frac{4\pi}{3}\mu_0 a U_i C_D. \end{aligned} \tag{36}$$

The last equality stems from (28). Thus, as expected, the drag force exerted on the particle is independent of y (or z) and depends solely upon the single yet undetermined constant of integration C_D.

4. Particular cases for various viscosity distributions. In the following section we focus on several particular cases exploiting the general formulas derived in sections 2 and 3 for the torque and drag force exerted on a rotating or translating sphere in an unbounded flow field with variable viscosity.

4.1. A rotating particle in an exponentially varying viscosity. Assume that λ varies exponentially from $1-\alpha$, near the rotating sphere, to 1 as r approaches infinity; namely,

$$\lambda = 1 - \alpha e^{-\beta(r-a)/a}. \tag{37}$$

Here β is a positive known parameter, and $\alpha < 1$ could be either a positive or a negative parameter. The correction factor C_T is calculated numerically for various values of α and β and is illustrated in Figure 1. For α and β values smaller than unity, (17) can be analytically evaluated, and the following result is obtained:

$$\begin{aligned} C_T^{-1} &= 1 + 3\sum_{n=1}^{\infty}(\alpha e^{\beta})^n E_4(n\beta) \\ &= \frac{1}{1-\alpha} - \frac{\beta\alpha e^{\beta}}{2}\sum_{n=1}^{\infty} n(\alpha e^{\beta})^{n-1}\left[e^{-n\beta}(1-n\beta) - n^2\beta^2 Ei(-n\beta)\right], \end{aligned} \tag{38}$$

where E_i and E_4 are the exponential integral functions,

$$Ei(-x) = -\int_1^{\infty}\frac{e^{-xt}}{t}dt, \qquad E_4(x) = \int_1^{\infty}\frac{e^{-xt}}{t^4}dt. \tag{39}$$

The RHS of (38) proves that, to leading order in α and β, the rotating sphere senses the viscosity field close to it. This is also verified by numerically calculating C_T (depicted in Figure 1).

4.2. A translating particle. Three different cases are addressed. In case A, a multiple-scale approximate solution of (30) is described for flow fields with weakly varying viscosity fields. In case B, an exact solution of (30) is obtained for flow fields with a rapidly varying viscosity field; namely, the viscosity field grows or decays exponentially in z-space (or algebraically in r-space). In case C we outline an analytical/numerical method for viscosity fields that increase asymptotically at a smaller than exponential rate in z-space.

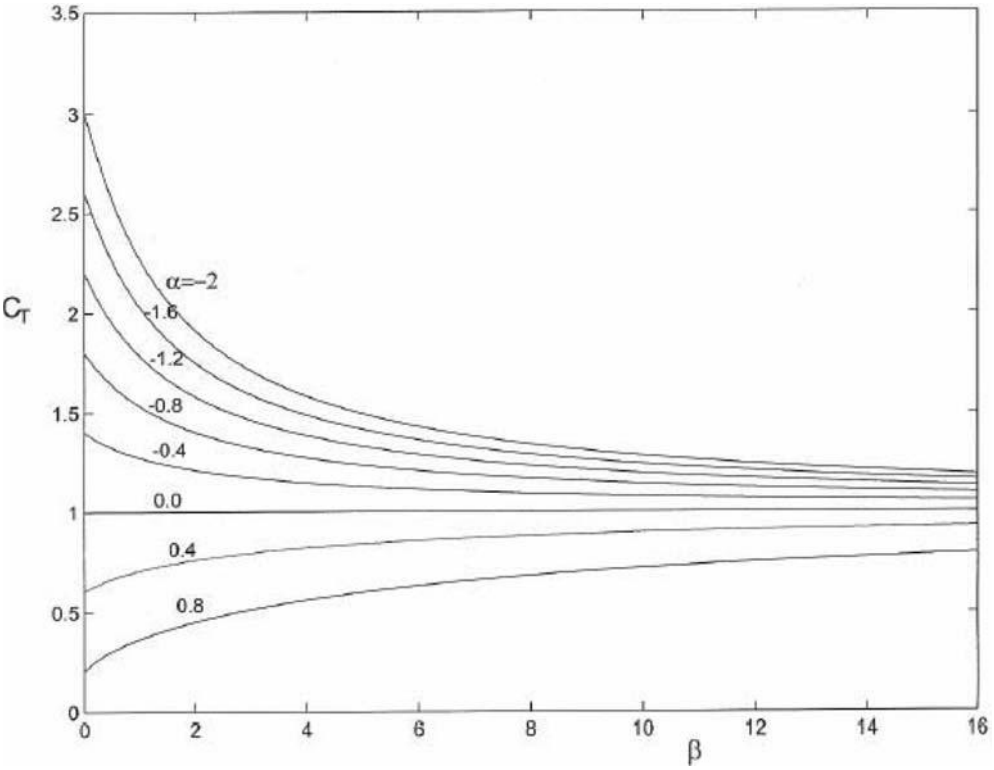


FIG. 1. *The dimensionless torque* $C_T = T_o/(-8\pi\mu_0 a^3\omega)$ *exerted on a spherical particle immersed in an unbounded quiescent fluid with an exponentially varying viscosity* $\mu_\infty = \mu_0[1 - \alpha e^{-\beta(r-a)/a}](\alpha < 1)$.

Case A: A translating particle in a weakly varying viscosity field. In case the viscosity field varies slowly with position, the viscosity field depends upon the "slow" variable:

$$\zeta = \varepsilon z, \tag{40}$$

where $\varepsilon < 1$ is a smallness parameter. Thus,

$$\lambda = \lambda(\zeta), \qquad \lambda_z = \varepsilon\lambda_\zeta. \tag{41}$$

The multiple-scale approximate method (Bender and Orszag (1999)) is applied to solve (30). Substitution of (40) and (41) into (30) and (31) and treating z and ζ as independent variables yields the following partial differential equations for G and h:

$$G_{zz} - G_z - 6G + \varepsilon[2G_{z\zeta} - G_\zeta + (\lambda_\zeta/\lambda)G_z] + \varepsilon^2[G_{\zeta\zeta} + (\lambda_\zeta/\lambda)G_\zeta] = C_D/\lambda, \tag{42a}$$

$$h_z + e^{-z}(G + G_z) + \varepsilon[h_\zeta + e^{-z}G_\zeta] = 0. \tag{42b}$$

The solutions for G, h, and the constant C_D can be approximately represented by a power series in ε:

$$G = G^{(0)} + \varepsilon G^{(1)} + \varepsilon^2 G^{(2)} + \cdots, \tag{43a}$$

$$h = h^{(0)} + \varepsilon h^{(1)} + \varepsilon^2 h^{(2)} + \cdots, \tag{43b}$$

$$C_D = C_D^{(0)} + \varepsilon C_D^{(1)} + \varepsilon^2 C_D^{(2)} + \cdots. \tag{43c}$$

Substituting (43) into (42) and collecting terms of zero, first, and second order in ε yields the following differential equations and boundary conditions for $G^{(i)}$ and $h^{(i)}$ $(i = 0, 1, 2)$:

For the zeroth order:

$$G^{(0)}_{zz} - G^{(0)}_{z} - 6G^{(0)} = C^{(0)}_D/\lambda(\zeta), \tag{44a}$$

$$h^{(0)}_z = -e^{-z}(G^{(0)} + G^{(0)}_z), \tag{44b}$$

$$[G^{(0)}]_{z=0} = 0, \quad [h^{(0)}]_{z=0} = 1, \quad [e^{-z}G^{(0)}]_{z\to\infty} \to 0, \quad [h^{(0)}]_{z\to\infty} \to 0. \tag{44c}$$

For the first order:

$$G^{(1)}_{zz} - G^{(1)}_{z} - 6G^{(1)} = -2G^{(0)}_{z\zeta} + G^{(0)}_{\zeta} - (\lambda_\zeta/\lambda)G^{(0)}_z + C^{(1)}_D/\lambda(\zeta), \tag{45a}$$

$$h^{(1)}_z = -e^{-z}(G^{(1)} + G^{(1)}_z) - h^{(0)}_\zeta - e^{-z}G^{(0)}_\zeta, \tag{45b}$$

$$[G^{(1)}]_{z=0} = 0, \quad [h^{(1)}]_{z=0} = 0, \quad [e^{-z}G^{(1)}]_{z\to\infty} \to 0, \quad [h^{(1)}]_{z\to\infty} \to 0. \tag{45c}$$

For the second order:

$$\begin{aligned} G^{(2)}_{zz} - G^{(2)}_{z} - 6G^{(2)} = {} & -2G^{(1)}_{z\zeta} + G^{(1)}_{\zeta} - (\lambda_\zeta/\lambda)G^{(1)}_z \\ & - G^{(0)}_{\zeta\zeta} - (\lambda_\zeta/\lambda)G^{(0)}_\zeta + C^{(2)}_D/\lambda(\zeta), \end{aligned} \tag{46a}$$

$$h^{(2)}_z = -e^{-z}(G^{(2)} + G^{(2)}_z) - h^{(1)}_\zeta - e^{-z}G^{(1)}_\zeta, \tag{46b}$$

$$[G^{(2)}]_{z=0} = 0, \quad [h^{(2)}]_{z=0} = 0, \quad [e^{-z}G^{(2)}]_{z\to\infty} \to 0, \quad [h^{(2)}]_{z\to\infty} \to 0. \tag{46c}$$

Further, drag force exerted on the particle is given by

$$\mathbf{F} = \frac{4}{3}\pi\mu_0 a\mathbf{U}(C^{(0)}_D + \varepsilon C^{(1)}_D + \varepsilon^2 C^{(2)}_D + \cdots). \tag{47}$$

The zeroth order solution. The general solution of (44a) is

$$G^{(0)}(z,\zeta) = A^{(0)}(\zeta)e^{3z} + B^{(0)}(\zeta)e^{-2z} - \frac{C^{(0)}_D}{6\lambda(\zeta)}, \tag{48}$$

where $A^{(0)}(\zeta)$ and $B^{(0)}(\zeta)$ are yet undetermined functions of ζ. Application of boundary conditions (44c) results in

$$\begin{aligned} & A^{(0)}(\zeta) = 0, \\ & B^{(0)}(\zeta = 0) - \frac{C^{(0)}_D}{6\lambda_S} = 0, \end{aligned} \tag{49}$$

where we define

$$\lambda_S = \lambda(\zeta = 0) = \lambda(r = a). \tag{50}$$

From (44b), (48), and (50), the solution for $h^{(0)}$ is

$$h^{(0)} = -\frac{1}{3}B^{(0)}(\zeta)e^{-3z} - \frac{C_D^{(0)}}{6\lambda(\zeta)}e^{-z}, \tag{51}$$

where from (44c),

$$-\frac{1}{3}B^{(0)}(\zeta=0) - \frac{C_D^{(0)}}{6\lambda_S} = 1. \tag{52}$$

Thus, from (50) and (52),

$$\begin{aligned} C_D^{(0)} &= -\frac{9}{2}\lambda_S, \\ B^{(0)}(\zeta=0) &= -\frac{3}{4}. \end{aligned} \tag{53a, b}$$

Consequently, from (47b) and (53a), the zeroth order approximation for the drag force exerted on a spherical particle translating in a quiescent fluid with a weakly radially varying viscosity is

$$\mathbf{F}^{(0)} = \frac{4}{3}\pi\mu_0 a\mathbf{U}C_D^{(0)} = -6\pi a\mathbf{U}\mu(r=a). \tag{54}$$

Equation (54) is the well-known Stokes law; only here, to a leading order, the particle senses the viscosity field near its surface.

It should be noted that an explicit expression for the function $B^{(0)}(\zeta)$ is not required to determine the zeroth order expression for the drag force. It must, however, be solved to obtain higher order corrections. Common to multiscale analyses, $B^{(0)}(\zeta)$ can be determined by demanding that secular terms in higher order approximations vanish. Upon substitution of (48) and (50) into (45a), the following differential equation for $G^{(1)}$ is obtained:

$$G_{zz}^{(1)} - G_z^{(1)} - 6G^{(1)} = \left(5B_\zeta^{(0)} + 2\frac{\lambda_\zeta}{\lambda}B^{(0)}\right)e^{-2z} + C_D^{(0)}\frac{\lambda_\zeta}{6\lambda^2} + \frac{C_D^{(1)}}{\lambda}. \tag{55}$$

To eliminate secular terms from the solution of $G^{(1)}$ we need that

$$5B_\zeta^{(0)} + 2\frac{\lambda_\zeta}{\lambda}B^{(0)} = 0. \tag{56}$$

A solution of (56) satisfying boundary condition (53b) is

$$B^{(0)} = -\frac{3}{4}\left(\frac{\lambda_S}{\lambda(\zeta)}\right)^{2/5}. \tag{57}$$

Consequently, the zeroth order solutions for $G^{(0)}$ and $h^{(0)}$ are

$$G^{(0)} = -\frac{3}{4}\left(\frac{\lambda_S}{\lambda(\zeta)}\right)^{2/5}e^{-2z} + \frac{3}{4}\frac{\lambda_S}{\lambda(\zeta)}, \tag{58a}$$

$$h^{(0)} = \frac{1}{4}\left(\frac{\lambda_S}{\lambda(\zeta)}\right)^{2/5}e^{-3z} + \frac{3}{4}\frac{\lambda_S}{\lambda(\zeta)}e^{-z}. \tag{58b}$$

The first order correction. A general solution of (55) subjected to (56) is

$$G^{(1)} = A^{(1)}(\zeta)e^{3z} + B^{(1)}(\zeta)e^{-2z} - \frac{C_D^{(0)}}{36}\frac{\lambda_\zeta}{\lambda^2} - \frac{C_D^{(1)}}{6\lambda}, \tag{59}$$

where $A^{(1)}$ and $B^{(1)}$ are yet undetermined functions of ζ. Application of boundary conditions (45c) results in

$$\begin{aligned} &A^{(1)}(\zeta) = 0, \\ &B^{(1)}(\zeta = 0) - \frac{C_D^{(1)}}{6\lambda_S} = -\frac{(\lambda_\zeta)_S}{8\lambda_S}, \end{aligned} \tag{60a, b}$$

where

$$(\lambda_\zeta)_S = \left[\frac{\partial\lambda}{\partial\zeta}\right]_{\zeta=0} = \frac{a}{\varepsilon}\left[\frac{\partial\lambda}{\partial r}\right]_{r=a}. \tag{61}$$

From (45b), (58), and (59), the solution for $h^{(1)}$ is

$$h^{(1)} = \frac{1}{15}\frac{\lambda_\zeta}{\lambda}\left(\frac{\lambda_S}{\lambda}\right)^{2/5}e^{-3z} - \left(\frac{C_D^{(1)}}{6\lambda} + \frac{11}{8}\frac{\lambda_\zeta\lambda_S}{\lambda^2}\right)e^{-z} - \frac{B^{(1)}(\zeta)}{3}e^{-3z}. \tag{62}$$

Hence, from (45c),

$$\frac{B^{(1)}(\zeta = 0)}{3} + \frac{C_D^{(1)}}{6\lambda_S} = -\frac{157}{120}\frac{(\lambda_\zeta)_S}{\lambda_S}. \tag{63}$$

Thus, from (60b) and (63),

$$C_D^{(1)} = -\frac{57}{10}(\lambda_\zeta)_S, \qquad B^{(1)}(\zeta = 0) = -\frac{43}{40}\frac{(\lambda_\zeta)_S}{\lambda_S}. \tag{64a, b}$$

Consequently, from (47b) and (53a), an approximate expression for the drag force exerted on a spherical particle is

$$\begin{aligned} \mathbf{F} &= \frac{4}{3}\pi\mu_0 a\mathbf{U}(C_D^{(0)} + \varepsilon C_D^{(1)}) = -6\pi a\mathbf{U}\mu(r = a)\left\{1 + \varepsilon\frac{19}{15}\frac{(\lambda_\zeta)_S}{\lambda_S} + O(\varepsilon^2)\right\} \\ &= -6\pi a\mathbf{U}\mu(r = a)\left\{1 + \frac{19}{15}a\left[\frac{1}{\mu}\frac{\partial\mu}{\partial r}\right]_{r=a} + O(\varepsilon^2)\right\}. \end{aligned} \tag{65}$$

To obtain higher order corrections a solution for $B^{(1)}(\zeta)$ must be provided. The differential equation governing $B^{(1)}(\zeta)$ is obtained by demanding that secular terms in the differential equation for $B^{(2)}(\zeta)$ vanish.

The latter is derived upon substitution of (59) and (60a) into (46a):

$$\begin{aligned} G_{zz}^{(2)} - G_z^{(2)} - 6G^{(2)} = e^{-2z}&\left(5B_\zeta^{(1)} + 2\frac{\lambda_\zeta}{\lambda}B^{(1)} - B_{\zeta\zeta}^{(0)} - \frac{\lambda_\zeta}{\lambda}B_\zeta^{(0)}\right) \\ &+ C_D^{(0)}\left(\frac{5}{18}\frac{\lambda_\zeta^2}{\lambda^3} - \frac{7}{36}\frac{\lambda_{\zeta\zeta}}{\lambda^2}\right) + C_D^{(1)}\frac{\lambda_\zeta}{6\lambda^2} + C_D^{(2)}\frac{1}{\lambda}. \end{aligned} \tag{66}$$

Thus, the governing differential equation for $B^{(1)}(\zeta)$ is

$$5B^{(1)}_{\zeta}+2\frac{\lambda_\zeta}{\lambda}B^{(1)}=B^{(0)}_{\zeta\zeta}+\frac{\lambda_\zeta}{\lambda}B^{(0)}_{\zeta}. \tag{67}$$

Subjected to boundary conditions (60b) and (46c), the solutions for $B^{(1)}$ and $h^{(1)}$ are easily obtained:

$$B^{(1)}=\frac{3}{50}\left(\frac{\lambda_S}{\lambda}\right)^{2/5}\int_0^\zeta \lambda^{-3/5}\frac{d(\lambda^{-2/5}\lambda_\zeta)}{d\zeta}d\zeta-\frac{43}{40}\frac{(\lambda_\zeta)_S}{\lambda_S^{3/5}\lambda^{2/5}}, \tag{68}$$

$$\begin{aligned} h^{(1)}=&\left(\frac{57}{60}\frac{(\lambda_\zeta)_S}{\lambda}-\frac{11}{8}\frac{\lambda_\zeta\lambda_S}{\lambda^2}\right)e^{-z}\\ &-\left[\frac{1}{50}\left(\frac{\lambda_S}{\lambda}\right)^{2/5}\int_0^\zeta \lambda^{-3/5}\frac{d(\lambda^{-2/5}\lambda_\zeta)}{d\zeta}d\zeta\right.\\ &\left.-\frac{43}{120}\frac{(\lambda_\zeta)_S}{\lambda_S^{3/5}\lambda^{2/5}}-\frac{1}{15}\frac{\lambda_\zeta}{\lambda}\left(\frac{\lambda_S}{\lambda}\right)^{2/5}\right]e^{-3z}. \end{aligned} \tag{69}$$

From (66) and (67), the general solutions for $G^{(2)}$, $h^{(2)}$ that satisfy the boundary conditions at an infinite distance from the sphere are

$$G^{(2)}=B^{(2)}(\zeta)e^{-2z}+\frac{5}{24}\frac{\lambda_S\lambda_\zeta^2}{\lambda^3}-\frac{7}{48}\frac{\lambda_S\lambda_{\zeta\zeta}}{\lambda^2}+\frac{19}{120}\frac{(\lambda_\zeta)_S\lambda_\zeta}{\lambda^2}-C_D^{(2)}\frac{1}{6\lambda}, \tag{70}$$

$$\begin{aligned} h^{(2)}=&e^{-3z}\left\{-\frac{43}{900}\left(\frac{\lambda}{\lambda_S}\right)^{3/5}\frac{(\lambda_\zeta)_S\lambda_\zeta}{\lambda^2}-\frac{7}{225}\left(\frac{\lambda_S}{\lambda}\right)^{2/5}\frac{\lambda_\zeta^2}{\lambda^2}\right.\\ &+\frac{1}{45}\left(\frac{\lambda_S}{\lambda}\right)^{2/5}\frac{\lambda_{\zeta\zeta}}{\lambda}+\frac{1}{375}\left(\frac{\lambda_S}{\lambda}\right)^{2/5}\frac{\lambda_\zeta}{\lambda}\int_0^\zeta\lambda^{-3/5}\frac{d}{d\zeta}(\lambda^{-2/5}\lambda_\zeta)d\zeta\\ &\left.-\frac{1}{150}\left(\frac{\lambda_S}{\lambda}\right)^{2/5}\lambda^{-3/5}\frac{d}{d\zeta}(\lambda^{-2/5}\lambda_\zeta)+\frac{1}{3}\frac{\partial B^{(1)}}{\partial\zeta}-3B^{(2)}\right\}\\ &-e^{-z}\left\{\frac{67}{24}\frac{\lambda_S\lambda_\zeta^2}{\lambda^3}+\frac{67}{48}\frac{\lambda_S\lambda_{\zeta\zeta}}{\lambda^2}+\frac{627}{360}\frac{(\lambda_\zeta)_S\lambda_\zeta}{\lambda^2}\right\}. \end{aligned} \tag{71}$$

Application of boundary conditions (46c) yields

$$C_D^{(2)}=-\frac{49703}{2000}\frac{(\lambda_\zeta^2)_S}{\lambda_S}-\frac{2011}{400}(\lambda_{\zeta\zeta})_S. \tag{72}$$

Hence, from (47b), a second order correction for the drag force exerted on a spherical

particle translating in a fluid with radially symmetric viscosity is

(73)
$$\begin{aligned}\mathbf{F} &= \frac{4}{3}\pi\mu_0 a\mathbf{U}(C_D^{(0)} + \varepsilon C_D^{(1)} + \varepsilon^2 C_D^{(2)}) \\ &= -6\pi a\mathbf{U}\mu(r=a)\left\{1 + \varepsilon\frac{19}{15}\frac{(\lambda_\zeta)_S}{\lambda_S} + \varepsilon^2\left(\frac{49703}{9000}\frac{(\lambda_\zeta^2)_S}{\lambda_S^2} + \frac{2011}{1800}\frac{(\lambda_{\zeta\zeta})_S}{\lambda_S}\right) + O(\varepsilon^3)\right\} \\ &= -6\pi a\mathbf{U}\mu(r=a)\left\{1 + \frac{19}{15}a\left[\frac{1}{\mu}\frac{\partial\mu}{\partial r}\right]_{r=a} + \frac{49703}{9000}a^2\left(\left[\frac{1}{\mu}\frac{\partial\mu}{\partial r}\right]_{r=a}\right)^2\right. \\ &\qquad \left. + \frac{2011}{1800}a^2\left[\frac{1}{\mu}\frac{\partial^2\mu}{\partial r^2}\right]_{r=a} + O(\varepsilon^3)\right\}.\end{aligned}$$

Case B: A particle translating in an algebraically varying viscosity fields in r-space. An analytic solution can easily be obtained in case the viscosity field varies algebraically in r-space (equivalently, varies exponentially in z-space), namely,

$$\lambda(z) = \exp(\alpha z) \equiv y^\alpha, \tag{74}$$

where α is an arbitrary constant. Notice that for α positive, the viscosity field increases without bound at infinity or becomes rigid far from the spherical particle. For α negative, the viscosity field vanishes at infinity, or the fluid becomes ideal far from the particle.

Introduction of (74) into (30) yields a simple linear nonhomogeneous second order equation with constant coefficients,

$$G_{zz} - (1-\alpha)G_z - 6G = C_D\exp(-\alpha z). \tag{75}$$

The general solution of (75) that satisfies boundary conditions (30c, d, e) at $z \to \infty$ is

$$G = \frac{C_D}{\alpha - 6}\exp(-\alpha z) + C_1'\exp(-sz) \quad \text{for } \alpha \neq 6 \quad \text{and} \quad \alpha > -1 \tag{76a}$$

and

$$G = (C_1'' - zC_D/7)\exp(-6z) \quad \text{for } \alpha = 6, \tag{76b}$$

where C_D, C_1', and C_1'' are constants of integration and

$$s = 0.5[\alpha - 1 + \sqrt{(\alpha-1)^2 + 24}] \tag{77}$$

is positive for any value of α.

Notice that the condition $\alpha > -1$ stems from (30e), i.e., that the viscosity field must decay to zero no faster than $1/r$ to obtain a velocity and pressure fields that vanish at an infinite distance from the particle.

Substituting (76) into (31) and integrating yields the respective expressions for h,

$$h = C_D\frac{1-\alpha}{(1+\alpha)(\alpha-6)}\exp[-(1+\alpha)z] + C_1'\frac{1-s}{1+s}\exp[-(s+1)z] \quad \text{for } \alpha \neq 6 \tag{78a}$$

and

$$h = \left[-\frac{5C_1''}{7} + C_D\left(\frac{5z}{49} - \frac{2}{343}\right)\right]\exp(-7z) \quad \text{for } \alpha = 6. \tag{78b}$$

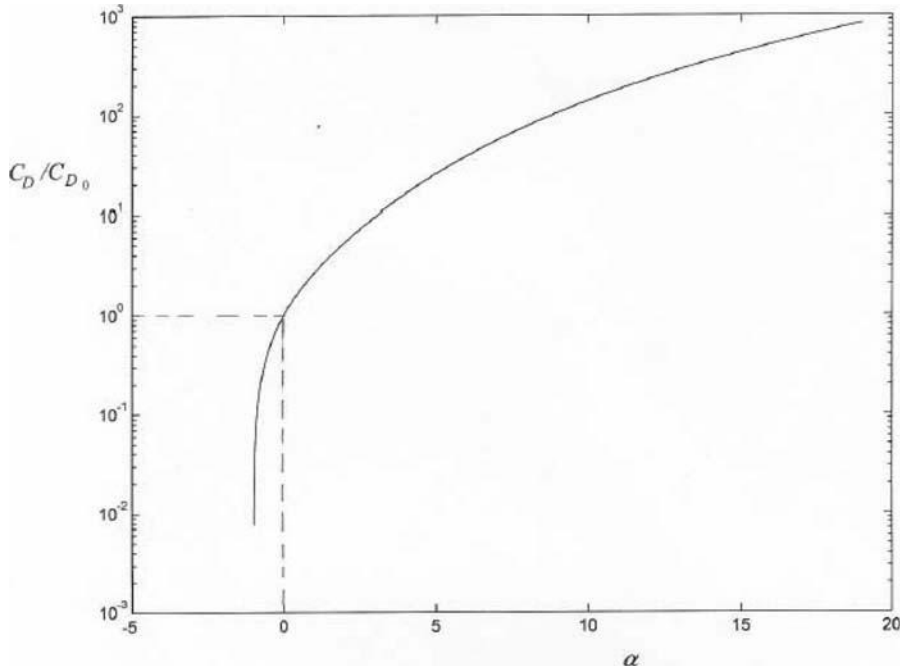


FIG. 2. *The dimensionless drag force* $C_D = F/(4\pi\mu_0 Ua/3)$ *exerted on a spherical particle immersed in an unbounded quiescent fluid with an algebraically varying viscosity* $\mu_\infty = \mu_0(r/a)^\alpha$ $(\alpha > -1)$ *and normalized with* $C_{D0} = -9/2$*, the corresponding dimensionless drag force exerted on a particle immersed in a fluid with uniform viscosity* $(\alpha = 0)$.

Utilizing boundary conditions (32a, b) makes it possible to determine the unknown coefficients C_1' and C_1'' and the drag coefficient C_D,

$$C_1' = -\frac{C_D}{(\alpha - 6)}, \qquad C_1'' = 0,$$
$$C_D = -\frac{(6-\alpha)(1+\alpha)(1+s)}{2(s-\alpha)} \quad \text{for } \alpha \neq 6 \text{ and } \alpha > -1 \tag{79a}$$

and

$$C_D = -\frac{343}{2} \quad \text{for } \alpha = 6. \tag{79b}$$

It is easy to show that C_D is negative for all values of $\alpha > -1$, and for $\alpha = 0$ (a uniform viscosity field) $C_D = C_{D0} = -9/2$ and Stokes law is recovered. Figure 2 illustrates the ratio between the drag coefficients for any viscosity parameter α and that of $\alpha = 0$. For α approaching -1 the drag coefficient decreases to zero. Obviously, for a very small fluid viscosity, the creeping flow approximation is no longer valid. Notwithstanding, the limit seems to approach the result of zero drag known from potential flow theory of ideal fluids.

Case C: A numerical algorithm. Equations (30), (31) and boundary conditions (32) constitute a boundary value problem where conditions must be satisfied simultaneously at $z = 0$ and at infinity. A numerical shooting scheme may encounter great difficulties despite the fact that the governing equations (30), (31) are linear. When

the asymptotic behavior of λ satisfies the relation $O(1) \leq \lambda \ll e^z$ as $z \to \infty$, the homogeneous part of (30) possesses a decaying (desired) and fast-growing (undesired) mode (see the appendix). Consequently, assuming the known values of G and h at $z = 0$ (see (32b, d)), an arbitrary value for the derivative of G at $z = 0$ is most likely to yield an exponentially divergent result, due to the overwhelming contribution of the undesired mode.

Nevertheless, a divergent solution (which may be obtained numerically from an arbitrary set of initial conditions for G and its derivative) can be utilized to construct convergent homogeneous and particular solutions of (30). Let us assume that $G_{H1}(z)$ is a *positive-definite* divergent solution of the homogeneous part of (30). Utilizing the method of variation of parameters, it is easy to show that a second independent homogeneous solution is given by

$$G_{H2}(z) = G_{H1}(z) \int_z^\infty \frac{e^x}{\lambda G_{H1}^2} dx \tag{80}$$

and that the particular solution of (30) with $C_D = 1$ is given by

$$G_P(z) = -G_{H1}(z) \int_z^\infty \frac{e^x}{\lambda G_{H1}^2} \left(\int_0^x G_{H1}(y) e^{-y} dy \right) dx. \tag{81}$$

By a simple application of l'Hopital's rule and accounting for the asymptotic behavior of G_{H1} at infinity (see the appendix), it is easy to show that G_{H2} vanishes at infinity and G_P approaches $-1/(6\lambda)$.

A general solution that satisfies (30) is given by

$$G = AG_{H1} + BG_{H2} + C_D G_P, \tag{82}$$

where A, B, and C_D are yet undetermined constants. However, A must vanish to satisfy the vanishing boundary condition at infinity. Thus, substituting (81) into (31) and integrating yields the general form of h,

(83)

$$\begin{aligned} h(z) = &\int_{w=z}^\infty (G_{H1} + G'_{H1}) e^{-w} \left[B \int_w^\infty \frac{e^x dx}{\lambda G_{H1}^2} - C_D \int_w^\infty \frac{e^x}{\lambda G_{H1}^2} \left(\int_0^x G_{H1} e^{-y} dy \right) dx \right] dw \\ &- \int_{x=z}^\infty \frac{1}{\lambda G_{H1}} \left[B - C_D \int_0^x G_{H1} e^{-y} dy \right] dx. \end{aligned}$$

Substituting boundary conditions (32b, d) into (82) and (83) yields two linear equations for the two unknowns B and C_D, which can readily be solved:

$$B \int_0^\infty \frac{e^x dx}{\lambda G_{H1}^2} - C_D \int_0^\infty \frac{e^x}{\lambda G_{H1}^2} \left(\int_0^x G_{H1} e^{-y} dy \right) dx = 0, \tag{84a}$$

$$\begin{aligned} &B \left\{ \int_0^\infty (G_{H1} + G'_{H1}) e^{-z} \left(\int_z^\infty \frac{e^x dx}{\lambda G_{H1}^2} \right) dz - \int_0^\infty \frac{dz}{\lambda G_{H1}} \right\} \\ &- C_D \left\{ \int_0^\infty (G_{H1} + G'_{H1}) e^{-z} \left[\int_z^\infty \frac{e^x}{\lambda G_{H1}^2} \left(\int_0^x G_{H1} e^{-y} dy \right) dx \right] dz \right. \\ &\left. - \int_0^\infty \frac{1}{\lambda G_{H1}} \left(\int_0^z G_{H1} e^{-y} dy \right) dz \right\} = 1. \end{aligned} \tag{84b}$$

Notice that we assume that both G_{H1} and its derivative G'_{H1} are known. Normally, this would *not* require a cumbersome numerical differentiation of G_{H1}. Indeed, the second order equation (30) would normally be replaced by a system of two first order equations for G_{H1} and G'_{H1}, so that a numerical solution for these functions would be obtained simultaneously.

Several terms in (84) include integration over an infinite domain, an operation that seems to pose some difficulty. However, due to the exponential behavior of G_{H1}, integration can be carried out over a finite domain without a significant loss of accuracy. A difficult case that involves a periodically varying viscosity is attempted in the next chapter.

5. Homogeneous suspensions. An intriguing example is the case in which the spherical particle rotates and translates in a spatially periodic viscosity, namely,

$$\lambda(r) = 1 - \alpha \cos\left[\frac{\beta(r-a)}{a}\right], \tag{85}$$

where α and β are known dimensionless parameters and a is the radius of the "test" particle that moves in the field.

Equation (85) can be loosely interpreted as the local viscosity of a homogeneous suspension. Frequently, a suspension is perceived as a single phase fluid with a homogeneous uniform effective viscosity that depends upon the viscosities of the constituent fluids and their volumetric fraction. For instance, the effective viscosity of a dilute suspension of droplets is $\mu_0 = \mu\big[1 + \big(\frac{1+2.5\mu_i/\mu}{1+\mu_i/\mu}\big)\phi\big]$, where μ and μ_i are the continuous fluid and droplet viscosities, respectively, and ϕ is the volumetric concentration of the droplets (Taylor (1932)). It is, however, plausible to view a suspension as a nonhomogeneous single phase fluid that possesses periodical fluctuations in the viscosity field. In this case, (85) represents a truncated Fourier series of the nonuniform viscosity field. If the effective mean viscosity of the suspension is employed as the characteristic viscosity μ_0, (e.g., $\mu_0 = \mu(1+2.5\phi)$ for a dilute suspension of rigid particles), then the parameter α stands for the amplitude of the first harmonic in a Fourier series expansion of the fluctuating viscosity field. We suggest that α is proportional to the viscosity difference $(\mu_0 - \mu)/\mu_0$. Thus, for a dilute suspension of rigid spheres, α is proportional to

$$\alpha \propto \frac{2.5\phi}{(1+2.5\phi)}. \tag{86}$$

Notice that, according to (85), the viscosity field near the rotating sphere is equal to the viscosity of the suspending fluid.

It is also plausible to assume that the length-scale of viscosity variations is determined by the distance between particles constituting the suspension. Thus, the parameter β is proportional to the reciprocal of the mean distance between particles,

$$\beta \propto \frac{a}{b\phi^{-1/3}}, \tag{87}$$

where b is the mean radius of a particle in the suspension. A positive α value pertains to a dense suspension of rigid particles, while a negative value pertains to a highly dense emulsion consisting of droplets (or bubbles) with viscosity lower than that of the surrounding continuous fluid. A large β value pertains to a large test particle $(a/b \gg 1)$ moving inside a dense suspension of smaller particles.

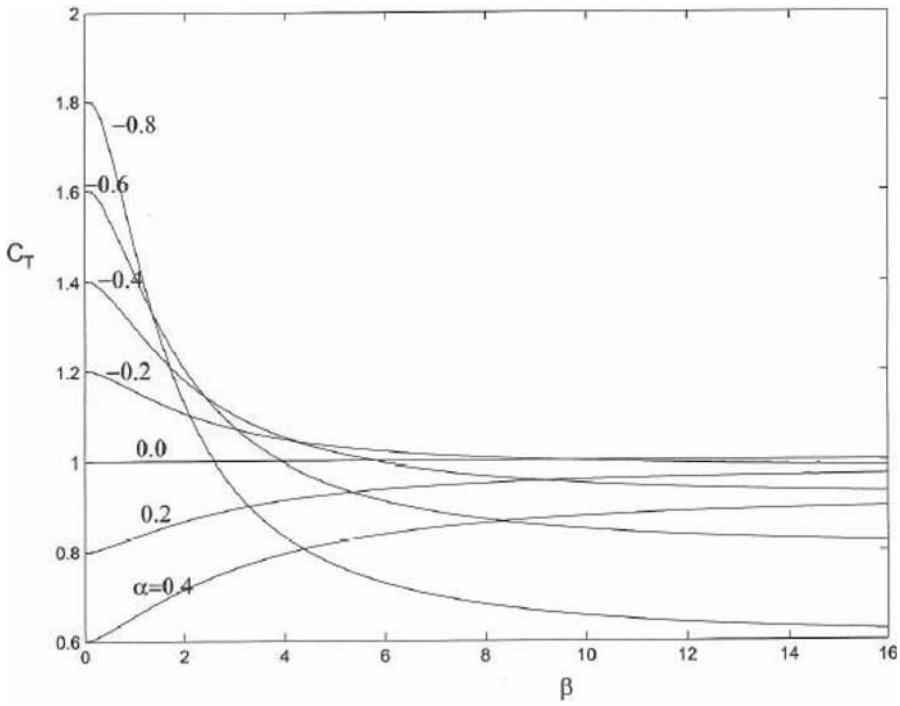


FIG. 3. *The dimensionless torque* $C_T = T_o/(-8\pi\mu_0 a^3\omega)$ *exerted on a spherical particle immersed in an unbounded quiescent fluid with a periodically varying viscosity* $\mu_\infty = \mu_0\{1-\alpha\cos[\beta(r-a)/a]\}$ $(|\alpha| < 1)$.

5.1. A particle rotating inside a suspension. Substitution of (85) into (17) yields the correction factor C_T depicted in Figure 3. For very small β's, the particles constituting the suspension are far from the rotating particle and have almost no effect on the moment exerted on the particle. Thus, the effective viscosity that the rotating particle senses is that of the fluid ($C_T \sim 1-\alpha$). However, for large β's (the mean distance between the suspended particles is smaller than the radius of the rotating test particle) the effective viscosity asymptotically approaches the value of the effective viscosity μ_0, irrespective of the sign of α. The latter fact coincides with known data. (See, for instance, Almog and Brenner (1998) for a rotating test particle in a dilute suspension.) In addition, it demonstrates that the results are insensitive to the exact location of the particle relative to the particles comprising the suspension (say, you replace the cosine with a sine function in (85)). For dilute suspensions or emulsions where $|\alpha| < 0.2$ and $\beta > 16$, the effective viscosity is higher than 0.98 μ_0. However, for high α values the solutions do not seem to reach the asymptotic value of μ_0. This is possibly due to the fact that the Fourier series (85) that contains only a first harmonic term is an insufficient representation of such strongly fluctuating viscosity fields.

5.2. A particle translating inside a dilute suspension. When α and β in (85) are small (which pertains to the case in which a small particle is embedded in a dilute suspension consisting of large particles), substitution of (85) into (73) yields

$$\mathbf{F} = -6\pi a\mathbf{U}\mu_0\left(1-\alpha+\frac{2011}{1800}\alpha\beta^2+O(\beta^3)\right). \tag{88a}$$

Thus, in such a case, a particle senses mainly the *fluid* viscosity with a second order correction in β. Notice, however, that if a *sine* distribution had been assumed in (85), the drag force would have been different,

$$\mathbf{F} = -6\pi a \mathbf{U}\mu_0 \left(1 - \frac{19}{15}\alpha\beta + \frac{49703}{9000}\alpha^2\beta^2 + O(\beta^3)\right), \tag{88b}$$

and the particle senses the *suspension* viscosity to a leading order in β. A possible explanation is that in the latter case the location of the small test particle is implicitly closer to the large particles comprising the suspension, increasing the drag force. This conclusion has also been noted by Almog and Brenner (1997), who showed in their two-sphere model that a small test particle would be strongly affected by a suspended particle adjacent to it, and as a result would spend a longer time traveling close to it.

We have shown in the Introduction that when a small particle moves inside a suspension its effect on the suspension configuration is negligible, provided that the Péclet number is smaller than unity. For a homogeneous suspension we hypothesize that the viscosity field retains its centrally symmetric configuration for all times, with the test particle always keeping its central position. Thus, if one desires to obtain the mean velocity of the test particle under a fixed external force, say gravity, or the mean force exerted on the test sphere, given that it moves with a fixed velocity, we suggest the following method: Instead of following the particle as it moves in the suspension, one may average over all realizations of the particle position with respect to the suspended particles. The viscosity field is centrally symmetric and is comprised of both sine and cosine radial distributions, accounting for a general location of the test particle with respect to the suspended particles, namely,

$$\frac{\mu}{\mu_0} = \lambda(r) = 1 - \alpha\cos\left[\frac{\beta(r-a)}{a} + \delta\right], \tag{89}$$

where $0 < \delta < 2\pi$ is an arbitrary phase angle with probability density $p(\delta)$. Implicitly, $\delta = 0$ pertains to the case in which the test particle is far from the suspended particles, while $\delta = \pi$ relates to a test particle placed near the suspended particles. Substituting (89) into (73) yields

$$\mathbf{F} = -6\pi a \mathbf{U}\mu_0 \Bigg(1 - \alpha\cos\delta + \frac{19}{15}\alpha\beta\sin\delta + \frac{2011}{1800}\alpha\beta^2\cos\delta + \frac{49703}{9000}\alpha^2\beta^2\frac{\sin^2\delta}{1-\cos\delta} + O(\beta^3)\Bigg).$$

Consequently, a small test particle of density ρ_p would move with a *mean* velocity

$$\bar{\mathbf{U}} = \int_0^{2\pi} \mathbf{U}p(\delta)d\delta = \frac{2(\rho_p - \rho)\mathbf{g}}{9\mu_0\sqrt{1-\alpha^2}}\left[1 - \frac{55373}{18000}\frac{\alpha^2\beta^2}{1-\alpha^2} + O(\beta^3)\right] \tag{90}$$

under gravity g in a suspension of density ρ, assuming that all realizations with respect to δ are equally probable (i.e., $p(\delta) = 1/2\pi$). Thus, the apparent viscosity is

$$\mu_{app.}(\text{constant force}) = \frac{\mu_0\sqrt{1-\alpha^2}}{\left[1 - \frac{55373}{18000}\frac{\alpha^2\beta^2}{1-\alpha^2} + O(\beta^3)\right]}. \tag{91}$$

The result (91) demonstrates that the sedimentation velocity of the test particle under gravity is practically determined by the suspension viscosity μ_0, and that *first*

order corrections in α and β vanish identically. Obviously, this result is based on the assumption that all realizations with respect to δ are equally probable. How good is that assumption? One may write the Fokker–Planck convection-diffusion equation for p, an approach used previously by Batchelor (1972) and others for a dilute suspension. However, if the Péclet number based on the diameter of the test particle is smaller than unity, the diffusion part of the Fokker–Planck equation dominates the probability density distribution, and the assumption that $p(\delta)$ is uniformly distributed is plausible. Surprisingly, Miliken et al. (1989), who performed experiments with large falling spheres of different diameters under gravity in a suspension comprised of neutrally buoyant spheres, obtained results that concur with our conclusions despite the fact that in those experiments the Péclet number was not small. They observed that only for dense concentrations ($\phi = 0.5, 0.55$) is there a noticeable effect of the ratio between test and suspension sphere diameters.

In the case when U is forced to be constant, one could calculate the mean force exerted on the particle to obtain the apparent viscosity. Assuming as before that all realizations with respect to δ are equally probable, we obtain

$$\bar{\mathbf{F}} = -6\pi a \mathbf{U} \mu_0 \left(1 + \frac{49703}{4500}\pi\alpha^2\beta^2 + O(\beta^3)\right), \tag{92}$$

and the apparent viscosity is given by

$$\mu_{app.}(\text{constant velocity}) = \mu_0 \left(1 + \frac{49703}{4500}\pi\alpha^2\beta^2 + O(\beta^3)\right). \tag{93}$$

Notice that the apparent viscosities in the foregoing two different cases (91) and (93) are identical up to first order terms in α and β. Moreover, these leading terms are independent of the ratio between the radii of suspension particles and test particles (as long as β is much smaller than unity). Thus, the apparent viscosities up to order ϕ are identical, a very gratifying result that implies that the viscosity μ_0 can be viewed as an intrinsic property of the suspension. Only second order terms in (91) and (93) appear to be different. However, these terms, which are of order $\alpha^2 \sim \phi^2$ for a dilute suspension, might be of limited value, since the expression for the viscosity field was assumed to retain only first order terms of the Fourier expansion.

Mondy, Graham, and Jensen (1986), who measured the velocity of a test sphere falling under gravity in a suspension of spherical rigid particles, also reached a similar conclusion. They used velocity measurements to calculate the apparent viscosity of the suspension (and wall effects induced by the suspension container). They concluded that in the absence of wall effects, and for a wide range of test sphere diameters larger than the suspended spheres, the apparent viscosity of a suspension is practically independent of the ratio of the diameters of the test sphere and the suspended spheres. In addition, for dilute suspensions, this apparent viscosity is essentially identical to the shear viscosity obtained by Couette or parallel-plates rheometers (see Figure 6 in Mondy, Graham, and Jensen (1986)).

A deviation of order ϕ in the apparent viscosities for constant velocity and constant force cases was obtained by Almog and Brenner (1997) for a very dilute suspension. The difference in our results is likely to stem from the different configurations assumed for the suspension. In our case the suspension is essentially perceived as a fixed three-dimensional lattice arranged around the test particle, and its position within the lattice is equally probable, while Almog and Brenner (1997) assumed that only a single particle in the suspension affects the test particle at any instant, and that

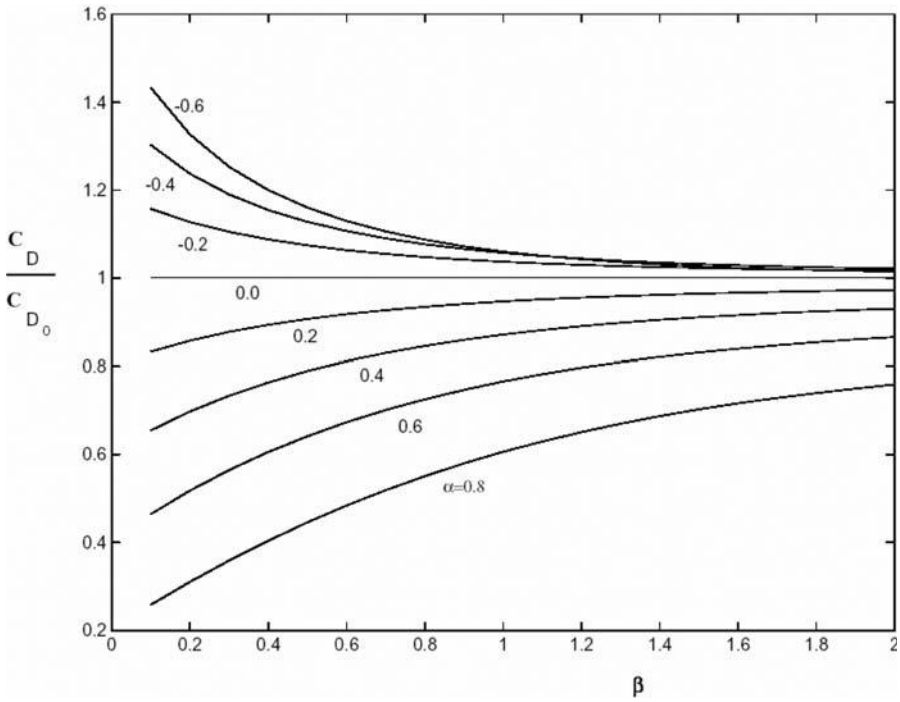


FIG. 4. *The dimensionless drag force* $C_D = F/(4\pi\mu_0 Ua/3)$ *exerted on a spherical particle immersed in an unbounded quiescent fluid with a periodically varying viscosity* $\mu_\infty = \mu_0\{1-\alpha\cos[\beta(r-a)/a]\}$ $(|\alpha| < 1)$ *and normalized with* $C_{D0} = -9/2$*, the corresponding dimensionless drag force exerted on a particle immersed in a fluid with uniform viscosity* $(\alpha = 0)$.

the probability density depends on the distance between the test and the suspended particles.

5.3. Computation of the drag force exerted on a particle translating inside a suspension. In this section we employ the numerical scheme suggested in section 4.2, case C, to obtain the drag force exerted on a rigid sphere translating in a flow field with periodically varying viscosity (89), for α and β not necessarily small and $\delta = 0$. Thus, only the drag force exerted on a particle positioned as far as possible from the suspended particles is considered.

Figure 4 depicts the numerical values obtained for the drag coefficient for a variety of α and β values. Evidently, in the case $\alpha = 0$ the viscosity field is uniform and the drag coefficient should be $C_D = C_{D0} = -4.5$, regardless of the value of β. Indeed, this known theoretical result was utilized to measure the accuracy of our numerical scheme. We observed that, replacing $z = \infty$ with $z = 8$ in (84), the error in C_{D0} was about 0.1%. For very small β's, the particles making up the suspension are far from the translating particle and have almost no effect on the force exerted on the particle. Thus, the effective viscosity that the translating particle senses is that of the fluid, and $C_D/C_{D0} \sim 1 - \alpha$. However, for larger β's (when the mean distance between the suspended particles is smaller than the radius of the translating test particle) the effective viscosity reaches asymptotically the value of the effective viscosity μ_0, irrespective of the sign of α. Comparison with known data can be done only qualitatively, since only one realization of the viscosity distribution according

to (89) was addressed. Miliken et al. (1989) examined experimentally the effect of falling test sphere diameter on the apparent viscosity of a dense suspension composed of spherical rigid particles. They observed (Figure 5 in their article) that for dense suspensions $\phi = (0.5, 0.55)$ the apparent viscosity deviates from its macroscopic average viscosity for test particles smaller than the suspended particles, while for larger test particles the apparent viscosity is practically independent of the ratio between the test and suspended particle radii. The value of β for $a/b = 1$ and $\phi = 0.5$ is 1.26. For this value of β, Figure 4 demonstrates that for $\alpha = 0.2$ the value of apparent viscosity deviates from the suspension viscosity by about 5%. For higher α values this deviation increases, possibly because the Fourier expansion of the viscosity field was limited to the first cosine term only. Nonetheless, for all α values the apparent viscosity asymptotically reaches the suspension viscosity for increasing values of β, a trend that was also observed by Miliken et al. (1989).

6. Summary. General solutions for the velocity and pressure fields were obtained for a rotating (section 2) and translating (section 3) spherical particle moving in an unbounded quiescent field with radially varying viscosity. An explicit formula, (17) was obtained for the torque exerted on a rotating particle. It manifests that only the near viscosity field affects the torque, whereas the far field is strongly screened. General analytical expressions (22) were obtained for the velocity and pressure fields, and a simplified expression (36) was obtained for the drag force exerted on a particle translating in an unbounded quiescent fluid. Three particular cases were addressed: In case A, weak radial variation of the viscosity field was assumed, and a second order approximate analytical solution (73) was derived for the drag force. In case B, the viscosity field increased or decreased algebraically, and an exact solution was obtained. It manifests that for a vanishing viscosity at infinity the drag force vanishes, a known result for the drag force exerted on a translating particle in ideal fluids. The latter result is quite surprising, in light of the fact that the Stokes equation for creeping flows is no longer a valid governing equation of the flow field. In case C, an effective numerical algorithm was suggested, by which the drag force can also be obtained for viscosity distributions that vary at a smaller than algebraic rate. In section 5 we addressed the difficult problem of a particle moving in a field with periodically varying viscosity. We hypothesized that such a viscosity field can be viewed as a continuous manifestation of a bimodal viscosity field of a suspension. For a rotating particle a screening effect of the near viscosity field was observed, a well-known fact for a particle rotating in a suspension. For a translating particle, the results demonstrate again that for large β's (i.e., the mean distance between the suspended particles is smaller than the radius of the rotating test particle) the effective viscosity asymptotically reaches the value of the shear suspension viscosity μ_0. For small values of β, namely when the test particle is much smaller than the suspended particles, the initial drag force exerted on the test particle is insensitive to the composition of the particles or droplets and "senses" the viscosity of the continuous liquid only, provided that it is positioned far from the suspended particles. However, if the particle is positioned with equal probability at an arbitrary distance from the suspended particles, the apparent viscosities of a dilute suspension, in case a constant external force is exerted on the test particle or in case it is forced to move with a constant velocity, are both equal to the shear viscosity up to first order in particle concentration.

Appendix. The asymptotic behavior of G at infinity under the asymptotic condition that $\lambda_z/\lambda \ll 1$ as $z \to \infty$ follows the general methods presented by Bender and Orszag (1999). We shall first investigate the asymptotic behavior of the homo-

geneous solution of (30). To this end, it is convenient to apply the transformation $w = 1/z$, rewrite (30) in terms of w, and investigate the asymptotic behavior of (30) near $w \to 0$.

Thus, (30) possesses the following form:

$$G_{ww} + \left(\frac{1}{w^2} + \frac{2}{w} + \frac{\lambda_w}{\lambda}\right) G_w - \frac{6G}{w^4} = 0. \tag{A1}$$

Since $w = 0$ is an irregular singular point, the controlling factor of G as $w \to 0$ is of the form of an exponential,

$$G \sim e^{S(w)}. \tag{A2}$$

Substitution of (A2) into the homogeneous part (A1) yields

$$S'' + S'^2 = -\left(\frac{1}{w^2} + \frac{2}{w} + \frac{\lambda_w}{\lambda}\right) S' + \frac{6}{w^4}. \tag{A3}$$

However, since $\lambda_w/\lambda \ll 1/w^2$ and $S'' \ll S'^2$ as $w \to 0$, the asymptotic differential equation that determines the controlling factor is

$$S'^2 \sim -\frac{S'}{w^2} + \frac{6}{w^4}.$$

Thus, $S'_{1,2} = 2/w^2, -3/w^2$ or $S_1 = -2/w$ and $S_2 = 3/w$. Thus, the homogeneous equation possesses convergent and divergent controlling factors near $z \to \infty$,

$$G \sim e^{3z} \quad \text{and} \quad G \sim e^{-2z}. \tag{A4}$$

A better leading behavior can be obtained by assuming that

$$S_1 = -\frac{2}{w} + C(w), \tag{A5}$$

given that $C \ll -2/w$ as $w \to 0$. Substituting (A5) into (A3) and collecting leading order terms yields

$$C' \sim -\frac{2}{5}\frac{\lambda_w}{\lambda} \quad \text{or} \quad C \sim \ln\lambda^{-2/5}. \tag{A6}$$

Hence

$$G_1 \sim \lambda^{-2/5} e^{-2z} \quad \text{as } z \to \infty, \tag{A7}$$

and similarly,

$$G_2 \sim \lambda^{-3/5} e^{3z} \quad \text{as } z \to \infty. \tag{A8}$$

Not surprisingly, this local behavior was also revealed in case A, which addressed the global behavior of G for all values of z.

The general homogeneous solution is given by

$$G_H = C_1 G_1 + C_2 G_2,$$

where C_1 and C_2 are arbitrary constants. Obviously, unless C_2 is identically zero, the homogeneous solution diverges for large values of z.

The asymptotic local behavior of the particular solution at $z \to \infty$ is easily derived by the method of dominant balance. The dominant term that balances the forcing term in (30) is the last term on the LHS of the equation. Thus,

$$G \sim -\frac{C_D}{6\lambda} \tag{A9}$$

and $G' \ll C_D/\lambda$ and $G'' \ll C_D/\lambda$, since $\lambda_z/\lambda \ll 1$ as $z \to \infty$. Equation (A9) is also the leading term in (62) of case A.

REFERENCES

A. Acrivos, R. Mauri, and X. Fan (1993), *Shear induced resuspension in a Couette device*, Int. J. Multiphase Flow, 19, pp. 797–802.

P. M. Adler, M. Zuzovski, and H. Brenner (1985), *Spatially periodic suspensions of convex particles in linear shear flows. II. Rheology*, Int. J. Multiphase Flow, 11, pp. 387–417.

Y. Almog and H. Brenner (1997), *Noncontinuum anomalies in the apparent viscosity experienced by a test sphere moving through an otherwise quiescent suspension*, Phys. Fluids, 9, pp. 16–22.

Y. Almog and H. Brenner (1998), *Apparent slip at the surface of a small rotating sphere in a dilute quiescent suspension*, Phys. Fluids, 10, pp. 750–752.

G. K. Batchelor (1972), *Sedimentation in a dilute suspension of spheres*, J. Fluid Mech., 52, pp. 245–268.

G. K. Batchelor and J. T. Green (1972a), *The hydrodynamic interaction of two small freely moving spheres in a linear flow field*, J. Fluid Mech., 56, pp. 375–400.

G. K. Batchelor and J. T. Green (1972b), *The determination of the bulk stress in a suspension of spherical particles to order c^2*, J. Fluid Mech., 56, pp. 401–427.

C. M. Bender and S. A. Orzag (1999), *Advanced Mathematical Methods for Scientists and Engineers*, Springer, New York.

J. F. Brady (1988), *Stokesian dynamics*, Ann. Rev. Fluid Mech., 20, pp. 111–157.

H. Brenner, A. L. Graham, J. R. Abbott, and L. A. Mondy (1990), *Theoretical basis for falling-ball rheometry in suspensions of neutrally buoyant spheres*, Int. J. Multiphase Flow, 16, pp. 579–596.

H. Brenner (1979), *Taylor dispersion in systems of sedimenting nonspherical Brownian particles*, J. Colloid Interface Sci., 71, pp. 189–208.

C. Chang and R. L. Powell (1993), *Dynamic simulation of bimodal suspensions of hydrodynamically interacting spherical particles*, J. Fluid Mech., 253, pp. 1–25.

F. Gadala-Maria and A. Acrivos (1980), *Shear induced structure in concentrated suspensions of solid spheres*, J. Rheol., 24, pp. 799–814.

K. Gurbebeck and W. Sprossig (1993), *Hypercomplex function theory for non linear Stokes problems with variable viscosity*, Complex Variables Theory Appl., 22, pp. 195–202.

S. Haber, H. Brenner, and M. Shapira (1990), *Diffusion, sedimentation and Taylor dispersion of a Brownian cluster subjected to a time periodic external force: A micromodel of AC electrophoretic phenomena*, J. Chem. Phys., 92, pp. 5569–5579.

J. Happel and H. Brenner (1983), *Low Reynolds Number Hydrodynamics*, Nijhoff, The Hague, The Netherlands.

R. Hassonjee, P. Ganatos, and R. Pfeffer (1988), *A strong interaction theory for the motion of arbitrary three dimensional clusters of spherical particles at low Reynolds number*, J. Fluid Mech., 197, pp. 1–37.

D. Leighton and A. Acrivos (1987a), *Measurement of shear-induced self-diffusion in concentrated suspensions of spheres*, J. Fluid Mech., 177, pp. 109–131.

D. Leighton and A. Acrivos (1987b), *The shear induced migration of particles in concentrated suspensions*, J. Fluid Mech., 181, pp. 415–439.

W. J. Miliken, L. A. Mondy, M. Gottlieb, A. L. Graham, and R. L. Powell (1989), *The effect of the diameter of falling balls on the apparent viscosity of suspensions of spheres and rods*, PhysicoChem. Hydrodyn., 11, pp. 341–355.

L. A. Mondy, A. L. Graham, and J. Jensen (1986), *Continuum approximation and particle interaction in concentrated suspensions*, J. Rheol., 30, pp. 1031–1051.

P. R. Nott and J. F. Brady (1994), *Pressure-driven flow of suspensions: Simulation and theory*, J. Fluid Mech., 275, pp. 157–199.

R. J. Phillips, R. C. Armstrong, R. A. Brown, A. L. Graham, and J. R. Abbott (1992), *A constitutive equation for concentrated suspensions that accounts for shear-induced particle migration*, Phys. Fluids A, 4, pp. 30–40.

G. I. Taylor (1932), *The viscosity of a fluid containing small droplets of another fluid*, Proc. Roy. Soc. London A, 138, pp. 41–48.

W. Wang, R. Mauri, and A. Acrivos (1998), *Transverse shear induced gradient diffusion in a dilute suspension of spheres*, J. Fluid Mech., 357, pp. 279–287.

M. Zuzovsky, P. M. Adler, and H. Brenner (1983), *Spatially periodic suspensions of convex particles in linear shear flows.* III. *Dilute arrays of spheres suspended in Newtonian fluids*, Phys. Fluids, 26, pp. 1714–1723.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 305–320

RESCUE OF THE QUASI-STEADY-STATE APPROXIMATION IN A MODEL FOR OSCILLATIONS IN AN ENZYMATIC CASCADE*

THOMAS ERNEUX† AND ALBERT GOLDBETER‡

We wish to dedicate this paper to the memory of L. A. Segel who contributed to many aspects of mathematical biology and for whom the quasi-steady-state hypothesis was a favorite topic of research

Abstract. A three-variable model describing the oscillatory activity of a cascade of enzyme reactions is analyzed. A quasi-steady-state approximation reduces the three equations to a system of two equations which admits only a stable steady state. This apparent failure of the quasi-steady-state approximation to describe the limit-cycle oscillations observed in the full, three-variable system is analyzed in detail. We first show that the oscillations occur in the full system provided the Michaelis constants are sufficiently small. We then develop a method for determining the correct limit for application of the quasi-steady-state approximation. The leading problem consists of two equations for a conservative oscillator, and a higher order analysis is required in order to determine the amplitude of the limit-cycle oscillations. Finally, we observe a good agreement when comparing exact numerical and approximate bifurcation diagrams.

Key words. enzyme reactions, limit-cycle oscillations, quasi-steady-state approximation, singular perturbation

AMS subject classifications. 34E05, 34E15, 92C45

DOI. 10.1137/060654359

1. Introduction.

The quasi-steady-state approximation (QSSA) of chemical kinetics is a mathematical way of simplifying the differential equations describing some chemical kinetic systems. This approximation is a powerful tool for analyzing the dynamics of enzymatic reactions [1, 2, 3] exhibiting a wide range of time scales. The QSSA often yields revealing analytic formulas, and it frequently circumvents problems of stiffness in the numerical integration of systems of differential equations. Originally devised by biochemists on the basis that enzymes as catalysts act with small concentrations compared to the concentrations of their substrates, the QSSA is now recognized as belonging to singular perturbation theory. Ideally, this theory provides a method for the correct use of the QSSA, but it is too complicated for general use. Various investigations of special cases, such as the Michaelis–Menten reaction [4, 5, 6], give some indications of the applicability of the QSSA. But further clarification is called for, especially since the QSSA is virtually unavoidable in introductory texts on chemical or biochemical kinetics [1, 2, 3, 7, 8]. Biological oscillations often exhibit different time or amplitude scales, and the QSSA is widely used to analyze excitability and limit-cycle oscillations in the phase plane (see [9, 10, 11] for biochemical examples). In this paper, we concentrate on a three-variable model for the oscillations in

*Received by the editors March 15, 2006; accepted for publication (in revised form) August 25, 2006; published electronically December 21, 2006.
http://www.siam.org/journals/siap/67-2/65435.html

†Optique Nonlinéaire Théorique, Université Libre de Bruxelles, Campus Plaine, C.P. 231, 1050 Bruxelles, Belgium (terneux@ulb.ac.be). The work of this author was supported by the Fonds National de la Recherche Scientifique (Belgium).

‡Unité de Chronobiologie Théorique, Université Libre de Bruxelles, Campus Plaine, C.P. 231, 1050 Bruxelles, Belgium (agoldbet@ulb.ac.be). The work of this author was supported by the European Union through the Network of Excellence BioSim, contract LSHB-CT-2004-005137, and by grant 3.4636.04 from the Fonds de la Recherche Scientifique Médicale (F.R.S.M., Belgium).

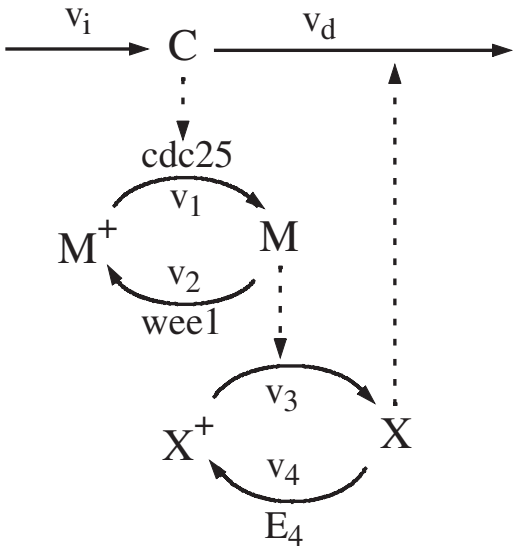


FIG. 1. *Bicyclic cascade model for the mitotic oscillator underlying the early cell cycle in amphibian embryos* [16, 17]. *Cyclin (C) is synthesized at a constant rate (v_i) and activates cdc*25 *phosphatase. The activated cdc*25 *phosphatase in turn activates cdc*2 *kinase (M) by dephosphorylating the inactive form (M^+). The activated cdc*2 *kinase is inactivated by the kinase wee*1. *In addition, a cyclin protease X is activated by cdc*2 *kinase and inactivated by an additional phosphatase (E_4). V_j ($j = 1 - 4$) denotes the effective maximum rate of each of the four converter enzymes; v_d denotes the maximum rate of cyclin degradation by protease X.*

the embryonic cell cycle and discover that the immediate application of the QSSA fails to describe these oscillations. Our main objective is to determine why the QSSA failed and how we may correctly use it in our problem.

Models for biochemical oscillations often contain one or more sigmoidal functions. Combined with positive and/or negative feedback loops, these functions allow the emergence of simple or complex oscillatory behavior. The sigmoidal dependence originates at the molecular level either from cooperative interactions in allosteric enzymes or from the phenomenon of zero-order ultrasensitivity in which two enzymes catalyzing opposite covalent modification reactions (e.g., phosphorylation-dephosphorylation) are saturated by their protein substrate [12]. While allosteric enzymes are abundant in metabolic regulation, covalent modification also plays an important role in biological signaling and cell regulation. Models for the oscillatory activity of allosteric enzymes have been studied analytically by taking advantage of the relatively large values of the allosteric constants [13, 14, 15]. However, the case of several enzymes working in a covalent modification cascade and exhibiting oscillatory activities has never been examined from an analytical point of view. In this paper, we consider a minimal model for biochemical oscillations underlying the embryonic cell cycle [16, 17]. See Figure 1. This model pertains to the situation encountered in early amphibian embryos, where the accumulation of cyclin suffices to trigger the onset of mitosis. In yeast and somatic cells, the mechanism involves additional checkpoints; see [2, 3, 18] and references therein. But what remains common to the various types of cell cycle mechanisms is the fact that they rely on the periodic activation of kinase cdc2 (also known as the cyclin-dependent kinase 1, cdk1). Cyclin activates the kinase cdc2, which promotes the degradation of cyclin. To produce oscillations, however,

activation and degradation cannot occur simultaneously and the negative feedback loop must be coupled to thresholds and time delays, which are naturally associated with phosphorylation-dephosphorylation cascades [16, 17].

The model is formulated in terms of the following three ordinary differential equations for the cyclin concentration C, the fraction of active cdc2 kinase M, and the fraction of active cyclin protease X [16, 17]:

$$\frac{dC}{dt} = v_i - v_d X \frac{C}{K_d + C} - k_d C, \tag{1}$$

$$\frac{dM}{dt} = V_{M1} \frac{C}{K_c + C} \frac{1 - M}{K_1 + 1 - M} - V_2 \frac{M}{K_2 + M}, \tag{2}$$

$$\frac{dX}{dt} = V_{M3} M \frac{1 - X}{K_3 + 1 - X} - V_4 \frac{X}{K_4 + X}. \tag{3}$$

In these equations, $1-M$ and $1-X$ represent the fractions of inactive cdc2 kinase and cyclin protease, respectively. v_i and v_d are the constant rate of cyclin synthesis and the maximum rate of cyclin degradation by protease X ($X = 1$), respectively. K_d and K_c denote the Michaelis constants for cyclin degradation and for cyclin activation of the cdc25 phosphatase acting on the phosphorylated form of the cdc2 kinase, respectively. k_d represents an apparent first-order rate constant related to nonspecific degradation of cyclin. The remaining parameters V_i and K_i ($i = 1$ to 4) denote the effective maximum rates and the Michaelis constants, respectively, for each of the enzymes E_i involved in the two cycles of phosphorylation-dephosphorylation. Moreover, the effective maximum rates $V_1(C)$ and $V_3(M)$ are given by $V_1 = V_{M1}C/(K_c + C)$ and $V_3 = V_{M3}M$.

The parameter K_d has been introduced to avoid the possibility that C becomes negative [16, 17]. For all our numerical solutions, however, we used $K_d = 0$ and found that C is always positive. Equation (1) with $K_d = 0$ is linear, and the nonlinearities necessary for the limit-cycle oscillations are given by the right-hand sides of (2) and (3). At steady state, the functions $M = M(C)$ and $X = X(M)$ obtained by setting the right-hand sides of (2) and (3) equal to zero are sigmoidal functions of M and X. The role of these functions for the oscillations is discussed in [16, 17]. Examples of limit-cycle oscillations for moderate and high values of V_{M1} and V_2 ($V_2/V_{M1} = 1/2$ fixed) are shown in Figure 2. They have been obtained by numerically solving (1)–(3). In Figure 2, top, the maximum rates V_{M1} and V_2 are moderate. The kinase M is activated as soon as C reaches a value close to 0.5. In Figure 2, bottom, the maximum rates V_{M1} and V_2 are large. In contrast to Figure 2, top, C remains close to 0.5. Time t, concentration C, and all parameters except the Michaelis constants have units as in [16, 17], where simulations are compared to experiments. We keep these equations in this form because they appear in all previous studies [16, 17, 30, 31].

We wish to describe the limit-cycle oscillations by using phase plane techniques. To this end, we propose to eliminate either M or X by using a QSSA. But, as we shall demonstrate, the reduced two-variable equations do no more than exhibit limit-cycle oscillations. By using a singular perturbation method, we then determine the correct two-variable limit that allows us to recover sustained oscillatory behavior. This analysis involves two steps. We first show that the leading approximation of the limit-cycle solution satisfies a two-variable conservative system of equations such as the Lotka–Volterra equations of chemical kinetics [1]. This system gives the correct relation

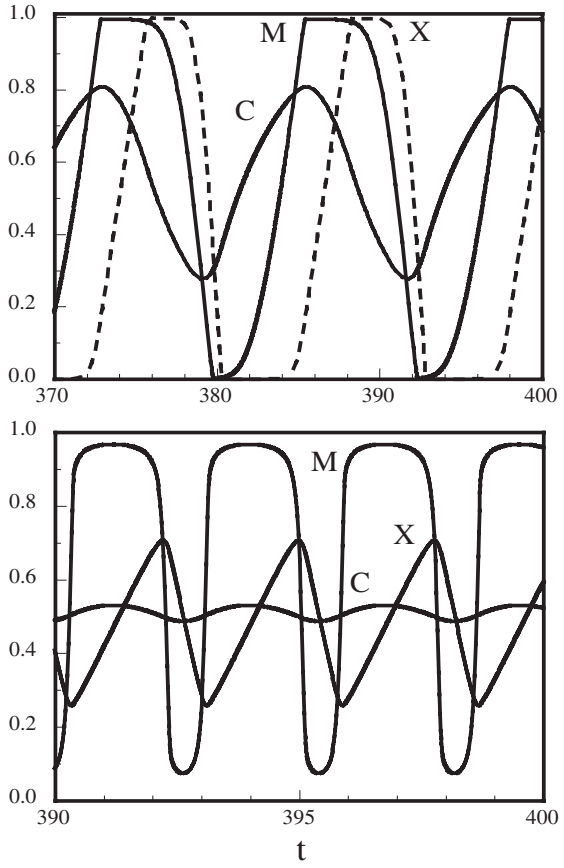


FIG. 2. *Limit-cycle oscillations in the enzymatic cascade model. M is dimensionless, C is measured in μM, and time t is measured in minutes. The values of the parameters are $K_j = K = 10^{-3}$ $(j = 1 - 4)$ and (in min^{-1}), $V_{M3} = 1$, $V_4 = 0.7$, $k_d = 0.25$; (in μM min^{-1}), $v_i = v_d = 0.25$; (in μM), $K_c = 0.5$, $K_d = 0$. Top: $V_{M1} = 3$ min^{-1} and $V_2 = 1.5$ min^{-1}; bottom: $V_{M1} = 3000$ min^{-1} and $V_2 = 1500$ min^{-1}.*

between the period and the amplitude of the oscillations. But in order to find how the amplitude changes with a given control parameter, a higher order analysis leading to a condition for bounded periodic solutions is needed. Similar singular perturbation techniques for particular Hopf bifurcation problems have been studied for chemical and biochemical relaxation oscillations [20], pulsating solidification fronts [21], and pulsating laser oscillations [24, 25].

The plan of the paper is as follows. In section 2, we show the failure of the standard QSSA and identify the source of the problem. Section 3 summarizes the results of our analysis, leading to the correct limit. The bifurcation diagram of the reduced two-variable equations is compared to the bifurcation diagram of the original three-variable equations. The main results are summarized in section 4. Mathematical details are given in appendices.

2. Failure of the QSSA. Equations (1)–(3) are too complicated for phase space analysis. A popular technique for simplifying the problem is to apply a QSSA for one of the dependent variables. This approximation (also called a pseudo-steady-state

hypothesis [26], steady-state assumption [27], or adiabatic elimination [28, 29]) is based on the assumption that the enzyme reacts so fast with the substrate that it can be taken as being in equilibrium, that is to say, $dM/dt \approx 0$ or $dX/dt \approx 0$. This approximation has been highly documented for the Michaelis–Menten reaction [1, 2, 3, 4, 5, 6] but has been used successfully for more complex systems exhibiting several enzymatic intermediates [2, 11]. The approximation is justified mathematically if a small parameter multiplies the time derivative of one of the dependent variables. This occurs in our problem if we consider the case of large values of both V_{M1} and V_2 (or, similarly, if we consider large values of V_{M3} and V_4). The proper way to apply the QSSA is to introduce the large parameter V defined as

$$V \equiv V_{M1} \tag{4}$$

and scale V_2 as

$$V_2 = V v_2, \tag{5}$$

where the coefficient v_2 is assumed to be an order one quantity. We may then factorize V in the right-hand side of (2) and rewrite this equation as

$$V^{-1}\frac{dM}{dt} = \frac{C}{K_c + C}\frac{1 - M}{K_1 + 1 - M} - v_2\frac{M}{K_2 + M}. \tag{6}$$

The coefficient of dM/dt is small because V is large. Thus, unless dM/dt is large, we may neglect this term and formulate the following algebraic equation for M and C:

$$\frac{C}{K_c + C}\frac{1 - M}{K_1 + 1 - M} - v_2\frac{M}{K_2 + M} = 0. \tag{7}$$

The QSSA is the assumption $V^{-1}dM/dt = 0$. Solving (7) for M and introducing $M = M(C)$ into (1) and (3) leads to two equations for only C and X. The two-variable problem represents a major simplification of our original three-variable equations, and we wonder if this reduced problem still admits limit-cycle oscillations. To this end, we examine the linear stability of the unique steady state $(C, X) = (C_s, X_s)$. We find that the coefficients of the characteristic equation for the growth rate σ are always positive. This means that $\mathrm{Re}(\sigma)$ is negative, implying stability of the steady state. A similar conclusion is obtained if we consider the case when V_{M3} and V_4 are large and eliminate the variable X.

We have thus found that a naive QSSA which allowed us to eliminate either M or X fails to describe the limit-cycle oscillations. But this approximation is nothing else than the leading term of an asymptotic solution and, like any asymptotic solution, it may admit different limits depending on the values of the other parameters in the problem. Returning to the original three-variable equations (1)–(3), we numerically investigate the behavior of the limit-cycle oscillations for progressively larger values of $V = V_{M1}$ and V_2 ($v_2 = V_2/V_{M1}$ fixed). We find that these oscillations persist only if we decrease the Michaelis constants K_j. The importance of these constants can be substantiated analytically by analyzing the Hopf bifurcation conditions in the double limits $V \to \infty$ and $K_j = K \to 0$ ($j = 1, 4$). The detailed analysis is given in Appendix A, where we show that the Hopf bifurcation point $K = K_H(V)$ scales like

$$K_H \sim V^{-1/2} \tag{8}$$

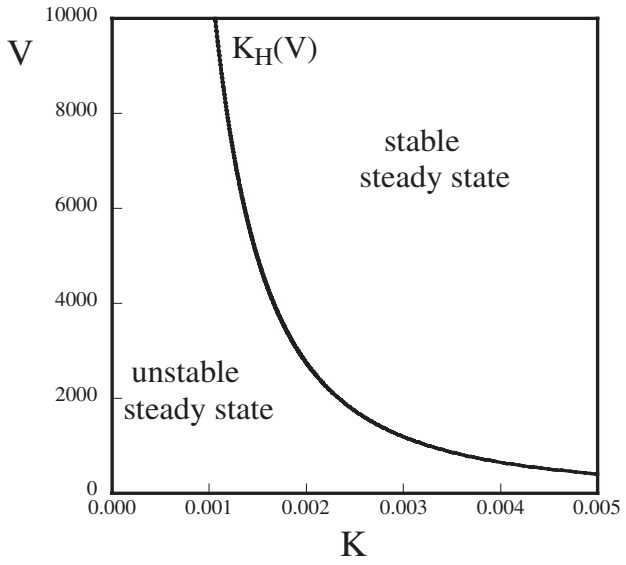


FIG. 3. *Hopf bifurcation line in the (K, V) parameter space. $V \equiv V_{M1}$ and $V_2 = V_{M1}/2$. The values of the other parameters are the same as in Figure 2. For progressively larger values of the maximum rate V, it is necessary to decrease the Michaelis constants (all equal to K) in order to keep the limit-cycle oscillations.*

as $V \to \infty$. Figure 3 shows the exact numerical Hopf bifurcation line for large values of V. It is in very good agreement with the approximation (47) derived in Appendix A. At a fixed value of V, the unique steady state undergoes a Hopf bifurcation at $K = K_H$, and the steady-state solution is unstable if $K < K_H$. From the Hopf conditions, we also learn that the frequency ω_H of the oscillations at the Hopf bifurcation point scales like

$$\omega_H \sim V^{1/4} \tag{9}$$

as $V \to \infty$, suggesting a short period oscillation for V large. We conclude that the QSSA based on the sole limit $V \to \infty$ cannot describe the oscillations unless we scale K as a $V^{-1/2}$ quantity and introduce a new time proportional to $V^{1/4}t$. In [16, 17], the value of V was much lower, allowing a larger K, for the observation of sustained oscillations.

3. Rescue of the QSSA. Phosphorylation-dephosphorylation cascades were already analyzed in the limit of small values of the Michaelis constants [19]. The difficulty here is that we need to scale both the maximum velocities and the Michaelis constants in order to describe the oscillations. Using the definitions (4) and (5), and motivated by the scaling laws (8) and (9), we introduce a small parameter ε, defined by

$$\varepsilon \equiv V^{-1/4}, \tag{10}$$

and scale the parameters V_2 and K_j $(j = 1 - 4)$ with respect to ε as

$$V_2 = \varepsilon^{-4} v_2 \text{ and } K_j = \varepsilon^2 k_j \ (j = 1 - 4). \tag{11}$$

In (11), v_2 and k_j are assumed to be order one coefficients. We also take into account (9) by introducing a new basic time T defined by

$$T \equiv \varepsilon^{-1} t. \tag{12}$$

Inserting (10)–(12) into (1)–(3) gives

$$\varepsilon^{-1} \frac{dC}{dT} = v_i - v_d X \frac{C}{K_d + C} - k_d C, \tag{13}$$

$$\frac{dM}{dT} = \varepsilon^{-3} \left[\frac{C}{K_c + C} \frac{1 - M}{\varepsilon^2 k_1 + 1 - M} - v_2 \frac{M}{\varepsilon^2 k_2 + M} \right], \tag{14}$$

$$\varepsilon^{-1} \frac{dX}{dT} = V_{M3} M \frac{1 - X}{\varepsilon^2 k_3 + 1 - X} - V_4 \frac{X}{\varepsilon^2 k_4 + X}. \tag{15}$$

Our QSSA now means the solution of these equations in the limit ε small. The analysis is long and tedious and is relegated to Appendix B. The results of our analysis are, however, simple and are summarized as follows for the case $K_d = 0$.

The leading approximation of the solution of (13)–(15) is described in terms of the variables M, U, and W, where U and W are defined as the deviations of C and X from their steady-state values

$$U \equiv \varepsilon^{-2}(C - C_0) \text{ and } W \equiv \varepsilon^{-1}(X - X_0). \tag{16}$$

In (16), C_0 and X_0 represent the steady-state values of C and X evaluated at $K_j = 0$ $(j = 1 - 4)$. They are defined as

$$C_0 \equiv \frac{K_c v_2}{1 - v_2} \text{ and } X_0 \equiv \frac{(v_i - k_d C_0)}{v_d}. \tag{17}$$

The leading order equations for M and W are then given by

$$F'(M) \frac{dM}{dT} = -v_d W, \tag{18}$$

$$\frac{dW}{dT} = V_{M3} M - V_4, \tag{19}$$

where

$$F(M) \equiv \frac{K_c v_2}{(1 - v_2)^2} \left[\frac{k_1}{1 - M} - \frac{k_2}{M} \right] \tag{20}$$

and

$$U = F(M). \tag{21}$$

Equations (18) and (19) form a conservative system of equations which admits a one-parameter family of periodic solutions. See Figure 4. For each point in the phase plane (W, M), there exists a closed orbit surrounding the center located at $(W, M) = (0, V_4/V_{M3})$. As the amplitude of the orbit increases, the period increases. The orbit becomes more and more rectangular and spends most of its time near $M = 1$ and $M = 0$.

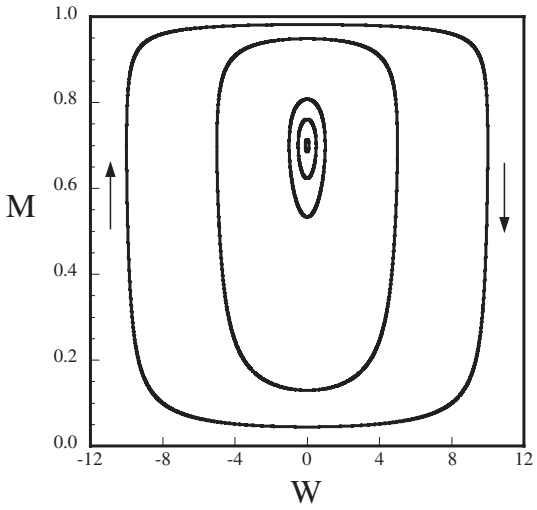


FIG. 4. *Orbits in the phase plane* (W, M) *obtained numerically from* (18)–(20). *The values of the parameters are* $v_2 = 0.5$, $V_{M3} = 1$, $V_{M4} = 0.7$, $v_d = 0.25$, $k_1 = k_2 = 1$, $K_c = 0.5$. *Each orbit corresponds to a periodic solution starting from a different initial point.* $M(0) = V_4/V_{M3} = 0.7$ *and* $W(0) = -0.1, -1, -5, -10$ *from the smallest to the largest orbit. The arrows indicate the direction of the time evolution.*

We now select the orbit of period P in the family of periodic solutions of (18) and (19). It is denoted by $U = U_P(T)$, $W = W_P(T)$, and $M = M_P(T)$. In order to determine how the period P (or equivalently, $U_P(T)$, $W_P(T)$, or $M_P(T)$) changes as we change a parameter, a higher order analysis is needed. This analysis is detailed in Appendix B. It leads to a solvability condition for bounded periodic solutions given by

$$v_d V_{M3} \frac{K_c}{(1-v_2)^2} \int_0^P \left(\frac{dM_P}{dT}\right)^2 dT - k_d \int_0^P \left(\frac{dU_P}{dT}\right)^2 dT = 0. \tag{22}$$

The two integrals must be computed numerically. This condition is the bifurcation equation since it relates the period of the oscillations and the physical parameters. In the limit of small amplitude solutions, $dU/dT = F'(M_0)dM/dT$, where $M_0 = V_4/V_{M3}$, and with $k_1 = k_2 = k$, (22) reduces to

$$\left[v_d V_{M3} \frac{K_c}{(1-v_2)^2} - k_d F'^2(M_0)\right] \int_0^P \left(\frac{dM_P}{dT}\right)^2 dT = 0 \tag{23}$$

which implies, using (20), that

$$k = k_H \equiv \sqrt{\frac{v_d V_{M3}}{k_d K_c}} \frac{(1-v_2)}{v_2} \left[\frac{1}{(1-M_0)^2} + \frac{1}{M_0^2}\right]^{-1}. \tag{24}$$

Using now $K = V_{M1}^{-1/2} k$ and $v_2 = V_2/V_{M1}$, expression (24) exactly matches the expression of the Hopf bifurcation point (47) obtained from the linearized theory in Appendix A. The expression (47) also is in excellent agreement with the exact numerical Hopf bifurcation line shown in Figure 3. We have thus verified that the bifurcation equation (22) correctly leads to the Hopf bifurcation point in the limit of small amplitude periodic solutions.

We next wish to find from (18)–(22) how the amplitude of the oscillations changes as we change the deviation $k - k_H$. To this end, we introduce

$$W = \sqrt{k}w \text{ and } T = \sqrt{k}s \tag{25}$$

into (18), (19) and obtain two equations for M and w that do not depend on k. They are of the form

$$f'(M)\frac{dM}{ds} = -v_d w, \tag{26}$$

$$\frac{dw}{ds} = V_{M3}M - V_4, \tag{27}$$

where

$$f(M) \equiv \frac{K_c v_2}{(1-v_2)^2}\left[\frac{1}{(1-M)} - \frac{1}{M}\right]. \tag{28}$$

Furthermore, substituting (25) into the bifurcation equation (22) leads to an expression for k^2 given by

$$k^2 = \frac{v_d V_{M3} K_c}{k_d(1-v_2)^2} \frac{\int_0^P \left(\frac{dM_P}{ds}\right)^2 ds}{\int_0^P f'^2(M)\left(\frac{dM_P}{ds}\right)^2 ds}. \tag{29}$$

The bifurcation equation (29) is now ready to be solved numerically: k appears only in the left-hand side, and the right-hand side is a function of the amplitude of the solution. Practically, we determine a P-periodic solution of (26) and (27) using the initial conditions

$$M(0) = V_4/V_{3M} \text{ and } w(0) = E, \tag{30}$$

where E is the parameter (in Figure 4, the orbits of different solutions are shown for $E = -0.1$, -1, -5, and -10). We then compute the two integrals in (29) and evaluate k^2. From an analysis of (26) and (27) in the phase plane, we note that $w = \pm E$ are the two extrema of $w(s)$. By gradually changing E from zero, we determine the function $E = E(k^2)$. Knowing E, we determine the extrema of W and X using first (25) and then (16). The bifurcation diagram of the extrema of X is shown in Figure 5 by the solid lines. As the amplitude of the periodic solutions increases, the Hopf bifurcation branch is first subcritical ($K > K_H$) and then folds back.

4. Discussion. The QSSA is widely used in the study of oscillations in biological and physical systems as a means to analyze limit-cycle behavior in the phase plane. Focusing on a biochemical model for limit-cycle oscillations in the embryonic cell cycle, we showed that a routine application of the QSSA leads to a two-variable system of equations that does not exhibit sustained oscillations. This apparent failure of the QSSA is, however, not a limitation of the method. We need to remember that the QSSA originates from an asymptotic method that considers a specific limit of a parameter (here the limit $V_{M1} = V$ large assuming V_2/V_{1M} fixed). From the Hopf bifurcation conditions, we showed that periodic solutions can be found only in the full system if we consider small values of the Michaelis constants. The correct scaling between these parameters and the other parameters in the model is provided by a careful analysis of the Hopf bifurcation conditions. A nonlinear analysis motivated by

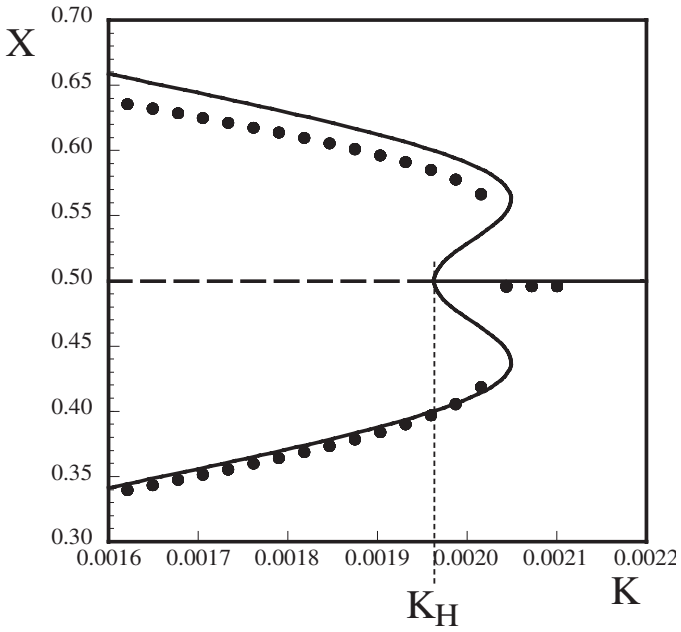


FIG. 5. *Maximum and minimum of the oscillations in X as a function of K. The approximation of the bifurcation diagram (solid lines) is compared to the bifurcation diagram of the original evolution equations* (1)–(3) *(dots). Parameter values are the same as in Figure* 4. *We represent the extrema of X as a function of K. The horizontal line is the steady state $X = 0.5$ which is unstable if $K < K_H \simeq 1.963 \times 10^{-3}$. The value of $\varepsilon = 0.135$. The agreement becomes better if we consider smaller values of ε (i.e., higher values of V_{M1} and V_2).*

these scalings then leads to a two-variable problem as the leading approximation. This problem admits conservative oscillations and therefore does not provide the bifurcation diagram of the amplitude of limit-cycle oscillations as a function of the control parameter. A higher order analysis was necessary in order to derive the bifurcation diagram. In other words, the limit-cycle solution is captured by a two-variable problem as one particular orbit in a family of periodic solutions, but the relation of this orbit to a specific value of the control parameter requires an extra solvability condition. In the strict spirit of the QSSA, the technique, even corrected by taking into account the scaling between parameters, is unable to provide a two-variable system exhibiting limit-cycle oscillations. However, it is not a limitation of the singular perturbation method, which tell us how to rescue the QSSA by supplementing it with a solvability condition.

Our results came from investigating two orders of a perturbation analysis after scaling parameters and variables with respect to a small parameter ε. It is not a local analysis near a Hopf bifurcation point because M is arbitrary even if C and X are assumed close to their steady-state values. This is why we obtained the global bifurcation diagram in Figure 5. An alternative to the perturbation theory is possible if we directly introduce the new variables T, U, and W into the original equations (1)–(3). Inserting (16) into (13)–(15) and simplying, we obtain the following equations for U, M, and W:

$$\frac{dU}{dT} = -v_d W - \varepsilon k_d U, \tag{31}$$

$$\frac{dM}{dT} = \varepsilon^{-1}\left[\begin{array}{c}\frac{U(1-v_2)^2(1-M)-v_2\left(K_c+\varepsilon^2U(1-v_2)\right)k_1}{(K_c+\varepsilon^2U(1-v_2))(1-M+\varepsilon^2k_1)}\\ +\ \frac{v_2k_2}{M+\varepsilon^2k_2}\end{array}\right], \tag{32}$$

$$\frac{dW}{dT} = V_{M3}\frac{M(1-X_0-\varepsilon W)}{1-X_0-\varepsilon W+\varepsilon^2k_3} - V_4\frac{X_0+\varepsilon W}{X_0+\varepsilon W+\varepsilon^2k_4}, \tag{33}$$

where X_0 is defined as in (17). The ε^{-1} term multiplying the right-hand side of (32) suggests that M will quickly approach a two-dimensional slow manifold. In this paper, we concentrated on the limit-cycle solution and did not investigate the evolution towards this slow manifold.

The standard QSSA failed in our three-variable model because the small parameter that motivated the elimination of one of the dependent variables also controlled the time scale of the remaining variables. This problem is known for two-variable models exhibiting relaxation oscillations [20] and for the standard laser rate equations [22]. A change of variables allows us to eliminate the singular perturbation difficulty. For three-variable systems, the problem has been analyzed for laser dynamical systems [23, 24, 25].

Appendix A. Linear theory. The Hopf bifurcation boundaries were investigated only numerically [30, 31]. In this appendix, we determine analytically the steady-state solution of (1)–(3) and its Hopf bifurcation point for the particular case when

$$K_d = 0 \text{ and } K_j = K \ (j = 1-4). \tag{34}$$

The steady-state solution $(C, M, X) = (C_s, M_s, X_s)$ satisfies the following three conditions:

$$v_i - v_dX - k_dC = 0, \tag{35}$$

$$\frac{V_{M1}C}{K_c+C}\frac{1-M}{K+1-M} - \frac{V_2M}{K+M} = 0, \tag{36}$$

$$V_{M3}M\frac{1-X}{K+1-X} - V_4\frac{X}{K+X} = 0. \tag{37}$$

If $K \to 0$, the steady state (C_s, M_s, X_s) approaches the limit (C_0, M_0, X_0). From (35)–(36) with $K = 0$, we find

$$C_0 = \frac{K_cV_2}{V_{M1}-V_2} > 0,\ M_0 = \frac{V_4}{V_{M3}},\ \text{and}\ X_0 = \frac{v_i-k_dC_0}{\nu_d} > 0. \tag{38}$$

Introducing the deviations $u = C - C_s$, $v = M - M_s$, and $w = X - X_s$, the linearized problem is given by

$$\left(\begin{array}{c}u'\\v'\\w'\end{array}\right) = \left(\begin{array}{ccc}-k_d & 0 & -v_d\\ \frac{V_{M1}K_c}{(K_c+C)^2}\frac{1-M}{K+1-M} & -F_1 & 0\\ 0 & V_{M3}\frac{1-X}{K+1-X} & -F_2\end{array}\right)\left(\begin{array}{c}u\\v\\w\end{array}\right), \tag{39}$$

where we have omitted the subscript s for the steady state. F_1 and F_2 are positive coefficients defined by

$$\begin{aligned}F_1 &\equiv K\left[\frac{C}{K_c+C}\frac{V_{M1}}{(K+1-M)^2} + \frac{V_2}{(K+M)^2}\right],\\ F_2 &\equiv K\left[\frac{V_{M3}M}{(K+1-X)^2} + \frac{V_4}{(K+X)^2}\right].\end{aligned} \tag{40}$$

From (39), we formulate the characteristic equation for the growth rate σ as

$$\sigma^3 - T_1\sigma^2 + T_2\sigma - T_3 = 0, \tag{41}$$

where

$$\begin{aligned} T_1 &= -k_d - F_1 - F_2, \\ T_2 &= k_d(F_1 + F_2) + F_1F_2, \\ T_3 &= -k_dF_1F_2 - V_{M3}V_{M1}v_d\frac{1-X}{K+1-X}\frac{K_c}{(K_c+C)^2}\frac{1-M}{K+1-M}. \end{aligned} \tag{42}$$

The conditions for a Hopf bifurcation are obtained by substituting $\sigma = i\omega$ into (41) and by separating the real and imaginary parts. We find

$$T_1T_2 - T_3 = 0 \text{ and } \sigma^2 = T_2 > 0. \tag{43}$$

The second condition is always verified. In order to satisfy the first condition, we determine $T_1T_2 - T_3$ and obtain

$$\begin{aligned} T_1T_2 - T_3 = &-k_d^2(F_1+F_2) - k_d(F_1+F_2)^2 - (F_1+F_2)F_1F_2 \\ &+ V_{M3}V_{M1}v_dK_c\frac{1-X}{K+1-X}\frac{1}{(K_c+C)^2}\frac{1-M}{K+1-M}. \end{aligned} \tag{44}$$

We wish to determine an approximation of (44) in the double limit K small and V large (V_{M1}/V_2 fixed). To this end, we need the leading expressions of F_1 and F_2 for K small. From (40) and using (38), we obtain

$$F_1 \simeq V_2K\left(\frac{1}{(1-M_0)^2} + \frac{1}{M_0^2}\right) \text{ and } F_2 \simeq V_4K\left(\frac{1}{(1-X_0)^2} + \frac{V_4}{X_0^2}\right). \tag{45}$$

We next consider the limit V_2 large (V_{M1}/V_2 fixed). In this limit, $F_1 \gg F_2$ and (44) simplifies as

$$\begin{aligned} T_1T_2 - T_3 &\simeq -k_dF_1^2 + V_{M3}V_{M1}v_d\frac{K_c}{(K_c+C_0)^2} \\ &= \left[\begin{array}{c} -k_dV_2^2K^2\left(\frac{1}{(1-M_0)^2} + \frac{1}{M_0^2}\right)^2 \\ + V_{M3}v_d\frac{(V_{M1}-V_2)^2}{K_cV_{M1}} \end{array}\right], \end{aligned} \tag{46}$$

where we eliminate C_0 using (38). The Hopf bifurcation point satisfies the condition $T_1T_2 - T_3 = 0$. Using (46) we find an expression of the Hopf bifurcation point $K = K_H$ given by

$$K_H = \frac{(V_{M1}-V_2)}{V_2\sqrt{V_{M1}}}\sqrt{\frac{V_{M3}v_d}{k_dK_c}}\left[\frac{1}{(1-M_0)^2} + \frac{1}{M_0^2}\right]^{-1}. \tag{47}$$

We have verified that this approximation compares well with the exact numerical solution shown in Figure 3. From (47), K_H scales like $V^{-1/2}$ as $V \to \infty$ ($V \equiv V_{M1}$, V_2/V_{M1} fixed). We also note from (43) that the frequency of the oscillations scales like

$$\omega = \sqrt{T_2} \simeq \sqrt{k_dF_1} = O(\sqrt{V_2K}) = O(V^{1/4}). \tag{48}$$

We conclude that a Hopf bifurcation is possible in the quasi-steady-state limit $V = V_{M1} \to \infty$ (V_2/V_{M1} fixed) provided that K is sufficiently small. The steady-state solution is unstable if $T_1T_2 - T_3 < 0$, which implies the inequality $K < K_H$.

Appendix B. Nonlinear bifurcation theory. Our starting point is (13)–(15), which we rewrite as

$$\frac{dC}{dT} = \varepsilon \left[v_i - v_d X \frac{C}{K_d + C} - k_d C \right], \tag{49}$$

$$\varepsilon^3 \frac{dM}{dT} = \frac{C}{K_c + C} \frac{1 - M}{\varepsilon^2 k_1 + 1 - M} - v_2 \frac{M}{\varepsilon^2 k_2 + M}, \tag{50}$$

$$\frac{dX}{dT} = \varepsilon \left[V_{M3} M \frac{1 - X}{\varepsilon^2 k_3 + 1 - X} - V_4 \frac{X}{\varepsilon^2 k_4 + X} \right] \tag{51}$$

so that all powers of ε are positive. We next seek a periodic solution of the form

$$\begin{gathered} C = C_0 + \varepsilon C_1 + \varepsilon^2 C_2 + \cdots,\ M = M_0 + \varepsilon M_1 + \varepsilon^2 M_2 + \cdots, \\ \text{and } X = X_0 + \varepsilon X_1 + \varepsilon^2 X_2 + \cdots . \end{gathered} \tag{52}$$

After introducing (52) into (49)–(51), we equate to zero the coefficients of each power of ε. This leads to a sequence of problems for the coefficients in (52). The leading order equations are

$$\frac{dC_0}{dT} = 0,\ \frac{C_0}{K_c + C_0} - v_2 = 0,\ \frac{dX_0}{dT} = 0. \tag{53}$$

Equation (53) admits the solution

$$C_0 = \frac{K_c v_2}{1 - v_2} \text{ and } X_0 = cst, \tag{54}$$

where X_0 is an unknown constant. The fact that X_0 and M_0 are unknown motivates the higher order analysis. The next problem is $O(\varepsilon)$ and is given by the following three equations:

$$\frac{dC_1}{dT} = v_i - v_d X_0 \frac{C_0}{K_d + C_0} - k_d C_0, \tag{55}$$

$$\frac{K_c C_1}{(K_c + C_0)^2} = 0, \tag{56}$$

$$\frac{dX_1}{dT} = V_{M3} M_0 - V_4. \tag{57}$$

From (56) and then from (55), we find C_1 and X_0 as

$$C_1 = 0 \text{ and } X_0 = \frac{(v_i - k_d C_0)(K_d + C_0)}{v_d C_0}. \tag{58}$$

We have determined X_0 but M_0 is still unknown. Thus, we consider the next problem, which is $O(\varepsilon^2)$:

$$\frac{dC_2}{dT} = -v_d \frac{C_0}{K_d + C_0} X_1, \tag{59}$$

$$\frac{K_c C_2}{(K_c + C_0)^2} - \frac{C_0}{K_c + C_0} \frac{k_1}{1 - M_0} + v_2 \frac{k_2}{M_0} = 0, \tag{60}$$

$$\frac{dX_2}{dT} - V_{M3} M_1 = 0. \tag{61}$$

From (60), we may determine M_0 as a function of C_2. Specifically, we define $M_0 = G(C_2)$ as the implicit solution of

$$C_2 = F(M_0) = \frac{K_c v_2}{(1-v_2)^2}\left[\frac{k_1}{1-M_0} - \frac{k_2}{M_0}\right]. \tag{62}$$

From (57) and (59), we eliminate X_1 and formulate a second-order differential equation for C_2:

$$\frac{d^2C_2}{dT^2} + \frac{v_d C_0}{K_d + C_0}\left[V_{M3}G(C_2) - V_4\right] = 0. \tag{63}$$

This equation is conservative and admits a one-parameter family of periodic solutions. This can be demonstrated in the phase plane by determining a first integral. The conservative nature of the oscillations means that the amplitude is arbitrary, and we still need to examine the higher order problem.

An equation for X_2 is already given by (61). From the $O(\varepsilon^3)$ equations, we obtain equations for C_3 and M_1 given by

$$\frac{dC_3}{dT} = -v_d X_2 \frac{C_0}{K_d + C_0} - v_d X_0 \frac{K_d}{(K_d + C_0)^2} C_2 - k_d C_2 \tag{64}$$

and

$$\frac{dM_0}{dT} = \frac{K_c C_3}{(K_c + C_0)^2} - \frac{C_0}{K_c + C_0}\frac{k_1 M_1}{(1-M_0)^2} - v_2 \frac{k_2 M_1}{M_0^2}. \tag{65}$$

Using (65), we determine M_1 as

$$M_1 = G'(C_2)C_3 - \frac{K_c}{(1-v_2)^2} G'(C_2)^2 \frac{dC_2}{dT}. \tag{66}$$

Then, using (61), (64), and (66), we obtain

$$\begin{aligned}\frac{d^2C_3}{dT^2} + v_d \frac{C_0}{K_d + C_0} V_{M3} G'(C_2) C_3 = v_d \frac{C_0}{K_d + C_0} V_{M3} \frac{K_c}{(1-v_2)^2} G'(C_2)^2 \frac{dC_2}{dT} \\ - \left[v_d X_0 \frac{K_d}{(K_d + C_0)^2} + k_d\right] \frac{dC_2}{dT}.\end{aligned} \tag{67}$$

By differentiating (63) with respect to T, we note that the homogeneous linear problem for C_3 admits the solution $C_{3H} = dC_2/dT$. The condition for a bounded periodic solution then implies that the right-hand side of (67) satisfies a solvability condition (Fredholm alternative [32]). Because the homogeneous problem is self-adjoint, this condition requires that the right-hand side is orthogonal to C_{3H}. This leads to the integral

$$\int_0^P \left[\begin{array}{c} v_d \frac{C_0}{K_d + C_0} V_{M3} \frac{K_c}{(1-v_2)^2} G'(C_2)^2 \\ - \left(v_d X_0 \frac{K_d}{(K_d + C_0)^2} + k_d\right) \end{array}\right] \left(\frac{dC_2}{dT}\right)^2 dT = 0 \tag{68}$$

and is the bifurcation equation. This equation can be further simplified, noting that

$$G'(C_2)^2 \left(\frac{dC_2}{dT}\right)^2 = \left(\frac{dM_0}{dT}\right)^2. \tag{69}$$

Equation (68) is then reformulated as

$$\int_0^P \left[\begin{array}{c} v_d \frac{C_0}{K_d+C_0} V_{M3} \frac{K_c}{(1-v_2)^2} \left(\frac{dM_0}{dT}\right)^2 \\ - \left(v_d X_0 \frac{K_d}{(K_d+C_0)^2} + k_d \right) \left(\frac{dC_2}{dT}\right)^2 \end{array} \right] dT = 0. \tag{70}$$

This condition determines the amplitude of the oscillations as a function of the physical parameters. If $K_d = 0$, it reduces to condition (22), which is analytically and numerically investigated.

REFERENCES

[1] J. D. Murray, *Mathematical Biology*, 3rd ed., Springer, New York, 2002.

[2] J. Keener and J. Sneyd, *Mathematical Physiology*, Interdiscip. Appl. Math. 8, Springer-Verlag, New York, 1998.

[3] C. P. Fall, E. S. Marland, J. M. Wagner, and J. J. Tyson, eds., *Computational Cell Biology*, Springer-Verlag, New York, 2002.

[4] L. A. Segel, *On the validity of the steady state assumption of enzyme kinetics*, Bull. Math. Biol., 50 (1988), pp. 579–593.

[5] L. A. Segel and M. Slemrod, *The quasi-steady-state assumption: A case study in perturbation*, SIAM Rev., 31 (1989), pp. 446–477.

[6] J. A. M. Borghans, R. J. De Boer, and L. A. Segel, *Extending the quasi-steady state approximation by changing variables*, Bull. Math. Biol., 58 (1996), pp. 43–63.

[7] C. C. Lin and L. A. Segel, *Mathematics Applied to Deterministic Problems in the Natural Sciences. With Material on Elasticity by G. H. Handelman*, 2nd ed., Classics Appl. Math. 1, SIAM, Philadelphia, 1988.

[8] I. R. Epstein and J. A. Pojman, *An Introduction to Nonlinear Chemical Dynamics*, Oxford University Press, Oxford, UK, 1998.

[9] A. Goldbeter, *Models for oscillations and excitability in biological systems*, in Mathematical Models in Molecular and Cellular Biology, L. A. Segel, ed., Cambridge University Press, Cambridge, UK, 1980, pp. 248–291.

[10] A. Goldbeter, T. Erneux, and L. A. Segel, *Excitability in the adenylate cyclase reaction in Dictyostelium discoideum*, FEBS Lett., 89 (1978), pp. 237–241.

[11] A. Goldbeter, *Biochemical Oscillations and Cellular Rhythms*, Cambridge University Press, Cambridge, UK, 1996.

[12] A. Goldbeter and D. E. Koshland, Jr., *An amplified sensitivity arising from covalent modification in biological systems*, Proc. Nat. Acad. Sci. U.S.A., 78 (1981), pp. 6840–6844.

[13] L. A. Segel and A. Goldbeter, *Scaling in biochemical kinetics: Dissection of a relaxation oscillator*, J. Math. Biol., 32 (1994), pp. 147–160.

[14] L. Holden and T. Erneux, *Slow passage through a Hopf bifurcation: From oscillatory to steady state solutions*, SIAM J. Appl. Math., 53 (1993), pp. 1045–1058.

[15] L. Holden and T. Erneux, *Understanding bursting oscillations as periodic slow passages through bifurcation and limit points*, J. Math. Biol., 31 (1993), pp. 351–365.

[16] A. Goldbeter, *A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase*, Proc. Nat. Acad. Sci. U.S.A., 88 (1991), pp. 9107–9111.

[17] A. Goldbeter, *Modeling the mitotic oscillator driving the cell division cycle*, Comments in Theoret. Biol., 3 (1993), pp. 75–107.

[18] J. J. Tyson, A. Csikasz-Nagy, and B. Novak, *The dynamics of cell cycle regulation*, BioEssays, 24 (2002), pp. 1095–1109.

[19] T. Erneux, R. D. Edstrom, and A. Goldbeter, *Enzyme sharing in phosphorylation-dephosphorylation cascades: The case where one protein kinase (or phosphatase) acts on two different substrates*, J. Theoret. Biol., 165 (1993), pp. 43–61.

[20] S. M. Baer and T. Erneux, *Singular Hopf bifurcation to relaxation oscillations*, SIAM J. Appl. Math., 46 (1986), pp. 721–739.

[21] G. J. Merchant, R. J. Braun, K. Brattkus, and S. H. Davis, *Pulsatile instability in rapid directional solidification: Strongly-nonlinear analysis*, SIAM J. Appl. Math., 52 (1992), pp. 1279–1302.

[22] G.-L. Oppo and A. Politi, *Toda potentials in laser equations*, Z. Phys. B, 59 (1985), pp. 111–115.

[23] G.-L. Oppo and A. Politi, *Center-manifold reduction for laser equations with detuning*, Phys. Rev. A (3), 40 (1989), pp. 1422–1427.
[24] T. Erneux and G. Kozyreff, *Nearly vertical Hopf bifurcation for a passively Q-switched microchip laser*, J. Statist. Phys., 101 (2000), pp. 543–552.
[25] G. Kozyreff and T. Erneux, *Singular Hopf bifurcation to strongly pulsating oscillations in lasers containing a saturable absorber*, European J. Appl. Math., 14 (2003), pp. 407–420.
[26] A. C. Fowler, *Mathematical Models in the Applied Sciences*, Cambridge Texts Appl. Math., Cambridge University Press, Cambridge, UK, 1997.
[27] P. C. Engel, *Enzyme Kinetics*, John Wiley & Sons, New York, 1977.
[28] H. Haken, *Synergetics*, 3rd ed., Springer, Berlin, 1983.
[29] S. H. Strogatz, *Nonlinear Dynamics and Chaos*, Addison-Wesley, New York, 1995.
[30] P. C. Romond, J. M. Guilmot, and A. Goldbeter, *The mitotic oscillator: Temporal self-organization in a phosphorylation-dephosphorylation enzymatic cascade*, Ber. Bunsenges. Phys. Chem., 98 (1994), pp. 1152–1159.
[31] A. Goldbeter and J.-M. Guilmot, *Thresholds and oscillations in enzymatic cascades*, J. Phys. Chem., 100 (1996), pp. 19174–19181.
[32] G. Iooss and D. D. Joseph, *Elementary Stability and Bifurcation Theory*, Undergrad. Texts Math., Springer-Verlag, New York, 1980, 2nd ed., Springer, New York, 1990.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 321–336

THE TRIPLE POINT PARADOX FOR THE NONLINEAR WAVE SYSTEM*

ALLEN M. TESDALL†, RICHARD SANDERS‡, AND BARBARA L. KEYFITZ†

Abstract. We present numerical solutions of a two-dimensional Riemann problem for the *nonlinear wave system* which is used to describe the Mach reflection of weak shock waves. Robust low order as well as high resolution finite volume schemes are employed to solve this equation formulated in self-similar variables. These, together with extreme local grid refinement, are used to resolve the solution in the neighborhood of an apparent but mathematically inadmissible shock triple point. Rather than observing three shocks meeting in a single standard triple point, we are able to detect a primary triple point containing an additional wave, a centered expansion fan, together with a sequence of secondary triple points and tiny supersonic patches embedded within the subsonic region directly behind the Mach stem. An expansion fan originates at each triple point. It is our opinion that the structure observed here resolves the von Neumann triple point paradox for the nonlinear wave system. These solutions closely resemble the solutions obtained in [A. M. Tesdall and J. K. Hunter, *SIAM J. Appl. Math.*, 63 (2002), pp. 42–61] for the unsteady transonic small disturbance equation.

Key words. weak shock reflection, self-similar solutions, nonlinear wave system, two-dimensional Riemann problems, von Neumann paradox

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

DOI. 10.1137/060660758

1. Introduction.

Experiments in which a weak shock wave reflects off a thin wedge appear to show a pattern of reflection in which three shocks meet at a triple point. However, the von Neumann theory of shock reflection [11] shows that Mach reflection, in which three shocks and a contact discontinuity meet at a triple point, is impossible for weak shocks. This apparent disagreement between theory and experiment was pointed out by von Neumann in 1943 and is referred to as the von Neumann, or triple point, paradox [8, 13].

In [13] numerical solutions were obtained of a problem for the unsteady transonic small disturbance equations that describes the reflection of weak shocks off thin wedges. The solutions were obtained in a parameter range where regular reflection is impossible, and contain a sequence of triple points in a tiny region behind the leading triple point, with a centered expansion fan originating at each triple point. It was shown that the triple points with expansion fans observed numerically are in fact consistent with theory, and that the presence of the expansion fans at the triple points resolves the paradox. A solution containing a supersonic patch and an expansion fan was first proposed by Guderley [5, 6]. Although Guderley did not offer evidence that

*Received by the editors May 24, 2006; accepted for publication August 25, 2006; published electronically December 21, 2006.

http://www.siam.org/journals/siap/67-2/66075.html

†Fields Institute, Toronto, ON M5T 3J1, Canada, and Department of Mathematics, University of Houston, Houston, TX 77204 (atesdall@fields.utoronto.ca, bkeyfitz@fields.utoronto.ca). The research of the first author was supported by National Science Foundation grant DMS 03-06307, NSERC grant 312587-05, and the Fields Institute. The research of the third author was supported by National Science Foundation grant DMS 03-06307, Department of Energy grant DE-FG02-03ER25575, and NSERC grant 312587-05.

‡Department of Mathematics, University of Houston, Houston, TX 77204 (sanders@math.uh.edu). The research of this author was supported by the National Science Foundation through grant DMS 03-06307.

this is what really occurs nor suggest that there is actually a sequence of expansion fans and triple points to resolve the triple point paradox, the term *Guderley Mach reflection* was chosen in [14] to name this new reflection pattern.

The nonlinear wave system is a simplification of the isentropic Euler equations obtained by dropping the momentum transport terms from the momentum equations [4]. Compared to the unsteady transonic small disturbance equations, the nonlinear wave system is closer in structure to the Euler equations: it is linearly well-posed in space and time, it has a characteristic structure similar to the Euler equations, and change of type takes the equations from a hyperbolic to a mixed-type system. These features make the nonlinear wave system a useful prototype for studying two-dimensional Riemann problems for the full Euler equations.

A problem for the nonlinear wave system that is the analogue of the reflection of weak shocks off thin wedges was studied in [3]. In a parameter range where regular reflection is not possible, the authors showed existence of the subsonic solution behind the Mach shock and reflected wave by solving a free boundary problem for the Mach shock. They did not find the actual reflected shock, but instead based their solution on modeling it as a continuous function with a singularity in the derivative at the sonic boundary. They showed that the composite solution they obtained is not a weak solution near the sonic line. The actual solution, therefore, is different from the construction they present, and they suggest two alternatives. Since triple point solutions do not exist for the nonlinear wave system, one possibility is that the reflected shock is a weak shock that has zero strength at the reflection point. Another possibility is Guderley Mach reflection, as obtained in [13].

Several numerical solutions of the weak shock reflection problem for the nonlinear wave system have been computed. In separate work, R. Sanders, A. Kurganov, and M. Lukacova-Medvidova (all unpublished; see [9]) computed numerical solutions of the problem studied in [3] over a wide range of parameter space where regular reflection is impossible. None of these solutions, however, are sufficiently well resolved to determine the nature of the solution near the apparent triple point. For example, it cannot be determined from any of these solutions whether the reflected shock has zero strength at the triple point, or if some other reflection pattern, such as Guderley Mach reflection, occurs. In fact, in the best resolved of these solutions, three shocks do appear to meet at a triple point—the triple point paradox.

In this paper we present high resolution numerical solutions of the shock reflection problem for the nonlinear wave system. Our most highly resolved solution shows that Guderley Mach reflection occurs at a set of parameter values where regular reflection is impossible: there is a sequence of supersonic patches behind the leading triple point, formed by a sequence of expansion fans and shocks that reflect between the sonic line and the Mach shock. This numerical solution is remarkably similar to those obtained for the unsteady transonic small disturbance equations in [13], and as in [13] the numerical results suggest that the sequence of triple points in an inviscid weak shock Mach reflection may be infinite.

Recent experimental evidence appears to confirm that the resolution of the triple point paradox obtained in [13] and again in the present paper is correct. Skews and Ashworth in [12] obtained schlieren photographs of shock reflection experiments which show a sequence of shocks and expansion waves behind the triple point in a weak shock Mach reflection. The supersonic region is extremely small, as discussed in [13], which is why it had never been observed before. Skews and Ashworth overcame this difficulty by using a specially designed shock tube and flow visualization enhancement techniques.

The numerical solutions of Sanders, Kurganov, and Lukacova-Medvidova were obtained by solving an initial-value problem for the unsteady nonlinear wave system. The problem of inviscid shock reflection off a wedge is self-similar, and there are advantages to solving the problem in self-similar, rather than unsteady, variables. In the unsteady formulation any waves which are present initially move through the numerical domain, making local grid refinement strategies difficult. By contrast, a solution of the self-similar equations is stationary, and local grid refinement near the triple point is much easier to implement. Moreover, in self-similar variables a global grid continuation procedure can be used in which a partially converged solution on a coarse grid is interpolated onto a fine grid, and then driven to convergence on the fine grid. In this paper we present numerical solutions of the shock reflection problem for the nonlinear wave system computed in self-similar coordinates. Procedures for solving the unsteady transonic small disturbance equations in self-similar variables were developed in [13], and are extended here to apply to the nonlinear wave system.

This paper is organized as follows. In section 2 we describe the shock reflection problem for the nonlinear wave system. In section 3 we discuss our approach to solving this problem numerically. The numerical results obtained are presented in section 4. In section 5 we discuss questions raised by our results. Finally, we summarize our findings in section 6.

2. The shock reflection problem for the nonlinear wave system. We consider a problem for the nonlinear wave system that is analogous to the reflection of weak shocks off thin wedges [3]. The shock reflection problem consists of the nonlinear wave system

$$\begin{aligned} &\rho_t + (\rho u)_x + (\rho v)_y = 0, \\ &(\rho u)_t + p(\rho)_x = 0, \\ &(\rho v)_t + p(\rho)_y = 0 \end{aligned} \tag{2.1}$$

in the half space $x > 0$ with piecewise constant Riemann data consisting of two states separated by a discontinuity located at $x = \kappa y$. Here, $\rho(x, y, t)$ is the density, $u(x, y, t)$ and $v(x, y, t)$ are the x and y components of velocity, respectively, and $p(\rho)$ is the pressure. For convenience, we assume a polytropic gas law

$$p(\rho) = C\rho^\gamma,$$

where C is a constant and γ is the ratio of specific heats. Letting $U = (\rho, m, n)$ denote the vector of conserved variables, where $m = \rho u$ and $n = \rho v$, the Riemann data are

$$U(x, y, 0) = \begin{cases} U_1 \equiv (\rho_1, 0, 0) & \text{if } x < \kappa y, \\ U_0 \equiv (\rho_0, 0, n_0) & \text{if } x > \kappa y. \end{cases} \tag{2.2}$$

We choose $\rho_0 > \rho_1$ to obtain an upward moving shock in the far field, and determine n_0 so that the one-dimensional wave between U_0 and U_1 at angle κ consists of a shock and a contact discontinuity with a constant middle state between them. The following expression for n_0 was obtained in [3]:

$$n_0 = \frac{1}{\kappa}\sqrt{(1 + \kappa^2)(p(\rho_0) - p(\rho_1))(\rho_0 - \rho_1)}. \tag{2.3}$$

Strictly speaking, data for reflection from a wedge of angle θ radians would explicitly include the wedge as a discontinuous change of slope, of angle θ, in the boundary

at the point $(0,0)$. In replacing a domain that imitates the physics by a half-plane $(x > 0)$, we are assuming that the reflection pattern near the apparent triple point is a local phenomenon. The physical wedge angle θ in this model is related to κ in (2.2), (2.3) by

$$\theta = \tan^{-1}(1/\kappa). \tag{2.4}$$

This problem depends on two parameters: the inverse slope κ of the incident shock, and the incident shock strength ρ_0/ρ_1 (see Appendix A). For values of κ greater than a critical value κ_R which depends on ρ_0 and ρ_1, a regularly reflected solution of (2.1)–(2.3) is impossible. In addition, triple point solutions of (2.1), in which three plane shocks separated by constant states meet at a point, do not exist (see Appendix B for a proof of this). We note that a self-similar solution in which three shocks and a linear wave meet at a point can be constructed. However, this is not consistent with the initial data, since (2.1) implies $(m_y - n_x)_t = 0$, even for weak solutions, and slip lines are characterized by nonzero values of $m_y - n_x$. Therefore, Mach reflection cannot occur when regular reflection becomes impossible, and the shock reflection problem for the nonlinear wave system embodies the triple point paradox in an essential form.

The problem (2.1)–(2.3) is self-similar, so the solution depends only on the similarity variables

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

We write (2.1) in the form

$$U_t + F_x + G_y = 0, \tag{2.5}$$

where

$$U = (\rho, m, n), \qquad F = (m, p, 0), \quad \text{and} \quad G = (n, 0, p).$$

Writing (2.5) in terms of ξ, η, and a pseudo-time variable $\tau = \log t$, we obtain

$$U_\tau - \xi U_\xi - \eta U_\eta + F_\xi + G_\eta = 0. \tag{2.6}$$

As $\tau \to +\infty$, solutions of (2.6) converge to a pseudosteady, self-similar solution that satisfies

$$-\xi U_\xi - \eta U_\eta + F_\xi + G_\eta = 0. \tag{2.7}$$

Equation (2.7) is hyperbolic when $c^2(\rho) < \xi^2 + \eta^2$, corresponding to supersonic flow in a self-similar coordinate frame, and of mixed type when $c^2(\rho) > \xi^2 + \eta^2$, corresponding to subsonic flow. Here, $c(\rho) = \sqrt{p_\rho}$ denotes the local sound speed. The equation changes type across the sonic line given by

$$\xi^2 + \eta^2 = c^2(\rho). \tag{2.8}$$

3. The numerical method. In order to solve (2.6) numerically, we write it in conservative form as

$$U_\tau + (F - \xi U)_\xi + (G - \eta U)_\eta + 2U = 0. \tag{3.1}$$

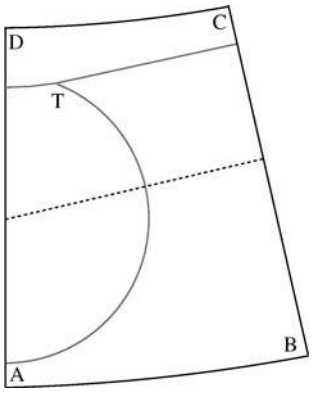


FIG. 1. *A schematic diagram of the computational domain. AD is the wall and ABCD is the far field numerical boundary. The incident shock enters the computational domain normal to BC. The incident (right of T), reflected (below T), and Mach (left of T) shocks meet at the triple point T.*

In self-similar variables, the nonlinear wave system has the form of the unsteady equations (2.5) with modified flux functions and a lower-order source term.

The essential feature of the numerical method is the capability to locally refine the grid in the area of the apparent triple point. We designed several nonuniform, logically rectangular, finite volume grids so that a given incident shock is aligned with the grid in the far field; see Figure 1. Specifically, each problem with a given incident shock angle has an associated fitted finite volume C-grid. Grid continuation is employed whereby partially converged numerical solutions are quadratically interpolated onto a refined grid. Inside a given box surrounding the triple point, uniform grid spacing is used. Outside of this box, the grid is exponentially stretched in both grid directions.

The basic finite volume scheme is quite standard. Each grid cell, Ω, is a quadrilateral, and using $\vec{\nu} = (\nu_\xi, \nu_\eta)$ to denote the normal vector to a typical side of Ω, numerical fluxes are designed to be consistent with

$$\widetilde{F}(U) = (F(U) - \xi U)\,\nu_\xi + (G(U) - \eta U)\,\nu_\eta = \begin{pmatrix} \nu_\xi m + \nu_\eta n - \bar{\xi}\,\rho \\ \nu_\xi p - \bar{\xi}\,m \\ \nu_\eta p - \bar{\xi}\,n \end{pmatrix},$$

where $\bar{\xi} = (\vec{\xi} \cdot \vec{\nu})$ and $\vec{\xi} = (\xi, \eta)$. Since $\vec{\xi}$ varies, our numerical flux formulae evaluate $\vec{\xi}$ frozen at the midpoint of each cell side. We use two distinctly different numerical fluxes in our results presented below: a first-order Lax–Friedrichs numerical flux and a high-order variant of the Roe numerical flux. The Lax–Friedrichs flux is

$$H_{LF} = \frac{1}{2}\left(\widetilde{F}(U_l) + \widetilde{F}(U_r) - \Lambda\,(U_r - U_l)\right),$$

where $\Lambda > 0$ is a scalar constant chosen to be larger than the fastest wave speed found on the computational domain. While the Lax–Friedrichs method yields only first-order accurate approximations, we regard it to be extremely robust. Our high-order Roe scheme is obtained from piecewise linear reconstruction with characteristic variable limiting, together with the Roe flux

$$H_{Roe} = \frac{1}{2}\left(\widetilde{F}(U_l) + \widetilde{F}(U_r) - R\Lambda L\,(U_r - U_l)\right),$$

where $\Lambda = \mathrm{diag}(|-\bar{\xi}-c|, |-\bar{\xi}|, |-\bar{\xi}+c|)$, and R and L are the matrices of right and left eigenvectors to the Jacobian of $\widetilde{F}$ evaluated at the midpoint $U_{Roe} = \frac{1}{2}(U_l + U_r)$. Below, we use the equation of state $p = 1/2\rho^2$. Therefore, using the midpoint for evaluation yields an exact Roe average since in this case $\widetilde{F}$ is quadratic.

Time integration is accomplished by the forward Euler method for the Lax–Friedrichs scheme:

$$\frac{U^{n+1} - U^n}{\Delta\tau} + \frac{1}{|\Omega|}\int_{\partial\Omega} H^n_{LF}\, ds + 2U^n = 0.$$

For reasons of linear stability, we use the explicit trapezoidal rule to integrate the high-order Roe scheme, as follows:

$$\frac{U^{n+1/2} - U^n}{\Delta\tau} + \frac{1}{|\Omega|}\int_{\partial\Omega} H^n_{Roe}\, ds + 2U^n = 0,$$
$$\frac{2U^{n+1} - U^{n+1/2} - U^n}{\Delta\tau} + \frac{1}{|\Omega|}\int_{\partial\Omega} H^{n+1/2}_{Roe}\, ds + 2U^{n+1/2} = 0.$$

3.1. The grid, initialization, and boundary conditions. We computed solutions of the half-space problem (2.1)–(2.3) in the finite computational domain shown schematically in Figure 1. We use a nonuniform grid that has a locally refined area of uniform grid very close to the triple point, and is stretched exponentially away from the triple point toward the outer numerical boundaries and the wall. (Exponential stretching of 1% means $\Delta x_{i+1} = 1.01\,\Delta x_i$.) In the solutions shown below, the nonuniform grids are stretched by amounts between 0.5% and 1%. The total number of grid points in our largest grid is approximately 11×10^6, of which approximately 2.5×10^6 cover a very small region surrounding the triple point. (See Figure 3(c) below where this small region is depicted.)

We impose reflecting boundary conditions, equivalent to the physical no-flow condition, on the wall AD. A standard first-order ghost cell implementation, with fictitious cells located to the left of the boundary AD, is given by

$$\begin{aligned} \rho_{-1} &= \rho_0, \\ m_{-1} &= -m_0, \\ n_{-1} &= n_0, \end{aligned} \tag{3.2}$$

where the subscripts -1 and 0 indicate values at ghost cells and at the first real cell adjacent to the boundary, respectively. In our higher-order computations we used a second-order formulation of this boundary condition. In addition, we require numerical boundary conditions on the outer computational boundaries.

In [3] expressions were given for the one-dimensional wave between U_0 and U_1 in the far field. The constant middle state $U_m = (\rho_0, m_m, n_m)$ between the contact discontinuity (the dotted line in Figure 1), located at $\xi = \kappa\eta$, and the incident shock, located at $\xi = \kappa\eta + \chi$, is given by

$$\begin{aligned} m_m &= -\sqrt{\frac{(p(\rho_0) - p(\rho_1))(\rho_0 - \rho_1)}{1+\kappa^2}}, \\ n_m &= -\kappa m_m, \\ \text{with} \quad \chi &= -\sqrt{1+\kappa^2}\sqrt{\frac{p(\rho_0) - p(\rho_1)}{\rho_0 - \rho_1}}. \end{aligned} \tag{3.3}$$

On the outer numerical boundary $ABCD$, we impose Dirichlet data corresponding to the incident shock/contact discontinuity solution in (2.2), (2.3), (3.3). We find that

$$U(\xi,\eta) = \begin{cases} U_1, & \xi < \kappa\eta + \chi, \\ U_m, & \kappa\eta + \chi < \xi < \kappa\eta, \\ U_0, & \xi > \kappa\eta. \end{cases} \tag{3.4}$$

We impose (3.4) as a boundary condition for (3.1) on $ABCD$.

4. Numerical results. We computed numerical solutions of (2.1)–(2.3) for κ equal to 1, 2, 4, and 8. In our computations we used ρ_0/ρ_1 equal to 64, 8, and 2. In the following figures we present solutions with $\rho_1 = 1$ and ρ_0/ρ_1 equal to 64. The solutions for other values of ρ_0/ρ_1 are similar to the ones presented here. For all computations, the polytropic gas law $p = \frac{1}{2}\rho^2$ was used. Figure 2 shows ρ-contour plots of the global solutions as a function of $(x/t, y/t)$. From (2.4), increasing κ corresponds to decreasing the wedge angle that is modeled by our problem. Hence, the sequence of plots in Figure 2(a)–(d) is a numerical representation of a series of shock reflection experiments in which the wedge angle is decreased, while holding the shock strength ρ_0/ρ_1 constant.

The numerical solutions appear to show a simple Mach reflection, with three shocks meeting at a triple point. The Mach shock becomes longer and weaker as κ increases, and the strength of the reflected shock also decreases when κ increases. For a fixed value of κ, the strength of the Mach shock increases as it moves away from the triple point, reaching a maximum at the wall $x = 0$.

For the value $\kappa = 1$, we used local grid refinement to obtain a highly resolved solution in the neighborhood of the triple point. In Figure 3(a)–(c), we show ρ-, m-, and n-contours and the numerically computed location of the sonic line, equation (2.8), near the triple point for $\kappa = 1$. The solution contains a small region of supersonic flow behind the triple point. The width $\Delta(x/t)$ of the patch is approximately 0.03, and the height $\Delta(y/t)$ is approximately 0.01. Here, the width $\Delta(x/t)$ is a numerical estimate of the difference between the maximum value of x/t on the sonic line and the minimum value of x/t at the rear sonic point on the Mach shock. The height $\Delta(y/t)$ is an estimate of the difference between the value of y/t at the triple point and the minimum value of y/t at the rear sonic point on the Mach shock. The width of the supersonic region is approximately 5% of the length of the Mach shock. The expansion fan centered at the leading triple point can be clearly seen. Behind the leading triple point, there is a sequence of shocks and expansion fans. The thickening of the incident shock as it moves away from the triple point in Figure 3(a)–(c) is caused by the use of a stretched grid.

The area covered by the most refined uniform grid is indicated by the box contained in Figure 3(c), and the figure caption gives the number of grid points in the most refined area of the grid. The box appears to be skewed because of the use of a C-grid. To illustrate the size and location of the refined uniform grid, in Figure 3(d) we plot ρ-contour lines over the entire numerical domain, for $\kappa = 1$. The refined grid area is too small to be visible in the main plot shown in Figure 3(d). The inset figure shows an enlargement of the solution contained within the small rectangular box centered about the reflection point, as indicated. The solution shown in the inset figure also contains a small box centered at the reflection point, indicating the approximate size and location of the region shown in Figure 3(a)–(c).

We found that, as in [13], a certain minimum grid resolution was required to resolve the supersonic region behind the triple point. As we refined the grid beyond

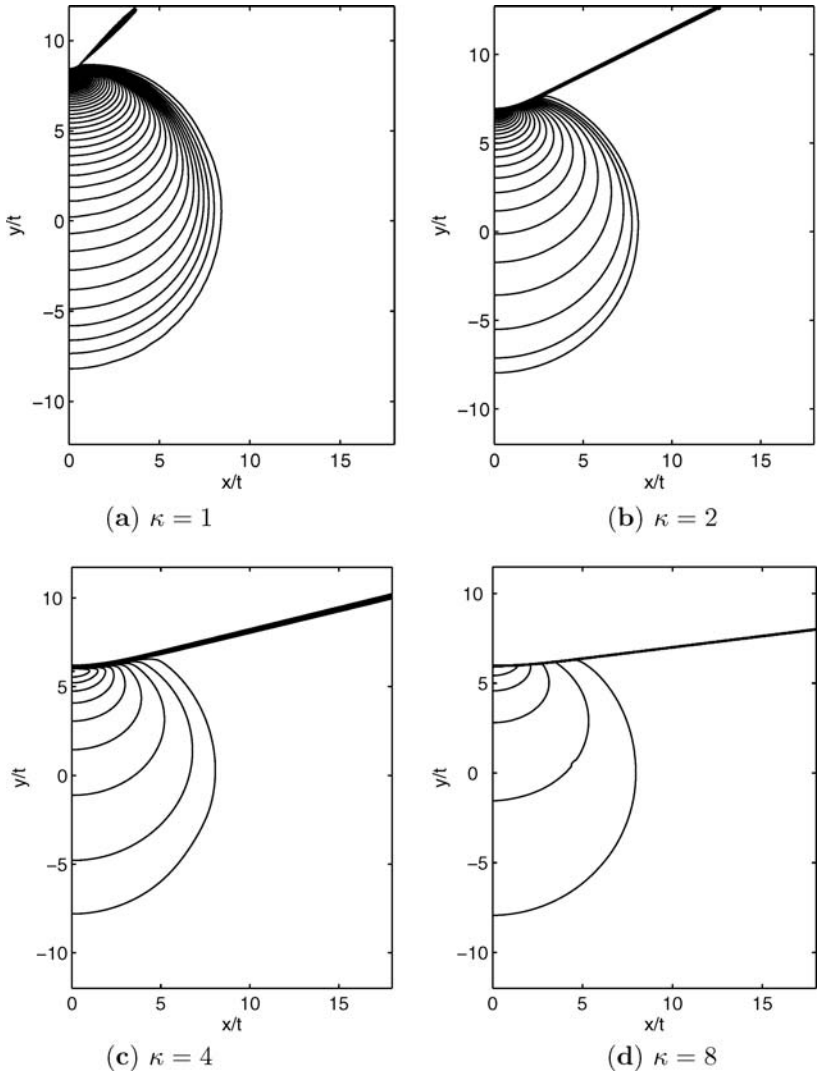


FIG. 2. *Contour plots of ρ for increasing values of κ. The ρ-contour spacing is* 1.0. *The shock strength $\rho_0/\rho_1 = 64$; $\rho_1 = 1$.*

this minimum level, details of the flow field near the triple point became clearer. Figure 4 shows ρ-contours and the sonic line near the triple point for a sequence of solutions for $\kappa = 1$, using a Lax–Friedrichs numerical flux. The sequence was computed using successively refined grids, with each grid refined by a factor of two in x/t and y/t in relation to the previous grid. The resolution of the locally refined areas is indicated on the plots. In Figure 4(a)–(b), the sonic line appears smooth. At the next level of refinement, shown in Figure 4(c), there is a steepening of the contours at the rear of the patch, and an indication of a shock forming there. Further shocks appear in our highest resolution solution in Figure 3. At resolutions lower than shown in the figure, the supersonic region disappears entirely.

There is a small discrepancy between the location of the triple point in these figures and the theoretical location of the incident shock, given in (3.3). The reason for

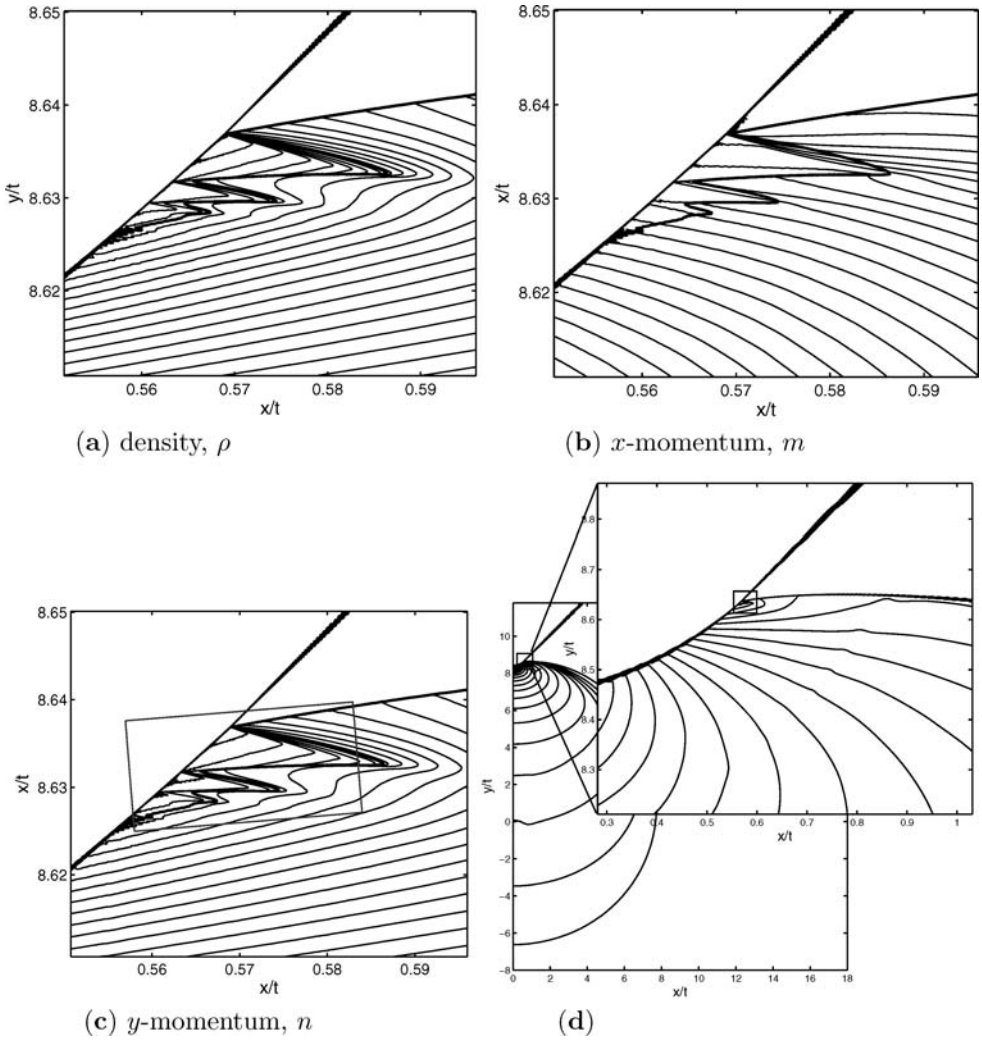


FIG. 3. *The contour plots in* (a)–(c) *show the true nature of the solution near the triple point for* $\kappa = 1$. *The* ρ*-contour spacing is* 0.5 *in* (a), *the* m*-contour spacing is* 1.5 *in* (b), *and the* n*-contour spacing is* 5.25 *in* (c). *The heavy line is the sonic line. The box in* (c) *indicates the area of the refined uniform grid, which has* 2048×1320 *grid points. A second-order Roe numerical flux was used. The plot in* (d) *is an illustration of the approximate size and location of the region shown in the plots in* (a)–(c), *which is contained in the small rectangular box shown in the inset figure; the plot shows contour lines of* ρ.

the discrepancy is that the numerical boundary conditions did not give an incident shock that was of exactly constant strength and exactly straight in the $(x/t, y/t)$ coordinates. However, the deviation of the numerical solution for the incident shock from the exact uniform solution was small. For example, in our numerical solution shown in Figure 3, the numerically computed value of the y/t coordinate of the triple point differs by 0.1% from the theoretical value obtained from (3.3) using the numerically computed value of x/t, and the nonuniformity in ρ in the state behind the incident shock near the triple point is about 0.6%. We tried a number of different implementations of the numerical boundary conditions and computational mesh, but none of

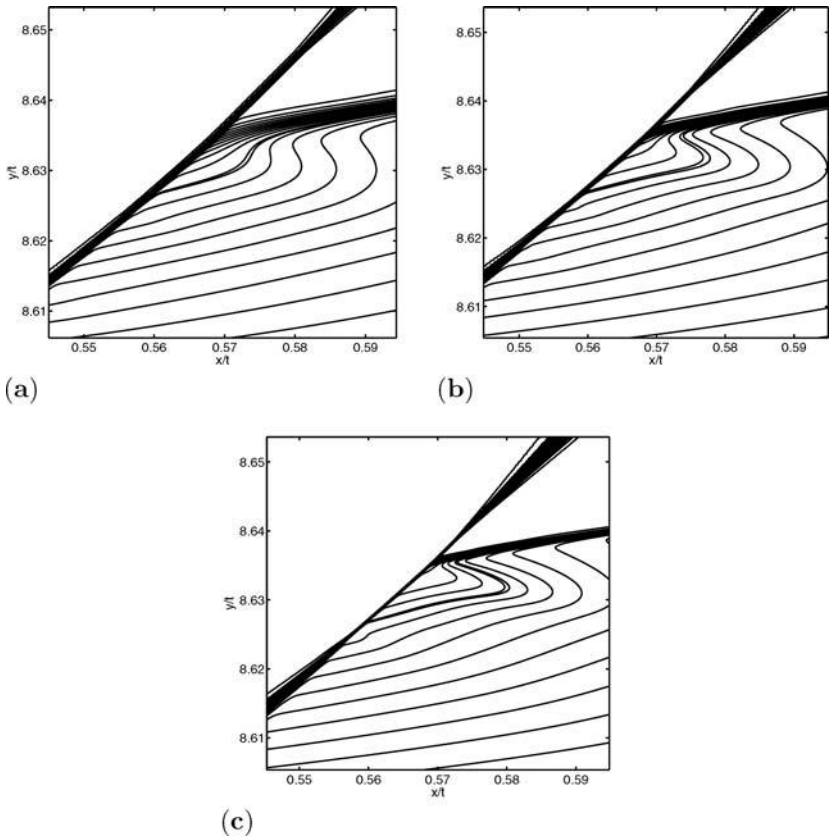


FIG. 4. *A sequence of contour plots illustrating the effect of increasing grid resolution on the numerical solution. The solutions plotted here are for* $\kappa = 1$. *The figures show* ρ*-contours in the refined grid area near the triple point, with a* ρ*-contour spacing of* 1.0. *Each grid is refined by a factor of two in relation to the previous grid. The region shown includes the refined uniform grid area. The heavy line is the sonic line. In* (a)*, the refined uniform grid contains* 760×760 *grid points. A supersonic region is visible as a bump in the sonic line, but it is poorly resolved. In* (b)*, the refined uniform grid area contains* 1280×1024 *grid points. The supersonic region appears to be smooth. In* (c)*, the refined uniform grid area contains* 2048×1320 *grid points. There is an indication of an expansion fan behind the leading triple point.*

them gave an incident shock that was of exactly constant strength. Nevertheless, the presence of a supersonic patch did not depend on the particular implementation.

In the computation for $\kappa = 1$, we partially converged a solution on a coarse grid, resampled the data onto a refined grid, and repeated the process until the necessary resolution was obtained. Three consecutive intermediate solutions in this computation are shown in Figure 4. Computations on less refined grids were made using a Lax–Friedrichs numerical flux, and after partial convergence on the most refined grid we switched to the more expensive Roe method. Figure 5 shows ρ-contours for a solution made using a first-order Roe scheme. Further computation using a second-order Roe scheme yielded the final solution shown in Figure 3. The solution on the final grid was evolved until no further change was observed in the details of the solution near the triple point. The solutions shown in Figures 4(c), 5, and 3 were obtained on the same grid using different methods. All three of the solutions contain a small supersonic region behind the triple point. The solutions shown in Figures 5 and 3,

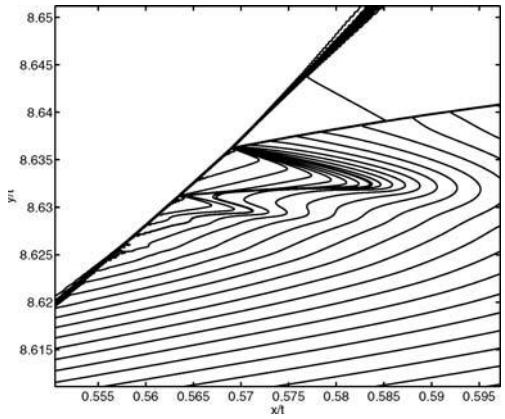


FIG. 5. *A contour plot of ρ near the triple point computed using a first-order Roe method. The number of points in the refined uniform grid is the same as in Figure* 4, *which shows a Lax–Friedrichs solution, and in Figure* 3(a)–(c), *which shows a second-order Roe method solution. The ρ-contours are plotted at the same levels of ρ as in Figure* 3(a).

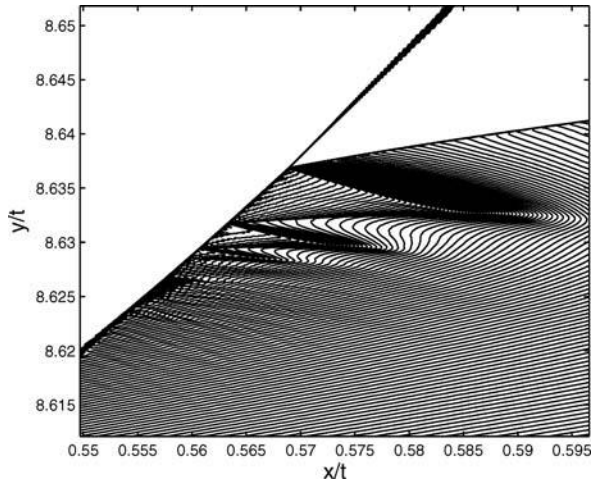


FIG. 6. *A detailed plot of contour lines for ρ illustrating Guderley Mach reflection. The ρ-contour spacing is* 0.1. *Three reflected shock/expansion wave pairs are clearly visible, with indications of a fourth. The region shown contains the refined uniform grid, which has* 2048×1320 *grid points.*

which are more highly resolved, contain a sequence of supersonic patches and triple points, which is better defined in Figure 3.

In Figure 6 we plot closely spaced ρ-contours to give a detailed picture of the sequence of shocks and expansion waves in a Guderley Mach reflection for $\kappa = 1$. Each shock-expansion pair in the sequence is smaller and weaker than the one preceding it. Three reflected shocks appear to be visible in the plot. From the numerical data, their approximate strengths, beginning with the leading reflected shock, are given in Table 1. The jump $[\rho]$ in ρ across a reflected shock is measured near the point where the flow behind the shock is sonic. This point is very close to the corresponding triple point on the Mach shock, as shown in Figure 3.

TABLE 1

Approximate values of the reflected shock strengths for the three reflected shocks visible in Figure 6, beginning with the leading reflected shock, from the numerical data. For each shock, ρ_1 and ρ_0 denote the approximate values of ρ ahead of and behind the shock, respectively.

Shock	ρ_1	ρ_0	$[\rho]$
1	64	76	12
2	72	75	3
3	74	75	1

5. Discussion. These numerical results are remarkably similar to the computed solutions of the shock reflection problem for the unsteady transonic small disturbance equations in [13]. In both cases, a weak shock reflection in a parameter range where regular reflection is impossible results in a sequence, possibly infinite, of triple points and supersonic patches embedded in the subsonic flow behind the Mach and reflected shocks. The unsteady transonic small disturbance equations can be derived from the full Euler equations by a systematic asymptotic expansion, and are considered to give an adequate description of the physical flow near the shock interaction point for weak shocks and small wedge angles. The nonlinear wave system, however, is not a systematic reduction of the Euler equations, and it does not appear to have any immediate physical relevance. It is therefore noteworthy that the shock reflection problem for the nonlinear wave system has a solution that resembles the solutions in [13], and is consistent with the experimental results in [12].

The nonlinear wave system has a characteristic structure similar to the two-dimensional Euler equations: nonlinear acoustic waves coupled (weakly) with linearly degenerate waves. The nonlinear wave system also respects the spatial (Euclidean) symmetries of gas dynamics, but not the space-time (Galilean) symmetry. In fact (see [10]), they are essentially the simplest system one can construct with these symmetries. The existence of a Guderley Mach reflection solution for a system that is only loosely related to gas dynamics suggests that the behavior may be typical of equations with this characteristic structure, and is not restricted to equations that describe gas dynamic phenomena. We conjecture that a sequence of supersonic patches and triple points is a generic feature of two-dimensional Riemann problems for some class of hyperbolic systems of conservation laws. Possibly this class is characterized by "acoustic waves," as defined in [2]. It is possible that numerical solutions of the weak shock reflection problem for the full Euler equations will contain a sequence of supersonic patches as well.

An important feature of the numerical solution is the small size of the supersonic region. In our solution for $\kappa = 1$, the width of the supersonic patch is approximately 5% of the length of the Mach shock. This is somewhat larger than the supersonic regions in the solutions in [13], which were obtained over a range of parameter values and varied in height from approximately 0.05% to 3% of the length of the Mach shock. Based on the dependence of patch size on wedge angle observed in [13], we expect solutions for larger values of κ to contain even larger supersonic regions. However, the strength of the reflected shock near the triple point decreases as κ increases, making it very difficult to resolve numerically the details of the solution near the triple point. We have displayed a solution with $\kappa = 1$ because it offers a good compromise between the size of the supersonic region and the strength of the sequence of reflected shocks and expansions.

One of the scenarios proposed in [3] for resolving the triple point paradox in the nonlinear wave system is that the reflected shock have zero strength at the shock

interaction point. In that case, there would be no triple point, and presumably no supersonic patch. We have obtained solutions, using different numerical schemes, which contain a supersonic region behind the triple point in a weak shock reflection. In these solutions, the reflected shocks have finite strength at the point where they collide with the Mach shock. Although we have not obtained numerical evidence of the zero strength reflected shock solution, we note that in the problem studied in [3], it is assumed that κ is large enough that the incident shock intersects the sonic circle, equation (2.8), corresponding to the state U_0 behind the shock. For shock reflection data with $\kappa = 1$, the incident shock does not intersect the sonic circle, so the partial solution presented in [3] is not available here. We also note, however, that in [13], several solutions were obtained in a parameter range for which the incident shock does intersect the sonic line for the state behind the incident shock. All of these solutions contained a reflected shock of nonzero strength at the triple point, and a supersonic region. For the nonlinear wave system, since we have obtained a solution containing a supersonic region at only one set of parameter values, we do not know if Guderley Mach reflection occurs over the entire set of parameter values for which regular reflection is impossible, or if solutions at large enough values of κ contain a reflected shock with zero strength at the triple point.

6. Conclusion. We have presented numerical evidence of a structure of reflected shocks and expansion waves, and a sequence of triple points and supersonic patches, in a small region behind the leading triple point in a shock reflection problem for the nonlinear wave system. This result is consistent with previous numerical solutions of a shock reflection problem for the unsteady transonic small disturbance equations, and with recent experimental results for weak shocks reflecting off thin wedges.

Appendix A. Symmetry. Equation (2.1) admits the usual Euclidean symmetries of gas dynamics (translation invariance and equivariance under rotation and reflection in the plane), but not the Galilean symmetry. For a polytropic gas law $p(\rho) = C\rho^\gamma$, where γ is the ratio of specific heats and C is a constant, it is also invariant under the scaling

$$(x, y) \mapsto \rho_1^{\frac{\gamma-1}{2}}(x, y), \qquad \rho \mapsto \rho_1\rho, \qquad (m, n) \mapsto \rho_1^{\frac{\gamma+1}{2}}(m, n).$$

Based on this, we see that solutions of the nonlinear wave system depend on the density only through a characteristic density ratio ρ_0/ρ_1, or equivalently, through the velocity ratio or the Mach number $M = c(\rho_0)/c(\rho_1) = (\rho_0/\rho_1)^{(\gamma-1)/2}$.

Appendix B. Nonexistence of triple points. To examine triple points in the nonlinear wave system we note, first, that this system does not have the Galilean invariance of the gas dynamics equations, so we cannot assume that the flow is stationary at a triple point. However, because of rotational symmetry we can assume that one of the shocks is horizontal, and we do so to simplify the calculation. We can also choose one set of momentum components to be zero.

We label the horizontal shock S_a, and proceeding counterclockwise, the other two are S_b and S_c (Figure 7). The state between S_a and S_b is $U_1 = (\rho_1, 0, 0)$; the other two states, also proceeding counterclockwise, are U_2 and $U_0 = (\rho_0, 0, n_0)$. The value of ρ_0 can be any number greater than ρ_1. Note that the component $m_0 = 0$ because S_a is horizontal. The equation of S_a is $\{\eta = \omega_a\}$, where $\omega_a = \sqrt{(p_0 - p_1)/(\rho_0 - \rho_1)}$, and $n_0 = \omega_a[\rho] = \sqrt{(p_0 - p_1)(\rho_0 - \rho_1)}$, using (2.3). (Note that $\kappa_a = \infty$ here and that U_0 corresponds to U_m in (3.3).)

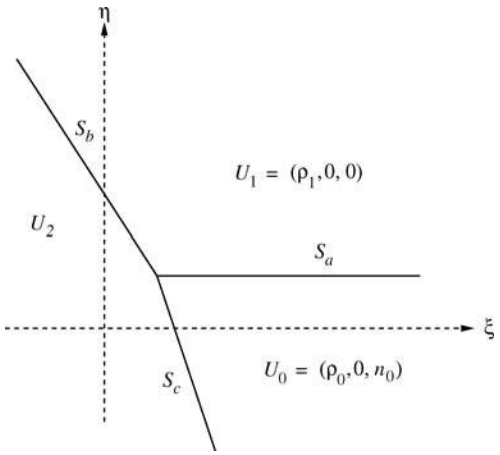


FIG. 7. *Triple point configuration.*

We introduce the notation $P_{ab} = \sqrt{(p_a - p_b)(\rho_a - \rho_b)}$. We have the following proposition.

PROPOSITION B.1. *For any convex equation of state $p(\rho)$, there is no nontrivial set of solutions to the Rankine–Hugoniot equations for constant states $\{U_0, U_1, U_2\}$ separated by shocks S_a, S_b, S_c, irrespective of whether the three shocks intersect in a point or not.*

Proof. The Rankine–Hugoniot equations at S_b and S_c imply (see [4, Appendix A])

$$m_2 - m_1 = \frac{P_{21}}{\sqrt{1+\kappa_b^2}}, \qquad m_2 - m_0 = \frac{P_{20}}{\sqrt{1+\kappa_c^2}},$$
$$n_2 - n_1 = -\frac{P_{21}}{\sqrt{1+\kappa_b^2}}\kappa_b, \quad n_2 - n_0 = -\frac{P_{20}}{\sqrt{1+\kappa_c^2}}\kappa_c.$$

Using the values $m_1 = m_0 = n_1 = 0, n_0 = P_{01}$, we get two equations:

$$m_2 = \frac{P_{21}}{\sqrt{1+\kappa_b^2}} = \frac{P_{20}}{\sqrt{1+\kappa_c^2}}, \quad n_2 = -\frac{P_{21}}{\sqrt{1+\kappa_b^2}}\kappa_b = P_{01} - \frac{P_{20}}{\sqrt{1+\kappa_c^2}}\kappa_c. \tag{B.1}$$

In principle, we can solve this pair of equations to obtain κ_b and κ_c as functions of the data ρ_0 and ρ_1. We get a 1-parameter family of solutions parameterized by ρ_2. However, the solutions obtained are not real numbers. For, if we substitute the first equation in (B.1) into the second, we get

$$-\kappa_b \frac{P_{20}}{\sqrt{1+\kappa_c^2}} = P_{01} - \kappa_c \frac{P_{20}}{\sqrt{1+\kappa_c^2}},$$

or

$$\kappa_b = \kappa_c - \sqrt{1+\kappa_c^2}\frac{P_{01}}{P_{20}}. \tag{B.2}$$

Now square the first relation in (B.1), write it as

$$1 + \kappa_b^2 = (1+\kappa_c^2)\frac{P_{21}^2}{P_{20}^2},$$

and substitute (B.2), to obtain

$$1+\left(\kappa_c-\sqrt{1+\kappa_c^2}\frac{P_{01}}{P_{20}}\right)^2=(1+\kappa_c^2)\frac{P_{21}^2}{P_{20}^2},$$

or

$$(P_{20}^2+P_{01}^2-P_{21}^2)\sqrt{1+\kappa_c^2}=2\kappa_c P_{01}P_{20}.$$

Square this and solve for κ_c^2:

$$\kappa_c^2=\frac{(P_{20}^2+P_{01}^2-P_{21}^2)^2}{4P_{01}^2P_{20}^2-(P_{20}^2+P_{01}^2-P_{21}^2)^2}.$$

Now, after a calculation,

$$P_{20}^2+P_{01}^2-P_{21}^2=(\rho_2-\rho_0)(p_1-p_0)+(\rho_1-\rho_0)(p_2-p_0).$$

So

$$4P_{01}^2P_{20}^2-(P_{20}^2+P_{01}^2-P_{21}^2)^2=-[(\rho_2-\rho_0)(p_1-p_0)-(\rho_1-\rho_0)(p_2-p_0)]^2\le 0.$$

In fact, this quantity is less than zero unless

$$\frac{p_2-p_0}{\rho_2-\rho_0}=\frac{p_1-p_0}{\rho_1-\rho_0}.$$

For a convex function p, this implies $\rho_2=\rho_1$, a degenerate case with only two distinct states.

Thus, no solutions exist. Note that the proof did not require the three shocks to intersect in a point, and that therefore this is a somewhat stronger result than simply the nonexistence of triple points. □

In a similar manner it is possible to show that, as in gas dynamics, a self-similar solution consisting of three shocks and a linear wave meeting at a point can be constructed. However, as mentioned in section 2, because of the invariance in time of the quantity (m_y-n_x), the linear characteristic coordinate, a linear wave cannot be present in the solution unless it is present in the data. For the data given in (2.2), therefore, solutions containing a triple point with a linear wave cannot exist. The same triple point paradox occurs in gas dynamics, of course, where solutions containing a triple point with a contact discontinuity cannot occur for sufficiently weak shocks (see [7] for a discussion of this).

REFERENCES

[1] M. Brio and J. K. Hunter, *Mach reflection for the two-dimensional Burgers equation*, Phys. D, 60 (1992), pp. 194–207.

[2] S. Čanić and B. L. Keyfitz, *Quasi-one-dimensional Riemann problems and their role in self-similar two-dimensional problems*, Arch. Ration. Mech. Anal., 144 (1998), pp. 233–258.

[3] S. Čanić, B. L. Keyfitz, and E. H. Kim, *Free boundary problems for nonlinear wave systems: Mach stems for interacting shocks*, SIAM J. Math. Anal., 37 (2006), pp. 1947–1977.

[4] S. Čanić, B. L. Keyfitz, and E. H. Kim, *Mixed hyperbolic-elliptic systems in self-similar flows*, Bol. Soc. Bras. Mat., 32 (2001), pp. 1–23.

[5] K. G. Guderley, *Considerations of the Structure of Mixed Subsonic-Supersonic Flow Patterns*, Air Materiel Command Tech. Report, F-TR-2168-ND, ATI 22780, GS-AAF-Wright Field No. 39, U.S. Wright-Patterson Air Force Base, Dayton, OH, 1947.

[6] K. G. Guderley, *The Theory of Transonic Flow*, Pergamon Press, Oxford, UK, 1962.
[7] L. F. Henderson, *Regions and boundaries for diffracting shock wave systems*, Z. Angew. Math. Mech., 67 (1987), pp. 73–86.
[8] J. K. Hunter and M. Brio, *Weak shock reflection*, J. Fluid Mech., 410 (2000), pp. 235–261.
[9] B. L. Keyfitz, home page, http://www.math.uh.edu/∼blk.
[10] B. L. Keyfitz and M. C. Lopes Filho, *A geometric study of shocks in equations that change type*, J. Dynam. Differential Equations, 6 (1994), pp. 351–393.
[11] J. von Neumann, *Collected Works*, Vol. 6, Pergamon Press, New York, 1963.
[12] B. Skews and J. Ashworth, *The physical nature of weak shock wave reflection*, J. Fluid Mech., 542 (2005), pp. 105–114.
[13] A. M. Tesdall and J. K. Hunter, *Self-similar solutions for weak shock reflection*, SIAM J. Appl. Math., 63 (2002), pp. 42–61.
[14] J. K. Hunter and A. M. Tesdall, *Weak shock reflection*, in A Celebration of Mathematical Modeling, D. Givoli, M. Grote, and G. Papanicolaou, eds., Kluwer Academic, New York, 2004.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 337–353

GLOBAL STABILITY FOR A VIRUS DYNAMICS MODEL WITH NONLINEAR INCIDENCE OF INFECTION AND REMOVAL*

PAUL GEORGESCU† AND YING-HEN HSIEH‡

Abstract. Global dynamics of a compartmental model which describes virus propagation in vivo is studied using the direct Lyapunov method, where the incidence rate of the infection and the removal rate of the virus are assumed to be nonlinear. In the case where the functional quotient between the force of infection and the removal rate of the virus is a nonincreasing function of the virus concentration, the existence of a threshold parameter, i.e., the basic reproduction number or basic reproductive ratio, is established and the global stability of the equilibria is discussed. Moreover, in the absence of the above-mentioned monotonicity property, estimations for the sizes of the domains of attraction are given. Biological significance of the results and possible extensions of the model are also discussed.

Key words. virus propagation, compartmental model, global stability, Lyapunov functional, endemic equilibrium, basic reproduction number

AMS subject classifications. 92D25, 92D30, 34D20, 34D23, 93D20

DOI. 10.1137/060654876

1. Introduction.

We consider a compartmental model for the propagation of a virus in vivo, in the form

$$\text{(S)} \qquad \begin{cases} S' = n(S) - c(S)f(V), \\ E' = c(S)f(V) - c_1 i(E), \\ I' = c_2 i(E) - c_3 p(I), \\ V' = c_4 p(I) - r(V). \end{cases}$$

Here, S denotes the concentration of the cells in the susceptible (i.e., uninfected) class, E denotes the concentration of cells in the exposed (i.e., latent) class, I denotes the concentration of cells in the infected class, and V denotes the concentration of the virus itself.

The intrinsic growth rate of the susceptible class, which includes both production of new cells and natural mortality of cells, is given by $n(S)$ with all the newly produced cells assumed to be susceptible. The movement of cells from the exposed class into the infected class and the production of free virus from infected cells are given by $c_2 i(E)$ and $c_4 p(I)$, respectively. By $c_1 i(E)$ and $c_3 p(I)$, we denote the removal of the exposed and infected classes, respectively, which include the mortality of cells in the above-mentioned classes.

It is assumed that the infection process is characterized by the incidence rate $c(S)f(V)$, where $c(S)$ denotes the contact function at concentration S and $f(V)$ denotes the force of infection by virus at concentration V. We note that our incidence

*Received by the editors March 23, 2006; accepted for publication (in revised form) August 31, 2006; published electronically December 21, 2006. This research was supported by National Science Council (NSC-Taiwan) research grant NSC-94-2115-M-005-006, which funded the first author's visit to National Chung Hsing University under an NSC research fellowship.
http://www.siam.org/journals/siap/67-2/65487.html
†Department of Mathematics, Technical University of Iaşi, Bd. Copou 11, 700506 Iaşi, Romania (vpgeo@go.com).
‡Department of Applied Mathematics, National Chung Hsing University, 250 Kuo-Kuang Rd, 402 Taichung, Taiwan, ROC (hsieh@amath.nchu.edu.tw).

rate is sufficiently general to encompass many forms of commonly used incidence rate, including simple mass action. The removal rate of the virus is denoted by $r(V)$. All functions c, f, i, p, r, n are allowed to be nonlinear and all constants c_1, c_2, c_3, c_4 are assumed to be positive.

We thereby assume that the major infection pathway is virus-to-cell, since the cell-to-cell pathway is sometimes less documented and therefore less considered, particularly in diseases such as AIDS (see Perelson and Nelson [18]).

While this model has been studied in Bonhoeffer et al. [1], Korobeinikov [7], Nowak and May [14], and Perelson and Nelson [18], among others, for linear c, f, i, p, r, n, it is perhaps important to account for a number of nonlinear features of the biological phenomena which are involved, especially for the nonlinearity of the incidence rate, which is influenced by the availability of susceptible cells and by the force of infection of viral cells. As the concentration of viral cells becomes higher, the simple mass action law βSV may not necessarily suffice. Moreover, the rate at which an infected cell or virus will die as a function of their concentrations is generally not known, and hence we make a further generalization by assuming that the removal rate is also nonlinear. For a detailed discussion on the virus dynamics of HIV, readers are referred to Perelson and Nelson [18].

We note that in (S), for $i(x) = x$ and $p(x) = x$, the constant $1/c_1$ represents the average time spent by a cell in the latent state, while $1/c_3$ represents the average lifetime of an infected cell. Also, $c_1 \geq c_2$ and $c_1 - c_2$ represents the mortality rate of the exposed cells, while c_4 relates to the production of virus from infected cells.

As noted by Korobeinikov in [7], if there is no exposed class E and consequently $c(S)f(V)$ represents the movement of cells from the susceptible class directly into the infected class, the (reduced three-dimensional) system (S) is equivalent to a SEIR model with a constant population assumption. It is therefore expected that the dynamics of our model will share some features with the dynamics of a SEIR model. Some perspectives and results from the global stability theory for SEIR models would also be relevant for our discussion. See Korobeinikov and Maini [8], Li et al. [11], Li and Muldowney [12], and Li, Muldowney, and van den Driessche [13] for global stability results for SEIR models. However, in [11, 12, 13] the approach is essentially geometrical, using a stability criteria which extends the Poincaré–Bendixson theorem and ruling out periodic orbits, rather than constructing a Lyapunov functional.

A related investigation pertaining to the dynamics of infectious disease models which incorporated nonlinear incidence rates of a very general form has recently been performed by Korobeinikov and Maini in [9] by using the Lyapunov method. In [9], the local stability of the equilibria for SIRS and SEIRS models has been considered assuming that the incidence rate is given by an arbitrary function $f(S, I, N)$, while the global stability of the equilibria for SIR and SEIR models has been considered assuming that the incidence rate is of the form $f(I)g(S)$. However, apart from the incidence rate, the other functions which appear in the models considered in [9] are linear and a constant population assumption is used, while for our model full nonlinearity is assured and a constant population assumption would not be an option. Moreover, the analysis performed in [9] is done in a somewhat different manner, with a focus on the role of the concavity of the nonlinear incidence rate in the existence and stability of the endemic equilibrium.

Substantial results regarding the global dynamics of a three-dimensional HIV model have been obtained by De Leenheer and Smith [3] using a different approach; their result distinguishes whether or not the term $-kVT$, which models the loss of a free virus particle once it enters the target cell, can be absorbed into the general loss

term $-\gamma V$. In [3], V is the concentration of free virus particles in the blood and T is the concentration of T cells. De Leenheer and Smith start with general assumptions on the function f which models T cell dynamics in a healthy individual and then specialize their results for two particular functions: $f_1(T) = \delta - \alpha T + pT(1 - T/T_{max})$ as used by Perelson and Nelson in [18] and $f_2(T) = \delta - \alpha T$ as used by Nowak and May in [14]. Certain linearity assumptions on some other functions appearing in the model are also made.

In the particular case in which the term $-kVT$ is absorbed into the general loss term (as done in [18] and [14])) and $f = f_2$, the model used in [3] can be thought of as a reduced version of our model, with no exposed class and extra linearity assumptions. However, the proof of our global stability result uses in an unavoidable manner the monotonicity assumption on n, which corresponds to f in [3], and therefore it can accommodate the case $f = f_2$ only and not the case $f = f_1$. In particular, our model does not admit orbitally asymptotically stable periodic solutions, which are obtained in [3] for $f = f_1$; see [3, Theorem 1] for details.

The paper is organized in the following manner. We propose the model to be studied in section 2 and discuss its well-posedness. In section 3 we give results on the stability of the disease-free equilibrium and persistence of the system, while sections 4 and 5 contain discussions on the existence, uniqueness, and global stability of the endemic equilibrium. Finally, in section 6, we give some remarks on the biological interpretation of our results, as well as some further extensions of the model one can make.

2. The model and its well-posedness. We assume that c, f, i, p, r are real locally Lipschitz functions defined at least on $[0, \infty)$ which satisfy

$$\begin{aligned} c(0) = f(0) = i(0) = p(0) = r(0) = 0, \\ c(t), f(t), i(t), p(t), r(t) > 0 \quad \text{for } t > 0 \end{aligned}$$

and that n is a real locally Lipschitz function defined at least on $[0, \infty)$ with $n(0) > 0$ such that the equation $n(S) = 0$ has a single solution S_0. We also assume that

$$\begin{aligned} (n(S) - n(S_0))(S - S_0) < 0 \quad \text{for } S \neq S_0, \\ (c(S) - c(S_0))(S - S_0) > 0 \quad \text{for } S \neq S_0 \end{aligned} \tag{2.1}$$

together with

$$\int_{0+}^{1} \frac{1}{\varphi(\tau)} d\tau = +\infty \quad \text{for all } \varphi \in \{c, f, i, p\}. \tag{D}$$

Note that (2.1) is satisfied if, for instance, n is strictly decreasing and c is strictly increasing. We also suppose that there are k_n, k_i, k_p, k_v, $\tilde{k}_n > 0$ such that

$$\begin{aligned} &n(S) \leq \tilde{k}_n - k_n S \text{ for } S \geq 0, \quad i(E) \geq k_i E \text{ for } E \geq 0, \quad p(I) \geq k_p I \text{ for } I \geq 0, \\ &r(V) \geq k_r V \text{ for } V \geq 0. \end{aligned} \tag{G}$$

The set of growth conditions (G) will be used to establish, in our general setting, the global existence of the solution for the Cauchy problem associated with the system (S). We note that these conditions may be dropped if the global existence property is known or the a priori boundedness of the solutions may be established by other methods. We shall indicate in section 6 how to remove conditions (G) at the expense

of other conditions on the behavior of c, f, i, p near $+\infty$ if f/r is nonincreasing on $(0, \infty)$.

First, it can be easily shown that a solution of the system (S) which starts in $[0, \infty)^4$ remains there on its whole interval of existence. To this purpose, we note that the vector (R_1, R_2, R_3, R_4) points inside $Q = [0, \infty)^4$ at all points of ∂Q, where R_1, R_2, R_3, and R_4 are the right-hand sides appearing in (S), and hence Nagumo's tangency conditions are satisfied. See [15] for details.

From our assumptions, it is clear that the system (S) has a unique saturated (i.e., nonextendable) solution for any initial data $(S(0), E(0), I(0), V(0))$. Using (G), it is possible to prove that all saturated solutions are global. To this aim, note that

$$\left(S + E + \frac{c_1}{2c_2} I + \frac{c_1 c_3}{4c_2 c_4} V\right)' \leq \tilde{k}_n - k_n S - \frac{c_1 k_i}{2} E - \frac{c_1 c_3}{4c_2} k_p I - \frac{c_1 c_3}{4c_2 c_4} k_r V,$$

it follows that there is $\delta = \delta(k_n, k_i, k_p, k_r, c_1, c_2, c_3, c_4) > 0$ small enough such that

$$\left(S + E + \frac{c_1}{2c_2} I + \frac{c_1 c_3}{4c_2 c_4} V\right)' + \delta \left(S + E + \frac{c_1}{2c_2} I + \frac{c_1 c_3}{4c_2 c_4} V\right) \leq \tilde{k}_n,$$

which implies that

$$\begin{aligned} &S + E + \frac{c_1}{2c_2} I + \frac{c_1 c_3}{4c_2 c_4} V - \frac{\tilde{k}_n}{\delta} \\ &\quad \leq \left(S(0) + E(0) + \frac{c_1}{2c_2} I(0) + \frac{c_1 c_3}{4c_2 c_4} V(0) - \frac{\tilde{k}_n}{\delta}\right) e^{-\delta t} \quad \text{for } t \geq 0, \end{aligned}$$

and therefore S, E, I, V are bounded on their maximal interval of existence. It follows that the functions $S(t)$, $E(t)$, $I(t)$, $V(t)$ are defined on $[0, \infty)$, and so the Cauchy problem with nonnegative initial data is well-posed for the system (S). Moreover, if we denote

$$F = \left\{(S, E, I, V) \in [0, \infty)^4; S + E + \frac{c_1}{2c_2} I + \frac{c_1 c_3}{4c_2 c_4} V \leq \frac{\tilde{k}_n}{\delta}\right\},$$

it follows that F is a feasible region for the system (S). Of course, the feasible region determined above is neither minimal nor unique, and the parameter δ above is obviously not uniquely determined. We shall simply choose

$$\delta = \min\left(k_n, \frac{c_1}{2} k_i, \frac{c_3}{2} k_p, k_r\right). \tag{2.2}$$

If S is small, then $S' = n(S) - c(S)F(V) > 0$ if V stays in a bounded set, since $n(0) > 0$ and $\lim_{S \to 0} c(S) = 0$, and we may infer that for any $S(0) > 0$ there is $\varepsilon_{S(0)} > 0$ such that $S(t) \geq \varepsilon_{S(0)}$ for all $t > 0$. This means that all solutions which start with positive $S(0)$ do not reach any point with $S = 0$ in future time. If $S(0) = 0$, then $S' > 0$ in a vicinity of 0 and, again, $S(t)$ raises over a certain minimum value (of course, the case in which $S(0) = 0$ does not make much biological sense). Also, it can be seen that the only w-limit point of (S) on the boundary of F is the disease-free equilibrium $(S_0, 0, 0, 0)$ and the only points on the boundary of $[0, \infty)^4$ which can be attained in finite time are situated on $[OS$, the positive S-semiaxis containing the origin.

3. Stability of disease-free equilibrium. Since the equation $n(S) = 0$ has a single solution S_0 and $f(0) = i(0) = p(0) = r(0) = 0$, it is easy to see that the system (S) admits a unique disease-free equilibrium $(S_0, 0, 0, 0)$. We now turn our attention to the study of its stability.

Consider the Lyapunov functional

$$U_1(S, E, I, V) = \int_{S_0}^{S} \frac{c(\tau) - c(S_0)}{c(\tau)} d\tau + E + \frac{c_1}{c_2} I + \frac{c_1 c_3}{c_2 c_4} V.$$

Since $(c(S) - c(S_0))(S - S_0) > 0$ for $S \neq S_0$, it is seen that U_1 increases whenever any of $|S - S_0|$, E, I, V increases and $U_1(S, E, I, V) \geq 0$ for all $S, E, I, V \geq 0$, while $U_1(S, E, I, V) = 0$ if and only if $(S, E, I, V) = (S_0, 0, 0, 0)$.

We now compute the time derivative of U_1 along the solutions of (S). It is seen that

$$\begin{aligned} \dot{U_1} = \left(1 - \frac{c(S_0)}{c(S)}\right)(n(S) - c(S)f(V)) + (c(S)f(V) - c_1 i(E)) \\ + \frac{c_1}{c_2}(c_2 i(E) - c_3 p(I)) + \frac{c_1 c_3}{c_2 c_4}(c_4 p(I) - r(V)), \end{aligned}$$

and since $n(S_0) = 0$, we can deduce that

$$\dot{U_1}(S, E, I, V) = \left(1 - \frac{c(S_0)}{c(S)}\right)(n(S) - n(S_0)) + \left[c(S_0)f(V) - \frac{c_1 c_3}{c_2 c_4} r(V)\right]. \tag{3.1}$$

Due to (2.1), it is easily seen that

$$\left(1 - \frac{c(S_0)}{c(S)}\right)(n(S) - n(S_0)) < 0 \quad \text{for } S \neq S_0, \tag{3.2}$$

and the first term in the right-hand side of (3.1) is negative. It is then seen that the stability of the disease-free equilibrium is related to the sign of the remaining term in the right-hand side of (3.1).

THEOREM 3.1. *Suppose that there is a number $V_R > 0$ such that*

$$c(S_0)\frac{f(V)}{r(V)}\frac{c_2 c_4}{c_1 c_3} \leq 1 \quad \textit{for } V \in (0, V_R), \tag{3.3}$$

and let $m = U_1(S_0, 0, 0, V_R)$. Then the disease-free equilibrium $(S_0, 0, 0, 0)$ is locally asymptotically stable and its domain of attraction includes the set

$$M_m = \left\{(S, E, I, V) \in (0, \infty) \times [0, \infty)^3; U_1(S, E, I, V) < m\right\}.$$

Proof. From (3.1), (3.2), and (3.3), it is seen that $\dot{U_1}(S, E, I, V) \leq 0$ for $0 \leq V < V_R$, with equality if and only if $S = S_0$ and either $V = 0$ or the equality in (3.3) holds.

Let us denote $\tilde{M} = \left\{(S, E, I, V) \in (0, \infty) \times [0, \infty)^3, 0 \leq V < V_R\right\}$ and take $k < m$ arbitrary. Since for all $V \geq V_R$ one has $U_1(S, E, I, V) \geq U_1(S_0, 0, 0, V_R)$, it is seen that $M_k \subset \tilde{M}$. Consequently, $U_1(S, E, I, V) \leq 0$ on M_k, with equality if and only if $S = S_0$ and the equality in (3.3) holds.

We now find the invariant subsets $\tilde{P}$ within the set

$$P = \left\{(S, E, I, V) \in M_k; \dot{U_1}(S, E, I, V) = 0\right\}.$$

Since $S = S_0$ on $\tilde{P}$ and consequently $S' = -c(S_0)f(V)$, it is seen that $V = 0$ and one similarly deduces that $E = I = 0$; that is, the only invariant subset of P is the singleton $\tilde{P} = \{(S_0, 0, 0, 0)\}$. From LaSalle's invariance principle (see LaSalle [10]) and the fact that $k < m$ was arbitrary, the conclusion follows. □

To complement Theorem 3.1, we further consider the case in which the disease-free equilibrium is unstable and give some remarks related to the persistence of the system. The system (S) is said to be *uniformly persistent* on F if there is a constant $\varepsilon_0 > 0$ such that any solution of (S) which starts in $(S(0), E(0), I(0), E(0)) \in \overset{\circ}{F}$ satisfies

$$\liminf_{t\to\infty} S(t) \geq \varepsilon_0, \quad \liminf_{t\to\infty} E(t) \geq \varepsilon_0, \quad \liminf_{t\to\infty} I(t) \geq \varepsilon_0, \quad \liminf_{t\to\infty} V(t) \geq \varepsilon_0.$$

See also Butler, Freedman, and Waltman [2] or Hofbauer and So [6].

Consider the Lyapunov function

$$U_2(S, E, I, V) = E + \frac{c_1}{c_2} I + \frac{c_1 c_3}{c_2 c_4} V.$$

Similar to the derivation of (3.1), the time derivative of U_2 along the solutions of (S) is given by

$$\dot{U}_2(S, E, I, V) = c(S)f(V) - \frac{c_1 c_3}{c_2 c_4} r(V). \tag{3.4}$$

Obviously, if (S) is uniformly persistent, then the disease remains endemic and stability for the disease-free equilibrium is excluded. In this regard, we have already observed that if (3.3) is satisfied on some interval $(0, V_R)$, then the disease-free equilibrium is locally asymptotically stable. If, on the other hand, the opposite of (3.3) is satisfied on some interval $(0, V_R)$, then the system (S) is uniformly persistent in the sense mentioned above.

THEOREM 3.2. *Assume that there is a number $V_R > 0$ such that*

$$c(S^0)\frac{f(V)}{r(V)}\frac{c_2 c_4}{c_1 c_3} > 1 \quad \text{for } V \in (0, V_R). \tag{3.5}$$

Then (S) *is uniformly persistent and the disease-free equilibrium* $(S_0, 0, 0, 0)$ *is unstable, with the positive semiaxis* $[OS$ *as its stable variety.*

Proof. From (3.4), (3.5), and the continuity of the function c at S_0, it follows that $\dot{U}_2 > 0$ on a small vicinity of $(S_0, 0, 0, 0)$, except for the points with $V = 0$. It then follows that any solution which starts in that vicinity remains away from $(S_0, 0, 0, 0)$, except for those starting on the positive semiaxis $[OS$, which tend to $(S_0, 0, 0, 0)$ while remaining on $[OS$. It may now be obtained, as in Proposition 3.3 in Li et al. [11], that the system (S) is uniformly persistent. This amounts to observing that $(S_0, 0, 0, 0)$ is the unique compact invariant set on the boundary of our feasible domain (so it is isolated) and its stable variety is the positive semiaxis $[OS$, which is contained in the boundary of the feasible domain. Then the use of Theorem 4.1 in Hofbauer and So [6], together with the remark that a flow and its time one map have the same maximal compact invariant set and the same stable set in a region, concludes the proof. □

It now remains to indicate some situations in which (3.3) or (3.5) are satisfied. Suppose for the moment that f/r is nonincreasing on $(0, \infty)$ and define a basic reproduction number R_0 of the system (S) by

$$R_0 = c(S_0)\frac{c_2 c_4}{c_1 c_3}\lim_{V\to 0}\frac{f(V)}{r(V)} \tag{3.6}$$

(note that the limit $\lim_{V\to 0}\frac{f(V)}{r(V)}$ does indeed exist, since f/r is monotone on $(0, \infty)$).

If $R_0 \leq 1$, then (3.3) is satisfied on $[0,\infty)$, while if $R_0 > 1$, then (3.5) is satisfied for V in a vicinity of 0. Also, it may be seen that $\lim_{V_R\to\infty} U_1(S_0,0,0,V_R) = +\infty$. One then obtains the following result, which establishes that R_0 is the threshold parameter for the stability of the disease-free equilibrium.

THEOREM 3.3. *Suppose that f/r is nonincreasing on $(0,\infty)$.*

1. *If $R_0 \leq 1$, then the disease-free equilibrium $(S_0,0,0,0)$ is globally asymptotically stable.*
2. *If $R_0 > 1$, then* (S) *is uniformly persistent and the disease-free equilibrium $(S_0,0,0,0)$ is unstable, with the positive semiaxis $[OS$ as its stable variety.*

In fact, if f/r is nonincreasing on $(0,\infty)$, more can be said for the case $R_0 > 1$, and it will be shown in sections 4 and 5 that, in this situation, the system (S) admits a positive endemic equilibrium, which is globally asymptotically stable.

We also note that if the functions f and r are of class C^1 and the limit $\lim_{V\to 0} \frac{f'(V)}{r'(V)}$ exists, then by the L'Hôpital theorem

$$R_0 = c(S_0)\frac{c_2c_4}{c_1c_3}\lim_{V\to 0}\frac{f'(V)}{r'(V)},$$

which is in agreement with the definition of the basic reproduction number given by van den Driessche and Watmough in [19] for a large class of compartmental models, including the present model. We do not need, however, to assume C^1 regularity for the functional coefficients throughout our proofs. We also note that, since no C^1 regularity is assumed, local stability analysis based on Jacobian matrices would fail.

4. Existence of endemic equilibrium. We now try to establish some sufficient conditions for the existence of the endemic equilibrium (S^*,E^*,I^*,V^*). Since it would be somehow unrealistic to attempt to solve the system (EQ) in its greatest generality, we impose some additional conditions on our functional coefficients. Let us suppose the following:

(4.1)
f/r is nonincreasing on $(0,\infty)$,

(4.2)
c, f, i, p are strictly increasing on $[0,\infty)$ and n is strictly decreasing on $[0,\infty)$,

(4.3)
$$\lim_{x\to\infty} i(x) = \lim_{x\to\infty} p(x) = +\infty.$$

Necessarily, $S^*,E^*,I^*,V^* > 0$, and the following equilibrium relations are satisfied:

$$\text{(EQ)}\qquad \begin{aligned} &n(S^*) = c(S^*)f(V^*), \quad c(S^*)f(V^*) = c_1 i(E^*), \quad c_2 i(E^*) = c_3 p(I^*),\\ &c_4 p(I^*) = r(V^*). \end{aligned}$$

To solve the equilibrium system (EQ), note first that from the last three equalities in (EQ) one obtains

$$c(S^*)f(V^*) = \frac{c_1c_3}{c_2c_4}r(V^*).$$

Let us define

$$F_1(S,V) = n(S) - c(S)f(V), \quad F_2(S,V) = c(S)f(V) - \frac{c_1c_3}{c_2c_4}r(V).$$

Since $S \mapsto F_1(S,V)$ is strictly decreasing and $F_1(0,V) \cdot F_1(S_0,V) < 0$ for all V, the equation $F_1(S,V) = 0$ can be uniquely solved with respect to S as a function of V for all V. That is, there is a function $S = \psi_1(V)$ which satisfies

$$\frac{n(\psi_1(V))}{c(\psi_1(V))} = f(V). \tag{4.4}$$

Since n/c is strictly decreasing and f is strictly increasing, it follows that ψ_1 is strictly decreasing. Note also that due to (4.4), $\lim_{V\to\infty} \psi_1(V) = 0$.

Similarly, $S \mapsto F_2(S,V)$ is strictly increasing and $F_2(0,V) < 0$ for all V. However, in this instance it is not necessarily true that $F_2(S_0,V) > 0$, and hence the same approach we used to solve the equation $F_2(S,V) = 0$ would not work. However, for our purpose we do not actually need the global solvability of the equation $F_2(S,V) = 0$, since we are searching for a unique endemic equilibrium and consequently for a single V^*. In some situations, local solvability may suffice.

To gain insight, suppose for the moment that the equation $F_2(S,V) = 0$ may also be uniquely solved with respect to S as a function of V (locally for V). That is, there is a function $S = \psi_2(V)$ which satisfies

$$c(\psi_2(V)) = \frac{c_1c_3}{c_2c_4}\frac{r(V)}{f(V)}.$$

Since c is strictly increasing, it follows that ψ_2 is strictly increasing.

Since ψ_1 is strictly decreasing, ψ_2 is strictly increasing and $\lim_{V\to\infty} \psi_1(V) = 0$, the curves defined by $S = \psi_1(V)$ and $S = \psi_2(V)$ have a common point (S^*, V^*) with $S^* > 0$ and $V^* > 0$ if and only if $\psi_1(0) > \psi_2(0)$, or equivalently, $c(\psi_1(0)) > c(\psi_2(0))$. Since $\psi_1(0) = S_0$ and $c(\psi_2(0)) = \frac{c_1c_3}{c_2c_4} \lim_{V\to 0} \frac{r(V)}{f(V)}$, the existence condition is $c(S_0) > \frac{c_1c_3}{c_2c_4} \lim_{V\to 0} \frac{r(V)}{f(V)}$. Using the basic reproduction number of the system (S) as defined in (3.6) (note again that f/r is monotone), this condition may be rewritten as $R_0 > 1$.

Up to now, we have shown that if the equation $F_2(S,V) = 0$ is solvable with respect to S as a function of V, then the necessary and sufficient condition for the existence of positive (S^*, V^*) is that $R_0 > 1$. In this case, we have

$$F_2(S,V) = \frac{c_1c_3}{c_2c_4} r(V) \left[c(S) \frac{c_2c_4}{c_1c_3} \frac{f(V)}{r(V)} - 1 \right];$$

and $F_2(S_0,V)$ is positive for V in a vicinity of 0. Since we have already noted that $F_2(0,V) < 0$ for all V, it follows that the equation $F_2(S,V) = 0$ is solvable with respect to S as a function of V (locally for V) if $R_0 > 1$, which is precisely what we needed. That is, we have shown that the existence of positive (S^*, V^*) is equivalent to the validity of condition $R_0 > 1$.

Also, if i, p are strictly increasing on $[0,\infty)$ and $\lim_{x\to\infty} i(x) = \lim_{x\to\infty} p(x) = +\infty$, then the equations $i(E) = \frac{1}{c_1} n(S^*)$ and $p(I) = \frac{c_2}{c_3c_1} n(S^*)$ will have unique positive solutions E^*, I^*, respectively. In view of the above, we can summarize our discussion with the following result.

THEOREM 4.1. *Assume that conditions* (4.1), (4.2), *and* (4.3) *are satisfied. Then there is a unique positive endemic equilibrium* (S^*, E^*, I^*, V^*) *of* (S) *if and only if* $R_0 > 1$, *where* R_0 *is the basic reproduction number for the system* (S), *as defined in* (3.6).

We note that conditions (4.1), (4.2), and (4.3) (combined with $R_0 > 1$) are sufficient for the existence of the endemic equilibrium but not necessary. Actually, if

one assumes that the removal rate $r(V)$ of the virus is influenced by treatment which is administered if an increase of the virus load over a certain value is observed, while the force of infection $f(V)$ is not, it is easy to think of a function f/r which is not monotone, for instance. In this situation, the disease-free equilibrium may coexist with multiple positive endemic equilibria. It is perhaps also worth noting that the stability of the equilibria depends essentially on the behavior of the function f/r and depends on the contact function c only through the basic reproduction number R_0.

5. Stability of endemic equilibrium. In this section we assume that the system (S) admits a positive endemic equilibrium (S^*, E^*, I^*, V^*) and study its stability. However, we do not assume that (4.1), (4.2), and (4.3) are satisfied and establish our results under somewhat weaker hypotheses. This is consistent with the remark that conditions (4.1), (4.2), and (4.3) are sufficient for the existence of the endemic equilibrium but not necessary. For our purpose, apart from the existence of the endemic equilibrium, we assume that

$$\text{(P)}\qquad \begin{aligned} (c(S)-c(S^*))\,(S-S^*) &> 0 \quad \text{for } S\neq S^*, S\geq 0,\\ (f(V)-f(V^*))\,(V-V^*) &> 0 \quad \text{for } V\neq V^*, V\geq 0,\\ (i(E)-i(E^*))\,(E-E^*) &> 0 \quad \text{for } E\neq E^*, E\geq 0,\\ (p(I)-p(I^*))\,(I-I^*) &> 0 \quad \text{for } I\neq I^*, I\geq 0 \end{aligned}$$

and

$$\text{(N)}\qquad (n(S)-n(S^*))\,(S-S^*)\leq 0 \quad \text{for all } S\geq 0.$$

Note that conditions (P) and (N) are satisfied if (4.2) holds. However, nonmonotone functions c, f, i, p, n can also satisfy (P) and (N).

We consider the Lyapunov function

$$\begin{aligned} U_3(S,E,I,V) = &\int_{S^*}^{S}\frac{c(\tau)-c(S^*)}{c(\tau)}d\tau + \int_{E^*}^{E}\frac{i(\tau)-i(E^*)}{i(\tau)}d\tau\\ &+\frac{c_1}{c_2}\int_{I^*}^{I}\frac{p(\tau)-p(I^*)}{p(\tau)}d\tau + \frac{c_1c_3}{c_2c_4}\int_{V^*}^{V}\frac{f(\tau)-f(V^*)}{f(\tau)}d\tau. \end{aligned}$$

Due to the sign conditions (P), it is seen that U_3 increases whenever any of $|S-S^*|$, $|E-E^*|$, $|I-I^*|$, $|V-V^*|$ increases and $U_3(S,E,I,V)\geq 0$ for all $S,E,I,V\geq 0$, while $U_3(S,E,I,V)=0$ if and only if $(S,E,I,V)=(S^*,E^*,I^*,V^*)$. We note that if any of S,E,I,V tends to 0, then $U_3(S,E,I,V)$ tends to ∞ due to the divergence condition (D). It then follows that all level sets of U_3 have no limit points on the boundary of $(0,\infty)^4$.

We now compute the time derivative of U_3 along the solutions of (S).

LEMMA 5.1. *The time derivative of U_3 with respect to the solutions of* (S) *is*

$$\begin{aligned} \dot{U_3}&(S,E,I,V)\\ &= (n(S)-n(S^*))\left(1-\frac{c(S^*)}{c(S)}\right) + c(S^*)r(V)\left(\frac{f(V^*)}{f(V)}-1\right)\left(\frac{f(V^*)}{r(V^*)}-\frac{f(V)}{r(V)}\right)\\ &\quad - c_1 i(E^*)\left[\frac{c(S^*)}{c(S)}+\frac{i(E^*)}{i(E)}\frac{c(S)}{c(S^*)}\frac{f(V)}{f(V^*)}+\frac{i(E)}{i(E^*)}\frac{p(I^*)}{p(I)}+\frac{f(V^*)}{f(V)}\frac{p(I)}{p(I^*)}-4\right]. \end{aligned}$$

If the inequality

$$c(S^*)r(V)\left(\frac{f(V^*)}{f(V)}-1\right)\left(\frac{f(V^*)}{r(V^*)}-\frac{f(V)}{r(V)}\right)\leq 0 \tag{5.1}$$

holds true for V in some given interval (V_L,V_R), then $\dot{U}_3(S,E,I,V)\leq 0$ for $V\in(V_L,V_R)$, with equality if and only if

$$S=S^* \quad \textit{and} \quad \frac{i(E)}{i(E^*)}=\frac{f(V)}{f(V^*)}=\frac{p(I)}{p(I^*)}.$$

Proof. It is seen that

$$\begin{aligned}
\dot{U}_3 &= \left(1-\frac{c(S^*)}{c(S)}\right)(n(S)-c(S)f(V))+\left(1-\frac{i(E^*)}{i(E)}\right)(c(S)f(V)-c_1i(E))\\
&\quad+\frac{c_1}{c_2}\left(1-\frac{p(I^*)}{p(I)}\right)(c_2i(E)-c_3p(I))+\frac{c_1c_3}{c_2c_4}\left(1-\frac{f(V^*)}{f(V)}\right)(c_4p(I)-r(V))\\
&= n(S)\left(1-\frac{c(S^*)}{c(S)}\right)+c(S^*)f(V)-\frac{i(E^*)}{i(E)}c(S)f(V)+c_1i(E^*)-c_1\frac{p(I^*)}{p(I)}i(E)\\
&\quad+\frac{c_1c_3}{c_2}p(I^*)-\frac{c_1c_3}{c_2c_4}r(V)-\frac{c_1c_3}{c_2}\frac{f(V^*)}{f(V)}p(I)+\frac{c_1c_3}{c_2c_4}\frac{f(V^*)}{f(V)}r(V).
\end{aligned}$$

Using the equilibrium relations (EQ), it follows that

$$\begin{aligned}
\dot{U}_3 &= n(S)\left(1-\frac{c(S^*)}{c(S)}\right)+c(S^*)f(V)-c_1i(E^*)\frac{i(E^*)}{i(E)}\frac{c(S)}{c(S^*)}\frac{f(V)}{f(V^*)}+c_1i(E^*)\\
&\quad-c_1i(E^*)\frac{i(E)}{i(E^*)}\frac{p(I^*)}{p(I)}+c_1i(E^*)-c_1i(E^*)\frac{r(V)}{r(V^*)}-c_1i(E^*)\frac{f(V^*)}{f(V)}\frac{p(I)}{p(I^*)}\\
&\quad+c_1i(E^*)\frac{f(V^*)}{f(V)}\frac{r(V)}{r(V^*)}\\
&= n(S)\left(1-\frac{c(S^*)}{c(S)}\right)+c(S^*)f(V)+c_1i(E^*)\left(\frac{f(V^*)}{f(V)}\frac{r(V)}{r(V^*)}-\frac{r(V^*)}{r(V)}\right)\\
&\quad-c_1i(E^*)\left[\frac{i(E^*)}{i(E)}\frac{c(S)}{c(S^*)}\frac{f(V)}{f(V^*)}+\frac{i(E)}{i(E^*)}\frac{p(I^*)}{p(I)}+\frac{f(V^*)}{f(V)}\frac{p(I)}{p(I^*)}-2\right]\\
&= n(S)\left(1-\frac{c(S^*)}{c(S)}\right)+c_1i(E^*)\frac{f(V)}{f(V^*)}+c_1i(E^*)\left(\frac{f(V^*)}{f(V)}\frac{r(V)}{r(V^*)}-\frac{r(V)}{r(V^*)}\right)\\
&\quad-c_1i(E^*)\left[\frac{c(S^*)}{c(S)}+\frac{i(E^*)}{i(E)}\frac{c(S)}{c(S^*)}\frac{f(V)}{f(V^*)}+\frac{i(E)}{i(E^*)}\frac{p(I^*)}{p(I)}+\frac{f(V^*)}{f(V)}\frac{p(I)}{p(I^*)}-4\right]\\
&\quad+c_1i(E^*)\frac{c(S^*)}{c(S)}-2c_1i(E^*).
\end{aligned}$$

This implies that

$$\begin{aligned}
\dot{U}_3 &= (n(S)-c_1i(E^*))\left(1-\frac{c(S^*)}{c(S)}\right)\\
&\quad+c_1i(E^*)\left(\frac{f(V^*)}{f(V)}\frac{r(V)}{r(V^*)}-\frac{r(V)}{r(V^*)}+\frac{f(V)}{f(V^*)}-1\right)\\
&\quad-c_1i(E^*)\left[\frac{c(S^*)}{c(S)}+\frac{i(E^*)}{i(E)}\frac{c(S)}{c(S^*)}\frac{f(V)}{f(V^*)}+\frac{i(E)}{i(E^*)}\frac{p(I^*)}{p(I)}+\frac{f(V^*)}{f(V)}\frac{p(I)}{p(I^*)}-4\right],
\end{aligned}$$

and since $c_1 i(E^*) = n(S^*)$, it follows that

$$\begin{aligned}\dot{U_3}&(S,E,I,V)\\ &= (n(S)-n(S^*))\left(1-\frac{c(S^*)}{c(S)}\right)+c_1 i(E^*)\left(\frac{f(V^*)}{f(V)}-1\right)\left(\frac{r(V)}{r(V^*)}-\frac{f(V)}{f(V^*)}\right)\\ &\quad - c_1 i(E^*)\left[\frac{c(S^*)}{c(S)}+\frac{i(E^*)}{i(E)}\frac{c(S)}{c(S^*)}\frac{f(V)}{f(V^*)}+\frac{i(E)}{i(E^*)}\frac{p(I^*)}{p(I)}+\frac{f(V^*)}{f(V)}\frac{p(I)}{p(I^*)}-4\right].\end{aligned}$$

Using the relation $c_1 i(E^*) = c(S^*)f(V^*)$, one gets the required conclusion. Now, from the sign condition (N) it is seen that

$$(n(S)-n(S^*))\left(1-\frac{c(S^*)}{c(S)}\right)\le 0 \quad \text{for } S\ge 0,$$

with equality if and only if $S = S^*$, and from the *AM-GM* inequality (which says that the algebraic mean is not smaller than the geometric mean) it is seen that

$$\frac{c(S^*)}{c(S)}+\frac{i(E^*)}{i(E)}\frac{c(S)}{c(S^*)}\frac{f(V)}{f(V^*)}+\frac{i(E)}{i(E^*)}\frac{p(I^*)}{p(I)}+\frac{f(V^*)}{f(V)}\frac{p(I)}{p(I^*)}\ge 4,$$

with equality if and only if

$$\frac{c(S^*)}{c(S)}=\frac{i(E^*)}{i(E)}\frac{c(S)}{c(S^*)}\frac{f(V)}{f(V^*)}=\frac{i(E)}{i(E^*)}\frac{p(I^*)}{p(I)}=\frac{f(V^*)}{f(V)}\frac{p(I)}{p(I^*)}=1. \tag{5.2}$$

It then follows that if the inequality

$$c(S^*)r(V)\left(\frac{f(V^*)}{f(V)}-1\right)\left(\frac{f(V^*)}{r(V^*)}-\frac{f(V)}{r(V)}\right)\le 0$$

holds true for $v\in(V_L,V_R)$, then $\dot{U_3}(S,E,I,V)\le 0$. For the equality case, we note that $c(S^*) = c(S)$ if and only if $S = S^*$, and substituting this into (5.2) one obtains that

$$\frac{i(E)}{i(E^*)}=\frac{f(V)}{f(V^*)}=\frac{p(I)}{p(I^*)}. \qquad \square$$

It is now obvious that the stability of the endemic equilibrium (S^*,E^*,I^*,V^*) is related to the validity of the inequality (5.1). Subsequently, we estimate the size of the domain of attraction associated with (S^*,E^*,I^*,V^*).

THEOREM 5.2. *Assume that the sign conditions* (P) *and* (N) *are satisfied and there are V_L and V_R such that*

$$\begin{aligned}&\frac{f(V)}{r(V)}\le\frac{f(V^*)}{r(V^*)} \quad \textit{for } V^*\le V<V_R,\\ &\frac{f(V)}{r(V)}\ge\frac{f(V^*)}{r(V^*)} \quad \textit{for } V_L<V\le V^*.\end{aligned} \tag{5.3}$$

Define $m = \min\left(U_3(S^*,E^*,I^*,V_L), U_3(S^*,E^*,I^*,V_R)\right)$. Then (S^*,E^*,I^*,V^*) is locally asymptotically stable and its domain of attraction includes the set

$$M_m=\left\{(S,E,I,V)\in(0,\infty)^4; U_3(S,E,I,V)<m\right\}.$$

Proof. Denote

$$\tilde{M} = \left\{(S, E, I, V) \in (0, \infty)^4; V_L < V < V_R\right\}.$$

From (5.3) it follows that (5.1) is satisfied for $V \in (V_L, V_R)$, and using Lemma 5.1 one may infer that $\dot{U}_3(S, E, I, V) \leq 0$ on $\tilde{M}$, with equality if and only if

$$S = S^* \quad \text{and} \quad \frac{i(E)}{i(E^*)} = \frac{f(V)}{f(V^*)} = \frac{p(I)}{p(I^*)}.$$

Take an arbitrary $k < m$. Since U_3 increases whenever any of $|S - S^*|$, $|E - E^*|$, $|I - I^*|$, $|V - V^*|$ increases, it follows easily that, for all V outside (V_L, V_R), one has $U_3(S, E, I, V) \geq m$ for all $S, E, I > 0$. Consequently $M_k \subset \tilde{M}$. Moreover, as noted previously, M_k is a bounded set which has no limit points on the boundary of $\tilde{M}$.

We now find the invariant subsets $\tilde{N}$ within the set

$$N = \left\{(S, E, I, V) \in M_k; \dot{U}_3(S, E, I, V) \leq 0\right\}.$$

Since $S = S^*$ on $\tilde{N}$ and consequently $S' = n(S^*) - c(S^*)f(V)$, it follows that $S' = c(S^*)(f(V^*) - f(V))$, and so $S' = 0$ if and only if $V = V^*$. From $\frac{i(E)}{i(E^*)} = \frac{p(I)}{p(I^*)} = 1$ we then deduce that $E = E^*$ and $I = I^*$ by using the sign condition (P).

Therefore, using LaSalle's invariance principle (see LaSalle [10]) one obtains that any trajectory which starts in M_k tends to (S^*, E^*, I^*, V^*) as $t \to \infty$. Then the endemic equilibrium (S^*, E^*, I^*, V^*) is locally asymptotically stable and the set M_k belongs to its domain of attraction. Since k was arbitrary and less than m, one obtains the required conclusion. □

We now continue with a few considerations on the inequalities (5.3). Since

$$\lim_{V_L \to 0} U_3(S^*, E^*, I^*, V_L) = \lim_{V_R \to \infty} U_3(S^*, E^*, I^*, V_R) = +\infty,$$

one obtains that if the following inequalities are satisfied,

$$\begin{aligned} &\frac{f(V)}{r(V)} \leq \frac{f(V^*)}{r(V^*)} \quad \text{for } V^* \leq V, \\ &\frac{f(V)}{r(V)} \geq \frac{f(V^*)}{r(V^*)} \quad \text{for } 0 < V \leq V^*, \end{aligned} \tag{5.4}$$

then (S^*, E^*, I^*, V^*) is globally asymptotically stable in $(0, \infty)^4$.

Regarding the inequalities (5.4) (or (5.3)), it is easy to see that they are verified if the function f/r is nonincreasing on $(0, \infty)$ (or on (V_L, V_R)); however, this monotonicity property is only sufficient and not necessary. If $r(V) = kV$, for some k, then the above monotonicity property is satisfied for three common incidence rates, namely $c_1(S)f_1(V) = \beta_1 SV$, $c_2(S)f_2(V) = \beta_2 S^p V^q$, where $0 < q \leq 1$, and $c_3(S)f_3(V) = \beta_3 SV/(1 + a_1 V)$.

We also remark that the inequalities (5.4) alone imply the uniqueness of the endemic equilibrium (S^*, E^*, I^*, V^*). To show this, suppose that there is another endemic equilibrium $(S_1^*, E_1^*, I_1^*, V_1^*)$. Apart from (EQ), one then has

$$\begin{aligned} (\text{EQ}') \qquad & n(S_1^*) = c(S_1^*)f(V_1^*), \quad c(S_1^*)f(V_1^*) = c_1 i(E_1^*), \quad c_2 i(E_1^*) = c_3 p(I_1^*), \\ & c_4 p(I_1^*) = r(V_1^*). \end{aligned}$$

It follows that

$$c(S^*) - c(S_1^*) = \frac{c_1 c_3}{c_2 c_4}\left(\frac{r(V^*)}{f(V^*)} - \frac{r(V_1^*)}{f(V_1^*)}\right), \tag{5.5}$$

$$n(S^*) - n(S_1^*) = \frac{c_1 c_3}{c_2 c_4}\left(r(V^*) - r(V_1^*)\right) \tag{5.6}$$

and therefore

$$\left(c(S^*) - c(S_1^*)\right)\left(V^* - V_1^*\right) \geq 0.$$

If $V^* > V_1^*$, then, from (5.5), $c(S^*) \geq c(S_1^*)$, $S^* \geq S_1^*$, which implies $n(S^*) \leq n(S_1^*)$ and consequently from (5.6), $r(V^*) \leq r(V_1^*)$, which is a contradiction. The case $V^* < V_1^*$ is dismissed in a similar manner, subsequently $V^* = V_1^*$ and from (5.5), $S = S_1^*$. Substituting these equalities into (EQ) and (EQ′) we obtain that $i(E^*) = i(E_1^*)$ and $p(I^*) = p(I_1^*)$, and hence $E^* = E_1^*$ and $I^* = I_1^*$; that is, the endemic equilibrium is uniquely determined. However, we should point out that inequalities (5.4) ensure the uniqueness of the endemic equilibrium only and not necessarily its existence.

6. Discussions and concluding remarks. The earlier analysis clearly indicates the importance of the quantity

$$c(S_0)\frac{f(V)}{r(V)}\frac{c_2 c_4}{c_1 c_3}$$

in the discussion on local stability of the disease-free equilibrium and persistence for the system. Moreover, under the monotonicity condition on $f(V)/r(V)$, we obtain the basic reproduction number

$$R_0 = c(S_0)\frac{c_2 c_4}{c_1 c_3}\lim_{V\to 0}\frac{f(V)}{r(V)}. \tag{6.1}$$

We will now give a biological interpretation of this result. From (S), it is obvious that the terms in the numerator denote the growth in the concentrations of the infected cells, E and I, and of the virus V. The terms in the denominator, on the other hand, denote the removal (or decrease in concentration) of these three same classes. Therefore, the ratio of the two can be considered as a measurement of the combined "productivity," perhaps more aptly, the *basic reproductive ratio* of the infected classes in the system. The fact that the stability of the disease-free equilibrium and the persistence of the system depend on whether this quantity is less than one or not (Theorems 3.1 and 3.2) further confirms our assertion.

The quantity $f(V)/r(V)$ is also important for our results. It can be interpreted as the efficiency of the virus, that is, the ratio of its infectivity to its removal, as a function of the virus concentration. Theorems 3.3, 4.1, and 5.2 require $f(V)/r(V)$ to be a nonincreasing function of V. Some recent studies (see, e.g., [16, 17]) let $f(V) = r(V) = V$, an assumption which is supported by some clinical data. We note that in this case $f(V)/r(V) = 1$, and hence our condition of nonincreasing ratio $f(V)/r(V)$, which generalizes to the models with nonlinear $f(V)$ and $r(V)$, is satisfied. For HIV, it has been observed that the productivity of the virus, $f(V)$, increases as the virus concentration increases. Our analysis is valid if the increase in removal of the virus $r(V)$ as virus concentration increases is at least to the same level as the increase in $f(V)$. Further studies are needed to verify whether our assertion holds.

On the other hand, if the function f/r is indeed increasing on $(0,\infty)$, then U_1 and U_3 are not necessarily global Lyapunov functionals and therefore do not create their own boundedness structure for the solutions of (S). For the global existence of the solutions, growth conditions (G) (see section 2) need to be imposed. If f/r is nonincreasing on $(0,\infty)$, however, the boundedness structures created by the level sets of U_1 and U_3 render the growth conditions unnecessary.

Suppose that f/r is nonincreasing on $(0,\infty)$ and $R_0 > 1$. Assume that the following conditions are satisfied:

$$\text{(B)} \qquad \lim_{y\to\infty}\left(y-\varphi(x)\int_x^y \frac{1}{\varphi(\tau)}d\tau\right)=+\infty \quad \text{for all } x>0 \text{ and } \varphi\in\{c,f,i,p\}\,.$$

Note that (B) is satisfied for a function φ such that $\lim_{y\to\infty}\varphi(y)=+\infty$, since in this situation

$$\lim_{y\to\infty}\frac{\int_x^y \frac{1}{\varphi(\tau)}d\tau}{y}=\lim_{y\to\infty}\frac{1}{\varphi(y)}=0 \quad \text{for } \varphi\in\{c,f,i,p\}\,.$$

However, condition (B) is also satisfied for $\varphi(x)=x^p/(1+ax^p), 0<p\le 1$ (this is, for instance, the case when $\varphi(V)=f(V)=V^p/(1+aV^p)$ is a nonlinear force of infection with saturation), which does not tend to $+\infty$ as $x\to+\infty$.

Regarding conditions (D), since the only points on the boundary of $[0,\infty)^4$ which can be reached in finite time are situated on $[OS$ and the only w-limit point of (S) on the boundary of $[0,\infty)^4$ is the disease-free equilibrium $(S_0,0,0,0)$, a less restrictive condition than (D) would suffice to avoid these situations, namely

$$\text{(D}')\qquad \int_{0+}^1 \frac{1}{\varphi(\tau)}d\tau=+\infty \quad \text{for some } \varphi\in\{f,i,p\}\,.$$

Then, by the results in the previous section, there is a unique positive endemic equilibrium which verifies relations (EQ). Take $(S(0),E(0),I(0),V(0))\in(0,\infty)^4$. Then $U_3\le 0$ for all t, and it follows that $(S(t),E(t),I(t),V(t))$ stays in a level set of U_3 on its whole interval of existence. Since the level sets of U_3 are bounded due to (B), it follows that the saturated solution which starts in $(S(0),E(0),I(0),V(0))$ exists on $[0,\infty)$. The growth conditions (G), which were used to obtain global existence, therefore become unnecessary and the proof proceeds in the same manner. Then, as in section 3, all solutions which start in $[0,\infty)^4$ tend to (S^*,E^*,I^*,V^*), except for those which start on $[OS$ and tend to $(S_0,0,0,0)$ as $t\to\infty$. The growth conditions become unnecessary for the proof of the uniform persistence result as well, since the system (S) admits an endemic equilibrium and it is obviously uniformly persistent.

If $R_0\le 1$, the reasoning is quite similar, with U_1 in place of U_3, and it is obtained again that all the saturated solutions are global and the stability result remains valid. We then summarize our discussion in the following result.

THEOREM 6.1. *Suppose that f/r is nonincreasing on $(0,\infty)$ and conditions* (4.2), (4.3), (B), *and* (D′) *are satisfied.*

1. *If $R_0\le 1$, then the disease-free equilibrium $(S_0,0,0,0)$ is globally asymptotically stable.*
2. *If $R_0>1$, then the system* (S) *admits a unique positive endemic equilibrium which is globally asymptotically stable. The disease-free equilibrium $(S_0,0,0,0)$ is unstable, with the positive semiaxis $[OS$ as its stable variety.*

Obviously, in statement 2 the stable variety of the endemic equilibrium actually excludes $[OS$.

As an example to illustrate the usefulness of our results, it is easy to see that a system which fits into our framework is

$$\text{(RS)} \qquad \begin{cases} S' = b - mS - \beta S \dfrac{V^p}{1 + aV^p}, \\ E' = \beta S \dfrac{V^p}{1 + a_1 V^p} - c_1 E, \\ I' = c_2 E - c_3 I, \\ V' = c_4 I - kV^\gamma \end{cases}$$

for $b, m, \beta, k > 0$, $a \geq 0$, and $0 < p \leq \gamma \leq 1$. In this situation, $c(S) = \beta S$, $f(V) = V^p/(1 + aV^p)$, $i(E) = E$, $p(I) = I$, $r(V) = V^\gamma$, $n(S) = b - mS$.

It follows that $f/r = 1/((1 + a_1 V^p)V^{\gamma - p})$ is nonincreasing on $(0, \infty)$,

$$\lim_{E \to \infty} E = \lim_{I \to \infty} I = \lim_{V \to \infty} kV^\gamma = +\infty,$$

and $\lim_{V \to \infty} V^p/(1 + aV^p) = +\infty$ if $a = 0$, while if $a > 0$, then

$$\lim_{V \to \infty} \left(V - \frac{x^p}{1 + ax^p} \int_x^V \frac{1 + a\tau^p}{\tau^p} d\tau \right) = +\infty \quad \text{for all } x > 0.$$

Also, $\int_{0+}^1 \frac{1}{E} dE = +\infty$. Note that if $a = 0$ and $p \in (0, 1)$, then $f(V) = V^p$ is not Lipschitzian on $[0, \infty)$ due to its behavior near 0. However, our solutions which start with $V > 0$ do not reach points for which $V = 0$ in finite time. Hence the uniqueness property is not impaired. The same remark applies to the function r. We can therefore apply the results in the previous sections and obtain the following result.

THEOREM 6.2.

1. *If $p < \gamma$, the basic reproduction number R_0 of the system* (RS) *is $+\infty$. The system* (RS) *admits a positive endemic equilibrium which is globally asymptotically stable. The disease-free equilibrium $(S_0, 0, 0, 0)$ is unstable, with the positive semiaxis $[OS$ as its stable variety.*
2. *If $p = \gamma$, the basic reproduction number R_0 of the system* (RS) *is*

$$R_0 = \frac{\beta b}{m} \frac{c_2 c_4}{c_1 c_3} \frac{1}{k}.$$

In this case, if $R_0 \leq 1$, then the disease-free equilibrium $(S_0, 0, 0, 0)$ is globally asymptotically stable, while if $R_0 > 1$, the system (RS) *admits a positive endemic equilibrium which is globally asymptotically stable. The disease-free equilibrium $(S_0, 0, 0, 0)$ is unstable, with the positive semiaxis $[OS$ as its stable variety.*

Again, the "global" stable variety of the endemic equilibrium is understood to exclude $[OS$. Note that for $p = \gamma = 1$ and $a = 0$ we obtain the results given in Korobeinikov [7].

As a final remark, we note that similar analysis can be extended to a system of the form

$$\text{(SE)}\qquad \begin{cases} S' = n(S) - c(S)f(V), \\ E' = c(S)f(V) - c_1 i(E), \\ I_1' = c_2 i(E) - k_1 p_1(I_1), \\ I_j' = \tilde{k}_{j-1} p_{j-1}(I_{j-1}) - k_j p_j(I_j), \quad 2 \le j \le n, \\ V' = \tilde{k}_n p_n(I_n) - r(V). \end{cases}$$

The associated Lyapunov functionals are in this case

$$U_1(S, E, I_1, \dots, I_n) = \int_{S_0}^{S} \frac{c(\tau) - c(S_0)}{c(\tau)} d\tau + E + \frac{c_1}{c_2} \sum_{i=1}^{n} \left(\prod_{j=1}^{i-1} \frac{k_j}{\tilde{k}_j} \right) I_i + \frac{c_1}{c_2} \prod_{j=1}^{n} \frac{k_j}{\tilde{k}_j} V,$$

$$U_2(S, E, I_1, \dots, I_n) = E + \frac{c_1}{c_2} \sum_{i=1}^{n} \left(\prod_{j=1}^{i-1} \frac{k_j}{\tilde{k}_j} \right) I_i + \frac{c_1}{c_2} \prod_{j=1}^{n} \frac{k_j}{\tilde{k}_j} V,$$

and

$$\begin{aligned} U_3(S, E, I_1, \dots, I_n) = {} & \int_{S^*}^{S} \frac{c(\tau) - c(S^*)}{c(\tau)} d\tau + \int_{E^*}^{E} \frac{i(\tau) - i(E^*)}{i(\tau)} d\tau \\ & + \frac{c_1}{c_2} \sum_{i=1}^{n} \left(\prod_{j=1}^{i-1} \frac{k_j}{\tilde{k}_j} \right) \int_{I_i^*}^{I_i} \frac{p_i(\tau) - p_i(I_i^*)}{p_i(\tau)} d\tau \\ & + \frac{c_1}{c_2} \left(\prod_{j=1}^{n} \frac{k_j}{\tilde{k}_j} \right) \int_{V^*}^{V} \frac{c(\tau) - c(V^*)}{c(\tau)} d\tau, \end{aligned}$$

with the convention $\prod_{j=1}^{0} \frac{k_j}{\tilde{k}_j} = 1$.

Again, related asymptotic stability can be obtained as in previous sections, and the size of the domain of attraction depends essentially on the behavior of the function f/r. If the function f/r is nonincreasing on $(0, \infty)$, the threshold parameter R_0 is given by

$$R_0 = c(S_0) \frac{c_2}{c_1} \left(\prod_{j=1}^{n} \frac{\tilde{k}_j}{k_j} \right) \lim_{V \to 0} \frac{f(V)}{r(V)}.$$

The first Lyapunov functional of type $\sum_{i=1}^{n} d_i \big(x_i - x_i^* - x_i^* \ln \frac{x_i}{x_i^*} \big)$, to which our functional U_3 reduces when c, f, i, p are linear functions, has been used by Volterra in [20] to treat a two-dimensional predator-prey model which describes the interaction between sharks and predated fish in the Mediterranean Sea. (See also Goh [4].) In [5], Harrison constructed a Lyapunov functional of this type for a two-dimensional predator-prey model which accounted for very general numerical and functional responses of the predator. The computation of the derivatives is straightforward and hence omitted for brevity.

Acknowledgment. The authors would like to thank the referees for their constructive comments.

REFERENCES

[1] S. Bonhoeffer, R. M. May, G. M. Shaw, and M. A. Nowak, *Virus dynamics and drug therapy*, Proc. Nat. Acad. Sci. U.S.A., 94 (1997), pp. 6971–6976.
[2] G. Butler, H. I. Freedman, and P. Waltman, *Uniformly persistent systems*, Proc. Amer. Math. Soc., 96 (1986), pp. 425–430.
[3] P. De Leenheer and H. L. Smith, *Virus dynamics: A global analysis*, SIAM J. Appl. Math., 63 (2003), pp. 1313–1327.
[4] B. S. Goh, *Global stability in many species systems*, Amer. Natur., 111 (1977), pp. 135–143.
[5] G. W. Harrison, *Global stability of predator-prey interactions*, J. Math. Biol., 8 (1979), pp. 159–171.
[6] J. Hofbauer and J. W. H. So, *Uniform persistence and repellors for maps*, Proc. Amer. Math. Soc., 107 (1989), pp. 1137–1142.
[7] A. Korobeinikov, *Global properties of basic virus dynamics models*, Bull. Math. Biol., 66 (2004), pp. 879–883.
[8] A. Korobeinikov and P. K. Maini, *A Lyapunov function and global properties for SIR and SEIR epidemiological models with nonlinear incidence*, Math. Biosci. Eng., 1 (2004), pp. 57–60.
[9] A. Korobeinikov and P. K. Maini, *Non-linear incidence and stability of infectious disease models*, Math. Med. Biol., 22 (2005), pp. 113–128.
[10] J. P. LaSalle, *The Stability of Dynamical Systems*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 25, SIAM, Philadelphia, 1976.
[11] M. Y. Li, J. R. Graef, L. K. Wang, and J. Karsai, *Global dynamics of a SEIR model with a varying total population size*, Math. Biosci., 160 (1999), pp. 191–213.
[12] M. Y. Li and J. S. Muldowney, *Global stability for the SEIR model in epidemiology*, Math. Biosci., 125 (1995), pp. 155–164.
[13] M. Y. Li, J. S. Muldowney, and P. van den Driessche, *Global stability for SEIRS models in epidemiology*, Can. Appl. Math. Q., 7 (1999), pp. 409–425.
[14] M. A. Nowak and R. M. May, *Virus Dynamics: Mathematical Principles of Immunology and Virology*, Oxford University Press, New York, 2000.
[15] N. H. Pavel, *Differential equations, flow invariance and applications*, Pitman Research Notes in Mathematics, Boston, MA, 1984.
[16] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho, *HIV*-1 *dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time*, Science, 271 (1996), pp. 1582–1586.
[17] A. S. Perelson, *Modelling viral and immune system dynamics*, Nat. Rev. Immunol., 2 (2002), pp. 28–36.
[18] A. S. Perelson and P. W. Nelson, *Mathematical analysis of HIV*-1 *dynamics in vivo*, SIAM Rev., 41 (1999), pp. 3–44.
[19] P. van den Driessche and J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci., 180 (2002), pp. 29–48.
[20] V. Volterra, *Leçons sur la Theorie Mathematique de la Lutte pour la Vie*, Gauthier-Villars, Paris, 1931.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 354–363

SUFFICIENT CONDITIONS FOR MONOTONE LINEAR STABILITY OF STEADY AND OSCILLATORY HAGEN–POISEUILLE FLOW*

VÍT PRŮŠA†

Abstract. Sufficient conditions for the monotone decay of disturbances to oscillatory and steady Hagen–Poiseuille flow are rigorously derived within the framework of linear stability theory. The conditions hold both for axisymmetric and nonaxisymmetric disturbances, whereas the result for nonaxisymmetric disturbances to the oscillatory flow is of particular importance because in this case even numerical results are not available in the literature. Furthermore, the conditions provide explicit bounds on the range of parameters that must be examined in any prospective search for instability by numerical means. The derivation of the sufficient conditions on monotone decay is based on a detailed analysis of the spectrum of the Stokes operator.

Key words. linear stability, nonaxisymmetric disturbances, Hagen–Poiseuille flow, oscillatory Hagen–Poiseuille flow

AMS subject classifications. 76E05, 76Z05

DOI. 10.1137/060652506

1. Introduction.

Hagen–Poiseuille flow is a flow of incompressible Newtonian fluid in a pipe of infinite length that is driven by pressure gradient in the direction of the pipe's axis. For steady flow the gradient is given by the formula

$$\frac{\partial p}{\partial z} = -\Delta_s,$$

where Δ_s is constant. In the oscillatory case the pressure gradient is a harmonic function of time

$$\frac{\partial p}{\partial z} = -\Delta_p e^{i\omega t}, \tag{1.1}$$

where Δ_p denotes the amplitude of the pressure gradient and ω denotes the frequency of oscillations. The flow generated by the pressure gradient (1.1) can serve as a simple model for the oscillating component of the blood flow in arteries (see McDonald [6] and Womersley [14] for details), and therefore a detailed examination of properties of the flow is important from a mathematical point of view as well as from a physiological point of view.

Governing equations for the flows are the well-known Navier–Stokes equations, and for both oscillatory and steady flow it is possible to find an analytical formula for the base flow. Having analytical formulas for the base flows, the question of stability of the base flows naturally arises. The present work is focused on the rigorous derivation of an explicit sufficient condition for monotone linear stability of the oscillatory flow, whereas special attention is given to stability of nonaxisymmetric disturbances. The reason for the focus on nonaxisymmetric disturbances is that stability characteristics of the oscillatory flow are—especially from the analytical point of view—mainly

*Received by the editors February 20, 2006; accepted for publication (in revised form) September 5, 2006; published electronically December 21, 2006. This work was partially supported by the Grant Agency of Charles University under grant GA UK 6/2005/R and by the Czech Science Foundation under grant GA CR 101/05/2537.

http://www.siam.org/journals/siap/67-2/65250.html

† Mathematical Institute, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, Prague 8, 186 75, Czech Republic (prusv@karlin.mff.cuni.cz).

unknown (see the discussion below). Nevertheless, the method presented here can also be used in the steady case, and its outcomes can be compared to previously achieved results.

Concerning stability of the steady flow, the pioneering works are that of Salwen and Grosch [9] and Salwen, Grosch, and Cotton [10], where the linear stability of the base flow was examined by numerical methods, and it was found that the steady Hagen–Poiseuille flow is stable for a wide range of parameters. The results of Salwen and Grosch [9] and Salwen, Grosch, and Cotton [10] were reproduced many times,[1] and now it is commonly accepted that the steady Hagen–Poiseuille flow is stable to all possible disturbances (as far as the disturbances can be described by the linear stability theory).

Important results in the analytical approach to the stability of the steady flow are the works of Herron [4], who proved that the base flow is linearly stable to axisymmetric disturbances,[2] and Joseph and Carmi [5], who proved (by semianalytical methods) that the base flow is monotonically stable to all possible disturbances if the Reynolds number is less than 81.49.

In the oscillatory case the situation is more complicated because the numerical calculations in the framework of the linear stability theory were—from the earliest one, that of Yang and Yih [15], to the most recent one by Blennerhassett and Bassom [2]—focused only on axisymmetric disturbances, and behavior of nonaxisymmetric disturbances remains unknown. Furthermore it seems that the oscillatory flow is in contrast to the steady flow linearly unstable for high Reynolds numbers.[3] The inadequacy of restriction of calculations only to axisymmetric disturbances is evident and was commented on by von Kerczek and Tozzi [12], but even in this work nonaxisymmetric disturbances to the oscillatory flow were not considered.

Analytical results comparable to Herron [4] and Joseph and Carmi [5] are not available for the oscillatory flow; therefore, a need to establish some analytical results is obvious. Furthermore, the lack of numerical results for nonaxisymmetric disturbances opens a possibility to extend analytical results to the area that was not yet examined by numerical means. In this work it is shown that it is indeed possible to establish a sufficient condition for monotone linear stability of the oscillatory flow and that this condition holds for both axisymmetric and nonaxisymmetric disturbances.

2. Problem formulation.

2.1. Governing equations for disturbances. In the framework of the linear stability theory (see, for example, Schmidt and Henningson [11]), the nondimensional governing equations for the disturbance $\vec{v}$ to the base flow $\vec{V}$ are

$$\frac{\partial \vec{v}}{\partial t} + [\nabla \vec{V}]\vec{v} + [\nabla \vec{v}]\vec{V} = -\nabla p + \frac{1}{\mathcal{R}}\Delta \vec{v}, \tag{2.1}$$

$$\operatorname{div} \vec{v} = 0, \tag{2.2}$$

together with the no-slip boundary condition on the pipe's wall and the periodic boundary condition on the ends of the pipe. The monotone stability is defined in the standard manner.

[1] See, for example, Meseguer and Trefethen [7] for a precise numerical calculation for high Reynolds numbers.

[2] Thus disturbances with $n = 0$; see (2.4) below.

[3] See Blennerhassett and Bassom [2] for details. Yang and Yih [15], however, claim that the flow is linearly stable, but they did not examine extreme parameter values as did Blennerhassett and Bassom [2].

DEFINITION 2.1 (monotone stability). *The flow is monotonically stable if for all disturbances* $\vec{v}$

$$\frac{1}{2}\frac{d}{dt}\|\vec{v}\|^2 < 0, \tag{2.3}$$

where $\|\vec{v}\|^2 = \int_\Omega \vec{v} \bullet \vec{v} dx$ *is kinetic energy of the disturbance in the volume* Ω.

The solution to the disturbance equations is sought in the form of waves

$$\vec{v}(r, \varphi, z, t) = \tilde{\vec{v}}(r, t) e^{i\alpha z + in\varphi}, \tag{2.4}$$

where $(\alpha, n) \in \mathbb{R} \times \mathbb{Z}$ is the wave vector of the disturbance and r, φ, and z are the cylindrical coordinates. In the pipe flow case, the scalar product is defined (see Salwen and Grosch [9] for a discussion) as integration over a section of the pipe that is bounded by two planes perpendicular to the pipe's axis and one wavelength apart[4]:

$$\langle \vec{v}, \vec{u} \rangle = \int_{r=0}^{1} \int_{\varphi=0}^{2\pi} \int_{z=0}^{\frac{2\pi}{\alpha}} \tilde{\vec{v}} \bullet \tilde{\vec{u}}^\star r dr d\varphi dz = \frac{4\pi^2}{\alpha} \int_{r=0}^{1} \tilde{\vec{v}} \bullet \tilde{\vec{u}}^\star r dr. \tag{2.5}$$

The norm in (2.3) is naturally the norm induced by the scalar product (2.5).

2.2. Base flow velocity. The base flow velocity has in both the oscillatory and steady cases a nonzero component only in the direction of the pipe's axis. For the steady flow the nondimensional base flow velocity is given by the well-known formula

$$V^{\hat{z}}(r) = (1 - r^2), \tag{2.6}$$

and the Reynolds number is defined in the standard manner—the characteristic length is equal to the pipe's radius R and the characteristic velocity U is equal to the centerline velocity $U = \frac{\Delta_s R^2}{4\rho\nu}$, where ρ is the density and ν is the kinematic viscosity.

In the oscillatory case there are several possibilities for introducing dimensionless variables. The approach of Womersley [14] is followed in this work. The Reynolds number is thus defined in the standard manner as $\mathcal{R} = \frac{RU}{\nu}$, where the characteristic velocity U is equal to $U = \frac{\Delta_p R^2}{4\rho\nu}$. The characteristic length is again the pipe's radius R. The second dimensionless parameter that is necessary in the oscillatory case is the Womersley number $\mathcal{W}$ defined by the formula $\mathcal{W} = \sqrt{\frac{\omega}{\nu}} R$.

Expressions for the base flow velocity and other characteristics of the oscillatory base flow were derived in Womersley [14] and McDonald [6]. Using the scaling introduced above, the dimensionless base flow is given by the formula

$$V^{\hat{z}}(r, t) = -i\frac{4}{\mathcal{W}^2}\left(1 - \frac{\mathrm{J}_0(i^{\frac{3}{2}}\mathcal{W}r)}{\mathrm{J}_0(i^{\frac{3}{2}}\mathcal{W})}\right) e^{i\omega t}, \tag{2.7}$$

where $\mathrm{J}_k(x)$ denotes the Bessel function of the first kind and of order k. Application of basic identities for the Bessel functions then immediately gives the following formula for the base flow velocity derivative:

$$\frac{\partial V^{\hat{z}}}{\partial r}(r, t) = -i^{\frac{5}{2}}\frac{4}{\mathcal{W}}\frac{\mathrm{J}_1(i^{\frac{3}{2}}\mathcal{W}r)}{\mathrm{J}_0(i^{\frac{3}{2}}\mathcal{W})} e^{i\omega t}. \tag{2.8}$$

[4] In the singular case $\alpha = 0$ the integration is only with respect to the r and φ variables.

3. Analytical results. The governing equation for the disturbance energy can be easily obtained by multiplying the governing equation for disturbances (2.1) by the disturbance itself and using the scalar product (2.5) to give

$$\left\langle \frac{\partial \tilde{\vec{v}}}{\partial t}, \tilde{\vec{v}} \right\rangle + \left\langle [\nabla \vec{V}]\tilde{\vec{v}} + [\nabla \tilde{\vec{v}}]\vec{V}, \tilde{\vec{v}} \right\rangle = \left\langle -\nabla \tilde{p} + \frac{1}{\mathcal{R}} \Delta \tilde{\vec{v}}, \tilde{\vec{v}} \right\rangle.$$

From the above equation it then follows that

$$\frac{1}{2}\frac{d}{dt}\|\tilde{\vec{v}}\|^2 = \Re\left(\left\langle -\nabla \tilde{p} + \frac{1}{\mathcal{R}} \Delta \tilde{\vec{v}}, \tilde{\vec{v}} \right\rangle - \left\langle [\nabla \vec{V}]\tilde{\vec{v}}, \tilde{\vec{v}} \right\rangle - \left\langle [\nabla \tilde{\vec{v}}]\vec{V}, \tilde{\vec{v}} \right\rangle \right), \tag{3.1}$$

where $\Re$ denotes real part.

The equation (3.1) is the governing equation for the disturbance energy measured in the norm induced by the scalar product. To prove monotone stability (monotone decay of disturbance energy) it is necessary to prove that the right-hand side of (3.1) is negative. The first term in the bracket is real and has a negative sign and it is even possible to estimate its magnitude—this can be done using eigenvalues of the Stokes operator (see section 3.1 for the properties of the Stokes operator). The last two terms in the bracket are complex and its real part has no definite sign. The real part must therefore be estimated by the absolute value. The steps described above are formally summarized in the following lemma.

LEMMA 3.1 (estimate of disturbance energy derivative). *Let $\alpha \in \mathbb{R}$, $n \in \mathbb{Z}$. Then the time derivative of kinetic energy of the disturbance with a wave vector (α, n) can be estimated as*

$$\frac{1}{2}\frac{d}{dt}\|\tilde{\vec{v}}\|^2 \leq \left(-\lambda_1 + |\alpha| \sup_{r\in[0,1],\ t\in[0,\frac{2\pi}{\omega}]} \left|V^{\hat{z}}\right| + \sup_{r\in[0,1],\ t\in[0,\frac{2\pi}{\omega}]} \left|\frac{\partial V^{\hat{z}}}{\partial r}\right| \right) \|\tilde{\vec{v}}\|^2, \tag{3.2}$$

where λ_1 is the lowest eigenvalue of the Stokes problem (3.6)–(3.7)*, and $V^{\hat{z}}$ is the component of the base flow velocity in the direction of the z-axis.*

Proof. Let us consider terms on the right-hand side of (3.1). For the first term it can be shown that (see section 3.1 for the properties of the Stokes operator)

$$\Re\left(\left\langle -\nabla \tilde{p} + \frac{1}{\mathcal{R}} \Delta \tilde{\vec{v}}, \tilde{\vec{v}} \right\rangle \right) = \left\langle -\nabla \tilde{p} + \frac{1}{\mathcal{R}} \Delta \tilde{\vec{v}}, \tilde{\vec{v}} \right\rangle \leq -\lambda_1 \left\langle \vec{v}, \vec{v} \right\rangle = -\lambda_1 \|\vec{v}\|^2, \tag{3.3}$$

where λ_1 is the lowest positive eigenvalue of the Stokes problem.

The last two terms on the right-hand side of (3.1) can be estimated by absolute value. The estimate for the first term is

$$\left| \left\langle [\nabla \vec{V}]\tilde{\vec{v}}, \tilde{\vec{v}} \right\rangle \right| = \left| \int_{r=0}^{1} \frac{\partial V^{\hat{z}}}{\partial r} \tilde{v}^{\hat{r}} \left(\tilde{v}^{\hat{z}}\right)^* r dr \right| \leq \sup_{r\in[0,1],\ t\in[0,\frac{2\pi}{\omega}]} \left|\frac{\partial V^{\hat{z}}}{\partial r}\right| \|\tilde{\vec{v}}\|^2, \tag{3.4}$$

and for the second term we find that

$$\left| \left\langle [\nabla \tilde{\vec{v}}]\vec{V}, \tilde{\vec{v}} \right\rangle \right| = \left| \int_{r=0}^{1} i\alpha V^{\hat{z}} \begin{bmatrix} \tilde{v}^{\hat{r}} \\ \tilde{v}^{\hat{\varphi}} \\ \tilde{v}^{\hat{z}} \end{bmatrix} \bullet \begin{bmatrix} \tilde{v}^{\hat{r}} \\ \tilde{v}^{\hat{\varphi}} \\ \tilde{v}^{\hat{z}} \end{bmatrix}^{\star} r dr \right| \leq |\alpha| \sup_{r\in[0,1],\ t\in[0,\frac{2\pi}{\omega}]} \left|V^{\hat{z}}\right| \|\tilde{\vec{v}}\|^2. \tag{3.5}$$

Substituting estimates (3.3)–(3.5) into the equation for the disturbance energy (3.1) yields the inequality in the lemma. □

To determine a sufficient condition for monotone stability it is necessary to investigate the dependence of the terms in the inequality (3.2) on the free parameters of the problem—namely on the Reynolds number $\mathcal{R}$, the Womersley number $\mathcal{W}$ (for the oscillatory case only), and the wave vector (α, n) of the disturbance.

The required estimates on the base flow velocity and the base flow velocity derivative in terms of free parameters are given in section 3.2, and the estimate of the lowest eigenvalue λ_1 of the Stokes problem is derived in section 3.1. In these paragraphs various inequalities and identities for the Bessel functions are used. The identities can be found in Abramowitz and Stegun [1] or—including proofs—in Watson [13].

3.1. Stokes problem. The eigenvalue problem for the Stokes operator in a circular pipe is given by

$$-\frac{1}{\mathcal{R}}\Delta\vec{v} + \nabla p = \lambda\vec{v}, \tag{3.6}$$

$$\operatorname{div}\vec{v} = 0, \tag{3.7}$$

together with the no-slip boundary condition on the pipe's wall and the periodic boundary condition on the ends of the pipe.

It can be proved (see, for example, Constantin and Foias [3]) that the spectrum of the Stokes operator consists only of a point spectrum and that the eigenvalues of the Stokes operator are all positive. Furthermore, the eigenvectors form a complete orthogonal basis in an appropriate function space.

The Fourier transformed version[5] of the eigenvalue problem for the Stokes operator is considered in Salwen and Grosch [9] and Rummler [8]. In can be shown that the Fourier transformed eigenvalue problem has the same properties as the eigenvalue problem in real space (the eigenvalues are real, positive, and simple and the eigenfunctions form a complete orthogonal basis in an appropriate function space), and furthermore the eigenfunctions and the eigenvalues of the Fourier transformed problem can be explicitly calculated for all wave vectors (α, n).

The eigenvalues of the Fourier transformed problem are given by a complicated implicit equation and can be found by numerical means. However the aim of this work is to obtain an explicit condition for stability, and the mere fact that it is possible to calculate the eigenvalues by numerical solution of the implicit equation is clearly not sufficient for this purpose. To obtain wanted explicit conditions it is therefore necessary to find some explicit estimate of the lowest eigenvalue of the Stokes operator—this is done in the following lemma.

LEMMA 3.2 (estimate of the lowest eigenvalue of the Stokes operator). *Let λ_1 be the lowest positive eigenvalue of the Fourier transformed Stokes operator and let $\mathrm{j}_{k,l}$ denote the lth positive root of the Bessel function $\mathrm{J}_k(x)$.*

If $\alpha = 0$ and $n = 0$ then

$$\lambda_1 = \frac{1}{\mathcal{R}}\mathrm{j}_{0,1}^2. \tag{3.8}$$

If $\alpha = 0$ and $n \neq 0$ then

$$\lambda_1 = \frac{1}{\mathcal{R}}\mathrm{j}_{n,1}^2. \tag{3.9}$$

[5]The solution is sought in the wavy form (2.4).

If $\alpha \neq 0$ *and* $n = 0$ *then*

$$\lambda_1 = \frac{1}{\mathcal{R}} \left(\alpha^2 + \mathrm{j}_{1,1}^2 \right). \tag{3.10}$$

If $\alpha \neq 0$ *and* $n \neq 0$ *then*

$$\lambda_1 \geq \frac{1}{\mathcal{R}} \left(\alpha^2 + \mathrm{j}_{n-1,1}^2 \right). \tag{3.11}$$

Proof. In Rummler [8] it is shown that the eigenvalues of the Fourier transformed Stokes operator are equal to $\lambda_k = \frac{1}{\mathcal{R}} \left(\alpha^2 + \beta_k^2 \right)$, where the parameters β_k are roots of certain characteristic equations according to the value of the wavevector (α, n).

For $\alpha = 0$ and $n = 0$ parameters β_k are positive roots[6] of equations $\mathrm{J}_1(\beta_k) = 0$ and $\mathrm{J}_0(\beta_k) = 0$. For the Bessel functions it is known that for $m, l \in \mathbb{N}$, $m < l$ there is $\mathrm{j}_{m,1} < \mathrm{j}_{l,1}$. It is therefore obvious that the lowest eigenvalue is in this case given by (3.8).

For $\alpha = 0$ and $n \neq 0$ parameters β_k are positive roots of equations $\mathrm{J}_n(\beta_k) = 0$ and $\mathrm{J}_{|n|+1}(\beta_k) = 0$. The result (3.9) is then obtained by the same argument as in the previous case.

For $\alpha \neq 0$ and $n = 0$ parameters β_k are positive roots of equations

$$\mathrm{J}_1(\beta_k) = 0 \tag{3.12}$$

and

$$|\alpha| \mathrm{I}_0(|\alpha|) \mathrm{J}_1(\beta_k) - \beta \mathrm{I}_1(|\alpha|) \mathrm{J}_0(\beta_k) = 0, \tag{3.13}$$

where $\mathrm{I}_k(x)$ denotes the modified Bessel function of the first kind and of order k.

Dividing (3.13) by $|\alpha| \mathrm{I}_0(|\alpha|)$ and taking the limit $|\alpha| \to +\infty$ or $|\alpha| \to 0+$ yield asymptotic equations $\mathrm{J}_1(\beta_k) = 0$ and $-\frac{\beta_k}{2} \mathrm{J}_2(\beta_k) = 0$ respectively. It is obvious that roots of these asymptotic equations cannot be roots of the original (3.13). From the asymptotic equations, continuity of the left-hand side of (3.13), and the fact that the eigenvalues of the Stokes problem must be simple for all α and n, it follows that the first positive root of (3.13) must be for all $|\alpha|$ localized in the interval $(\mathrm{j}_{1,1}, \mathrm{j}_{2,1})$. The first positive β_k is therefore the first positive solution to (3.12), and equality (3.10) indeed holds.

For $\alpha \neq 0$ and $n \neq 0$ parameters β_k are roots of the following equation:

$$\begin{aligned} -2|\alpha| \mathrm{I}_n(|\alpha|) \mathrm{J}_{n-1}(\beta_k) \mathrm{J}_{n+1}(\beta_k) + \beta_k \mathrm{I}_{n+1}(|\alpha|) \mathrm{J}_{n-1}(\beta_k) \mathrm{J}_n(\beta_k) \\ - \beta_k \mathrm{I}_{n-1}(|\alpha|) \mathrm{J}_n(\beta_k) \mathrm{J}_{n+1}(\beta_k) = 0. \end{aligned} \tag{3.14}$$

In this case it is possible to apply the same approach as above. Dividing the equation by $|\alpha| \mathrm{I}_n(|\alpha|)$ and taking the limit $|\alpha| \to +\infty$ yield

$$-2 \mathrm{J}_{n-1}(\beta_k) \mathrm{J}_{n+1}(\beta_k) = 0, \tag{3.15}$$

and dividing (3.14) by $\mathrm{I}_{n-1}(|\alpha|)$ and taking the limit $|\alpha| \to 0+$ lead to the asymptotic equation

$$-\beta_k \mathrm{J}_n(\beta_k) \mathrm{J}_{n+1}(\beta_k) = 0. \tag{3.16}$$

[6] Trivial root $\beta = 0$ would lead to the trivial solution of the eigenvalue problem for the Stokes operator.

The first positive root of (3.15) is $\mathrm{j}_{n-1,1}$ and the first positive root of (3.16) is $\mathrm{j}_{n,1}$. Again it is obvious that neither $\mathrm{j}_{n-1,1}$ nor $\mathrm{j}_{n,1}$ are roots of the original (3.14) and—as in the previous case—it follows that the first positive root of the original equation is localized in the interval $(\mathrm{j}_{n-1,1}, \mathrm{j}_{n,1})$, and therefore the inequality (3.11) indeed holds. ∎

3.2. Estimates on the base flow. The last step necessary to rewrite the right-hand side of the inequality (3.2) as an explicit function of $\mathcal{R}$, $\mathcal{W}$, α, and n is to find some estimates of the expressions appearing in the inequalities (3.5) and (3.4). The appropriate estimates are given below.

LEMMA 3.3 (estimate of base flow velocity). *Let $V^{\hat{z}}$ be the base flow velocity given by the formula* (2.7)*. Then*

$$\sup_{r\in[0,1],\ t\in[0,\frac{2\pi}{\omega}]} \left|V^{\hat{z}}(r,t)\right| \le \frac{8}{\mathcal{W}^2}. \tag{3.17}$$

Proof. The proof of the lemma is straightforward:

$$\left|V^{\hat{z}}(r,t)\right| \le \frac{4}{\mathcal{W}^2}\left(1+\left|\frac{\mathrm{J}_0(i^{\frac{3}{2}}\mathcal{W}r)}{\mathrm{J}_0(i^{\frac{3}{2}}\mathcal{W})}\right|\right) \le \frac{8}{\mathcal{W}^2}.$$

The last estimate in the inequality is obvious from the identity

$$\begin{aligned}\left|\mathrm{J}_\nu(i^{\frac{3}{2}}x)\right|^2 &= \left|\mathrm{ber}_\nu(x)+i\,\mathrm{bei}_\nu(x)\right|^2 \\ &= \left(\frac{x}{2}\right)^{2\nu}\sum_{k=0}^{+\infty}\frac{1}{\Gamma(\nu+k+1)\Gamma(\nu+2k+1)}\frac{\left(\frac{x^2}{4}\right)^{2k}}{k!}.\end{aligned} \tag{3.18}$$

∎

LEMMA 3.4 (estimate of base flow velocity derivative). *Let $V^{\hat{z}}$ be the base flow velocity given by the formula* (2.7)*. Then*

$$\sup_{r\in[0,1],\ t\in[0,\frac{2\pi}{\omega}]} \left|\frac{\partial V^{\hat{z}}}{\partial r}(r,t)\right| \le \frac{4}{\mathcal{W}}. \tag{3.19}$$

Proof. The estimate easily follows from the above-mentioned identity (3.18) and basic inequality $|\mathrm{J}_0(i^{\frac{3}{2}}x)| > |\mathrm{J}_1(i^{\frac{3}{2}}x)|$. ∎

3.3. Sufficient condition for monotone linear stability. Combining estimates from sections 3.1 and 3.2 leads to the sufficient condition for monotone linear stability of the oscillatory and the steady Hagen–Poiseuille flow.

THEOREM 3.5 (sufficient condition for monotone linear stability). *Let $\tilde{\vec{v}}$ be a disturbance to the oscillatory Hagen–Poiseuille flow determined by the Reynolds number $\mathcal{R}$ and the Womersley number $\mathcal{W}$, and let (α, n) be the wave vector of the disturbance. If $\alpha = 0$, $n = 0$, and*

$$-\frac{1}{\mathcal{R}}\mathrm{j}_{0,1}^2 + |\alpha|\frac{8}{\mathcal{W}^2} + \frac{4}{\mathcal{W}} < 0, \tag{3.20}$$

or if $\alpha = 0$, $n \neq 0$, and

$$-\frac{1}{\mathcal{R}}\mathrm{j}_{n,1}^2 + \frac{4}{\mathcal{W}} < 0, \tag{3.21}$$

or if $\alpha \neq 0$, $n = 0$, *and*

$$-\frac{1}{\mathcal{R}}\left(\alpha^2 + \mathrm{j}_{1,1}^2\right) + |\alpha|\frac{8}{\mathcal{W}^2} + \frac{4}{\mathcal{W}} < 0, \tag{3.22}$$

or if $\alpha \neq 0$, $n \neq 0$,

$$-\frac{1}{\mathcal{R}}\left(\alpha^2 + \mathrm{j}_{n-1,1}^2\right) + |\alpha|\frac{8}{\mathcal{W}^2} + \frac{4}{\mathcal{W}} < 0, \tag{3.23}$$

then the disturbance energy monotonically decays in time.

Proof. The proposition is a immediate consequence of the energy estimate (Lemma 3.1), the estimate of the lowest eigenvalue of the Stokes operator (Lemma 3.2), and estimates of base flow velocity and base flow velocity derivative (Lemmas 3.3 and 3.4). ∎

The first condition in the theorem is in fact exceedingly strict. It can be proved that the disturbance with $(\alpha, n) = (0, 0)$ monotonically decays in time for all values of the Reynolds number $\mathcal{R}$ and the Womersley number $\mathcal{W}$. Indeed, in this case—due to a special form of the eigenfunctions of the Stokes operator[7]—the second and third terms on the right-hand side of (3.2) are equal to zero, and therefore there is no need to balance positive and negative terms using the estimates (3.5) and (3.4).

A theorem similar to Theorem 3.5 can be also derived in the steady case. Estimates on the base flow velocity analogous to the estimates (3.17) and (3.19) are in this case trivial. For example, in the most general case $\alpha \neq 0$ and $n \neq 0$, the condition for monotone decay of disturbances to the steady Hagen–Poiseuille flow reads

$$-\frac{1}{\mathcal{R}}\left(\alpha^2 + \mathrm{j}_{n-1,1}^2\right) + |\alpha| + 2 < 0. \tag{3.24}$$

Using Theorem 3.5 it is easy to derive the following condition for monotone decay of all possible disturbances.

Corollary 3.6 (explicit bound on Reynolds number). *If the Reynolds number* $\mathcal{R}$ *for the oscillatory Hagen–Poiseuille flow satisfies the inequality* $\mathcal{R} < \mathcal{R}_{\mathrm{crit}}$, *where*

$$\mathcal{R}_{\mathrm{crit}} = \frac{\mathcal{W}^2}{8}\left(2\sqrt{\left(\frac{\mathcal{W}}{2}\right)^2 + \mathrm{j}_{0,1}^2} - \mathcal{W}\right), \tag{3.25}$$

then the flow is monotonically linearly stable to all possible disturbances.

If the Reynolds number $\mathcal{R}$ *for the steady Hagen–Poiseuille satisfies the inequality* $\mathcal{R} < \mathcal{R}_{\mathrm{crit}}$, *where*

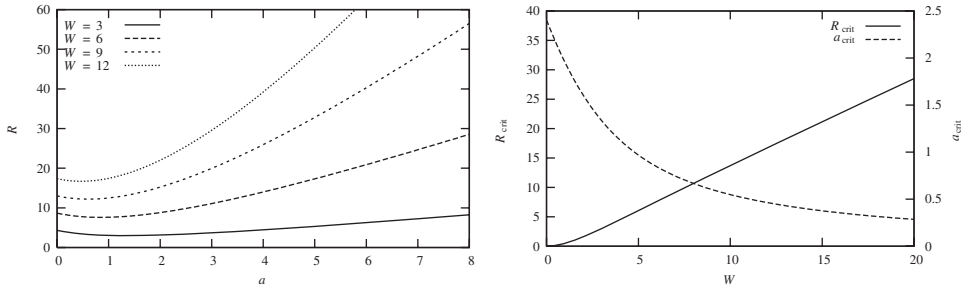
$$\mathcal{R}_{\mathrm{crit}} = 2\left(\sqrt{4 + \mathrm{j}_{0,1}^2} - 2\right), \tag{3.26}$$

then the flow is monotonically linearly stable to all possible disturbances.

Proof. The inequality $\mathrm{j}_{n,1} < \mathrm{j}_{n+1,1}$ that holds for all $n \in \mathbb{N}$, together with the estimates from Theorem 3.5, leads to the following general sufficient condition that ensures monotone decay for all n:

$$-\frac{1}{\mathcal{R}}\left(\alpha^2 + \mathrm{j}_{0,1}^2\right) + |\alpha|\frac{8}{\mathcal{W}^2} + \frac{4}{\mathcal{W}} < 0. \tag{3.27}$$

[7] See Rummler [8] for explicit formulas for the eigenfunctions in this special case.



(a) stable regions according to (3.23)

(b) critical values according to (3.25) and (3.29)

FIG. 3.1. *Sufficient condition for monotone stability of oscillatory Hagen–Poiseuille flow.*

The condition can be rewritten as

$$\mathcal{R} < \frac{\alpha^2 + \mathrm{j}_{0,1}^2}{\alpha\frac{8}{\mathcal{W}^2} + \frac{4}{\mathcal{W}}}, \tag{3.28}$$

and the corollary is then a straightforward consequence of minimization of the right-hand side with respect to α. The critical value of α where the function on the right-hand side attains its minimum is equal to

$$\alpha_{\mathrm{crit}} = \sqrt{\left(\frac{\mathcal{W}}{2}\right)^2 + \mathrm{j}_{0,1}^2} - \frac{\mathcal{W}}{2}, \tag{3.29}$$

and substitution of this value back into (3.28) gives the proposition.

In the steady case the proposition is proved by the same arguments; the key inequality analogous to (3.28) is in this case $\mathcal{R} < \frac{\alpha^2+\mathrm{j}_{0,1}^2}{\alpha+2}$. □

A plot of parameter regions which satisfy condition (3.23) for $n = 1$ (and thus also the general condition (3.23) that is uniform with respect to n) is shown in Figure 3.1(a). All pairs $(\alpha, \mathcal{R})$ that for a given Womersley number $\mathcal{W}$ lie below the corresponding curve in Figure 3.1(a) lead to monotonically decaying disturbances. The dependence of the critical Reynolds number $\mathcal{R}_{\mathrm{crit}}$ and the critical wavelength α_{crit} on the Womersley number $\mathcal{W}$ is shown in Figure 3.1(b).

4. Conclusion. Explicit sufficient conditions for monotone linear stability of the oscillatory and steady Hagen–Poiseuille flow were found by purely analytical means.

In the case of the steady Hagen–Poiseuille flow a numerical evaluation of the sufficient condition (3.26) leads to the conclusion that if the Reynolds number is less than approximately 2.26, then the flow is monotonically stable to all possible disturbances. This bound on monotone stability is low compared to the bound given by Joseph and Carmi [5], but the value of the present result is that it was achieved by purely analytical means. Herron [4] proved stability (but not monotone stability) for axisymmetric disturbances, and his result holds without any further conditions on the magnitude of the Reynolds number. In light of this result, the present sufficient condition (3.26) is clearly exceedingly restrictive for axisymmetric disturbances; however, the condition (3.26) becomes important if it is necessary to describe behavior of nonaxisymmetric disturbances that are beyond the scope of the Herron result.

Considering the oscillatory Hagen–Poiseuille flow, the condition (3.25) is more valuable because of the lack of analytical and numerical results particularly for nonaxisymmetric disturbances. Especially for nonaxisymmetric disturbances, where even numerical results are not available, the condition provides an important description of stability properties of the flow. However, the bound on the Reynolds number given in the condition (3.25) is very strict, and the condition is difficult—but not impossible—to meet in a practically important situation. For example, if the first harmonic component of oscillatory flow in the femoral artery of a dog (see Womersley [14]) is considered, then $\mathcal{W} = 3.3$ and $\mathcal{R} = 648$, while the condition (3.25) ensures stability for the Reynolds number less than 1.7. Nevertheless, for the sixth harmonic (sine) component, there are $\mathcal{W} = 8.2$ and $\mathcal{R} = 12.3$, while the Reynolds number sufficient for monotone stability is equal to 4.7.

Sufficient conditions set in Theorem 3.5 are important not only from the analytical point of view, but also from the numerical point of view, because they for a given Reynolds number and Womersley number provide rigorous bounds on the range of wavevectors (α, n) that must be examined in any future search for instability by numerical means. Furthermore, the critical wavelength α_{crit} provides a clue to the question of which wavelength α could be expected to be the least stable or even unstable.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, U.S. National Bureau of Standards, Applied Mathematics Series 55, Government Printing Office, Washington, D.C., 1964.

[2] P. J. Blennerhassett and Andrew P. Bassom, *The linear stability of high-frequency oscillatory flow in a channel*, J. Fluid Mech., 556 (2006), pp. 1–25.

[3] P. Constantin and C. Foias, *Navier-Stokes Equations*, University of Chicago Press, Chicago, IL, 1988.

[4] I. M. Herron, *Observations on the role of vorticity in the stability theory of wall bounded flows*, Stud. Appl. Math., 85 (1991), pp. 269–286.

[5] D. D. Joseph and S. Carmi, *Stability of Poiseuille flow in pipes, annuli and channels*, Quart. Appl. Math., 26 (1969), pp. 575–599.

[6] D. A. McDonald, *The relation of pulsatile pressure to flow in arteries*, J. Physiol., 127 (1955), pp. 533–552.

[7] Á. Meseguer and L. N. Trefethen, *Linearized pipe flow to Reynolds number* 10^7, J. Comput. Phys., 186 (2003), pp. 178–197.

[8] B. Rummler, *The eigenfunctions of the Stokes operator in special domains*, Z. Angew. Math. Mech., 77 (1997), pp. 619–627.

[9] H. Salwen and C. E. Grosch, *The stability of Poiseuille flow in a pipe of circular cross section*, J. Fluid Mech., 54 (1972), pp. 93–112.

[10] H. Salwen, C. E. Grosch, and F. W. Cotton, *Linear stability of Poiseuille flow in circular pipe*, J. Fluid Mech., 98 (1980), pp. 273–284.

[11] P. J. Schmidt and D. S. Henningson, *Stability and Transition in Shear Flows*, Appl. Math. Sci. 142, Springer-Verlag, New York, 2001.

[12] C. H. von Kerczek and J. T. Tozzi, *The stability of oscillatory Hagen-Poiseuille flow*, Trans. ASME J. Appl. Mech., 53 (1986), pp. 187–192.

[13] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, Cambridge, UK, 1944. Reprinted 1980.

[14] J. R. Womersley, *Method for calculation of velocity, rate of flow and viscous drag in arteries when the pressure gradient is known*, J. Physiol., 127 (1955), pp. 225–245.

[15] W. H. Yang and Chia-Shun Yih, *Stability of time-periodic flows in a circular pipe*, J. Fluid Mech., 82 (1977), pp. 497–505.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 364–386

AN ASYMPTOTIC FRAMEWORK FOR FINITE HYDRAULIC FRACTURES INCLUDING LEAK-OFF*

S. L. MITCHELL†, R. KUSKE†, AND A. P. PEIRCE†

Abstract. The dynamics of hydraulic fracture, described by a system of nonlinear integro-differential equations, is studied through the development and application of a multiparameter singular perturbation analysis. We present a new single expansion framework which describes the interaction between several physical processes, namely viscosity, toughness, and leak-off. The problem has nonlocal and nonlinear effects which give a complex solution structure involving transitions on small scales near the tip of the fracture. Detailed solutions obtained in the crack tip region vary with the dominant physical processes. The parameters quantifying these processes can be identified from critical scaling relationships, which are then used to construct a smooth solution for the fracture depending on all three processes. Our work focuses on plane strain hydraulic fractures on long time scales, and this methodology shows promise for related models with additional time scales, fluid lag, or different geometries, such as radial (penny-shaped) fractures and the classical Perkins–Kern–Nordgren (PKN) model.

Key words. asymptotic solutions, crack tip, critical scales, hydraulic fractures, integral-differential equations, leak-off

AMS subject classifications. 45K05, 35A20, 35B40, 76D08, 74B05, 74R10

DOI. 10.1137/04062059X

1. Introduction.

Hydraulic fractures are propagated in an elastic material due to the pressure exerted by a viscous fluid on the fracture. These fractures occur naturally in volcanic dikes where magma causes fracture propagation below the surface of the earth [37, 38, 55]. In the oil and gas industry hydraulic fractures are deliberately propagated in reservoirs to increase production. Hydraulic fracture models need to account for the primary physical mechanisms involved: deformation of the rock, fracturing of the rock, flow of viscous fluid within the fracture, and leak-off of the fracturing fluid into the permeable rock. The parameters that characterize these processes are, respectively, Young's modulus E and Poisson's ratio ν, the rock toughness K_{Ic}, the fluid viscosity μ, and the leak-off coefficient C_l.

The challenges for analysis of these models originate from the nonlinearity of the equation describing the flow of fluid in the fracture, the nonlocal character of the elastic response of the fracture, and the history-dependence of the equation governing the exchange of fluid between the fracture and the rock. The singular tip behavior, which can be difficult to resolve numerically, dominates these solutions and is highly dependent on the relative importance of the contributing physical processes. Therefore, the objectives of analytic treatment of these models are as follows: to characterize the structure of the near-tip solution that can be embedded in numerical algorithms, to provide benchmark solutions to test numerical codes, and to determine the parameter values and length scales that characterize the transitions between distinct combinations of physical processes. In this paper we use a novel asymptotic framework that enables us to characterize the different propagation regimes and provide asymptotic solutions when more than two physical processes are competing simultaneously. This

*Received by the editors December 9, 2004; accepted for publication (in revised form) August 14, 2006; published electronically January 12, 2007.
http://www.siam.org/journals/siap/67-2/62059.html
†Department of Mathematics, University of British Columbia, Vancouver, BC, V6T 1Z2, Canada (sarah@iam.ubc.ca, rachel@math.ubc.ca, peirce@math.ubc.ca).

is distinct from previous analytic work on such models, which have been restricted to considering at most two competing physical processes [6, 23, 24, 26].

There has been a significant amount of work in the last half century involving the mathematical modeling of hydraulic fractures [1, 8, 14, 28, 30, 31, 32, 33, 45, 48, 55]. As discussed in [21] and the references therein, the aim of these models is to calculate the fluid pressure, opening, and size of the fracture given the properties of the rock, the injection rate, and the fluid characteristics. More recent work has been concerned with developing numerical algorithms to simulate three-dimensional propagation of hydraulic fractures in layered strata [5, 7, 12, 46, 47, 53]; this is in contrast to earlier work where approximate solutions were found for simple fracture geometries [1, 8, 28, 33, 45, 48, 55]. A substantial difference between hydraulic fracturing and other studies of fracture (see [22, 49, 52]) is the coupling with the equations for the fluid and fracture geometry. Most models in hydraulic fracturing only consider planar fractures rather than kinked or curved cracks [13, 41, 42].

The relevant fracture geometry that we consider in this paper, known as the KGD (plane strain) model, was developed independently by Khristianovic and Zheltov [33] and Geertsma and de Klerk [28]. The fracture is assumed to be an infinite vertical strip so that horizontal cross-sections are in a state of plane strain. This model is applicable to large aspect ratio rectangular planar fractures and was extended in [54] to include toughness. A major contribution to this mathematical modeling was made by Spence and Sharp [54], who initiated the work on self-similar solutions and scaling for a KGD crack propagating in an elastic, impermeable medium with finite toughness. This approach has been continued through asymptotic analyses of near-tip processes, yielding the results from [15] for zero toughness in an impermeable rock, and from [36] for zero toughness when leak-off is dominant. Several papers [16, 19, 27] have extended this analysis to include toughness and fluid lag, where regions devoid of fluid develop close to the crack tip, along with transitional regions. In this paper we assume that fluid lag is negligible and so these effects can be ignored.

Certain phases of hydraulic fracture propagation are characterized within a dimensionless parametric space [21, 20], with boundaries controlled by the dominant processes, namely, viscosity, toughness, or leak-off. This framework has been the basis for semianalytical solutions for simple geometries (KGD and penny-shaped) and benchmarks for numerical simulators. These include the following asymptotic regimes: impermeable with zero toughness [2, 10, 50], small toughness [24], finite toughness [3, 54], and large toughness [26, 50]; and permeable with zero toughness [4].

Since much of our analysis is closely related and complementary to these most recent studies, we outline the context here, with further discussion given in section 1.1 in terms of the specific model. Previous analyses [2, 9, 10, 24, 25, 26, 27] have been limited to parameter regimes where one or two physical processes dominate the dynamics, with the remainder of the related nondimensional quantities set to unity. In each case there is a different set of scaling parameters defined, depending on the dominant process(es), corresponding to the edges and corners of the parameter space [6, 21]. These methods lead to asymptotic expansions for the tip behavior, where the terms in the expansion involve powers of the distance from the tip. In the case of vanishing leak-off, this method has also been used to describe the transition in behavior between different power law expansions [2]. Recent preliminary studies [23, 34, 35] have also used combinations of power law expansions in the context of a semi-infinite approximation for the fracture, combined with numerical methods to understand transitions between different scaling regimes.

In this paper we present a unifying scaling framework based on singular perturbation techniques which analyzes how the physical processes, namely, viscosity, toughness, and leak-off, all influence the KGD crack behavior. Avoiding semi-infinite approximations, it involves the simultaneous scaling of all three processes relative to the distance from the fracture tip: this means that the approach is applicable for different combinations of the dominant physical processes. It has been used in [44] for the impermeable case in which only the processes of viscous dissipation and energy release compete. Thus it provides a construction of the solution in the crucial tip region, identifying the parameter combinations which quantify spatial transitions in the behavior of the fracture. The scaling exponents of the physical processes are determined as part of the method, so that it can be applied to construct approximate solutions in intermediate parameter regimes where several processes are in balance. The resulting asymptotic approximation provides verification of the conditions under which self-similar solutions are appropriate, and indicates regimes in which a more complicated time-dependence is involved, as discussed in section 4. We also briefly outline how the technique can be generalized to regimes where there is more than one transition in the behavior near the tip.

The fracture propagation is formulated as a system of coupled integrodifferential equations, and our method proves to be very beneficial in understanding the nonlocal and local effects that arise. It can be applied to different geometries, such as the classical Perkins–Kern–Nordgren (PKN) model [43], and we expect that it can be extended to model other effects such as stress jumps and fluid lag.

1.1. Problem formulation and dimensional results. The solution of the KGD hydraulic fracture problem (shown in Figure 1.1) consists of determining the fracture opening w and the net pressure p (the difference between the fluid pressure p_f and the far-field stress σ_o) as functions of space and time, as well as the fracture half-length, $l(t)$. These functions depend on the volumetric fluid injection rate Q_0, assumed constant in this paper, and on the four material parameters E', μ', K', and

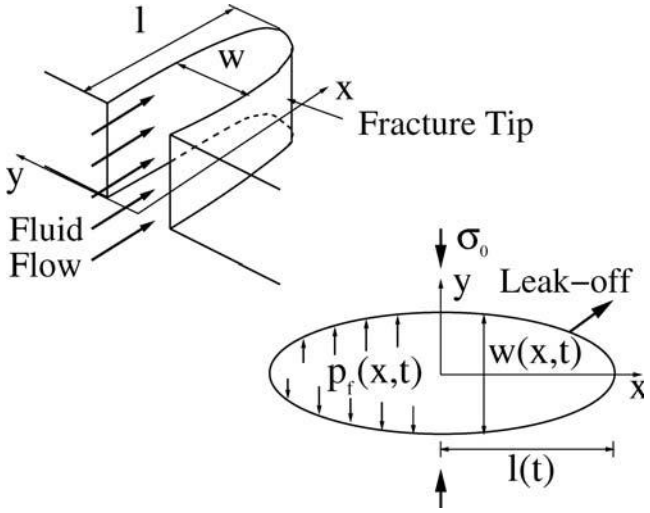


FIG. 1.1. *Diagrams showing the KGD crack and its cross-section.*

C', respectively, defined as

$$E' = \frac{E}{1-\nu^2}, \qquad \mu' = 12\mu, \qquad K' = 4\left(\frac{2}{\pi}\right)^{1/2} K_{Ic}, \qquad C' = 2C_l, \tag{1.1}$$

which are combinations of the parameters quantifying the primary physical mechanisms described at the beginning of this section. The rock toughness K_{Ic} is assumed to be equal to the stress intensity factor K_I which, for this geometry, can be expressed as an integral of the pressure

$$K_{Ic} = K_I = 2\sqrt{\frac{l}{\pi}} \int_0^l \frac{p}{\sqrt{l^2 - x^2}}\, \mathrm{d}x. \tag{1.2}$$

The equations for the KGD fracture are as follows:

$$\text{Reynolds' (lubrication) equation:} \quad \frac{\partial w}{\partial t} + g = \frac{1}{\mu'} \frac{\partial}{\partial x}\left[w^3 \frac{\partial p}{\partial x}\right] + Q_0 \delta(x), \tag{1.3}$$

which describes the conservation of fluid mass for an incompressible fluid. Note that g is the leak-off term which describes the fluid infiltration into the surrounding rock.

$$\text{Elasticity equation:} \qquad p(x,t) = -\frac{E'}{4\pi} \int_{-l}^{l} \frac{\partial w}{\partial s} \frac{\mathrm{d}s}{s-x}, \tag{1.4}$$

which describes the balance of forces and is a nonlocal equation relating the fracture opening w and net pressure p for a state of plane strain.

$$\text{Propagation condition:} \quad w = \frac{K'}{E'}\sqrt{l-x} + O\big[(l-x)^{3/2}\big], \qquad x \longrightarrow \pm l, \tag{1.5}$$

which accounts for the energy required to break the rock and is the condition that the fracture is in mobile equilibrium.

$$\text{Boundary conditions:} \qquad w = 0, \qquad w^3 \frac{\partial p}{\partial x} = 0, \qquad \text{at } x = \pm l. \tag{1.6}$$

$$\text{Global volume balance condition:} \quad Q_0 t = \int_{-l}^{l} w(s,t)\, \mathrm{d}s + \int_0^t \int_{-l}^{l} g(s,\tau)\, \mathrm{d}s\, \mathrm{d}\tau, \tag{1.7}$$

which equates the crack volume to the volume of injected fluid and amount lost to the surrounding rock mass, obtained by integrating (1.3) and applying (1.6). If $t_0(x)$ is the time at which the crack tip arrived at the point x, and t is the current time, then the leak-off function g is defined as

$$\text{Carter's leak-off model:} \qquad g(x,t) = \frac{C' H(t - t_0)}{\sqrt{t - t_0(x)}}. \tag{1.8}$$

The memory term $t_0(x)$ implies that the leak-off function $g(x,t)$ depends on the entire history of the fracture front locations, which significantly complicates the analysis.

Since its introduction in 1957, Carter's leak-off model [11] has been widely accepted and successfully used in the oil and gas industry to design hydraulic fracturing

treatments and has been referred to as "the standard model of fracturing fluid loss" (see [36]). We briefly summarize the steps involved in the derivation of the model and discuss its applicability for high confinement geological situations, which are becoming more important as deeper reserves are being exploited.

The first assumption made in the derivation of Carter's leak-off model is that the hydraulic load $\Delta p = p_f - p_0$ driving the leak-off process is approximately constant, where p_0 is the reservoir pore pressure. This assumption can be justified in high confinement reservoirs where $p_f \approx \sigma_0 \gg p_0$. In this case the hydraulic load is much larger than the net pressure $p = p_f - \sigma_0$ and is approximately constant, i.e., $\Delta p \approx \sigma_0 - p_0$. The second assumption made in the derivation of (1.8) is in approximating the leak-off process by a one-dimensional flow perpendicular to the crack propagation axis that does not account for any lateral interaction. Modeling this gradient-driven flow involves incorporating the growth of an impermeable *filter cake* layer via the deposition of polymer molecules by the leaking fluid, the growth of an *invaded zone* of fluid that penetrates the filter cake, and a pressure diffusion zone within the reservoir. Combining these three physical processes in series yields (1.8), in which the lumped coefficient C' is known as the *Carter leak-off coefficient* (see [6, 11, 39, 51] and the references therein).

Assuming that $l(t) = at^{1/2}$, Gordeyev and Entov [29] derived a similarity solution to the two-dimensional pressure diffusion equation, which yields a leak-off velocity of the same form as (1.8). In this case the fracture is growing sufficiently rapidly for the leak-off process to be one-dimensional, a situation that is likely to persist for power laws in which the fracture evolves more rapidly: $l(t) = at^{\lambda}$, where $\lambda \geq \frac{1}{2}$. Carter's model (1.8), which is based on the pressure diffusion equation, ignores feedback coupling between the reservoir pressure field and the elastic strain in the rock. This pure diffusion approximation can be justified using poroelasticity theory [17, 18] in which the elastic strain feedback due to the hydraulic load Δp is shown to vanish identically. Moreover, for high confinement reservoirs the mechanical load effect (due to the net pressure p which forces the crack to open) on the reservoir pressure is insignificant compared to that of the hydraulic load, since $\Delta p \gg p$.

There may be a question as to the validity of Carter's model right at the crack tip. However, the analysis presented in this paper is based on the fact that the dominant physical process governing the behavior of the fracture at the tip, which we refer to as the *near-tip region*, is the energy released in the breaking of the rock as characterized by the fracture toughness. Since the leak-off process is subdominant to this and only manifests itself a distance away from the tip in the *intermediate-tip region*, we make use of the model only in a region where it is still valid. We could include other higher order effects to model the leak-off more carefully in the near-tip region, but this is neglected in our analysis since it is not the dominant process.

The method presented in this paper involves an iterative construction of the asymptotic solution: the lubrication and elasticity equations (1.3)–(1.4) alternatively give the form of the solution. The volume balance equation (1.7) is applied to complete the solution: it verifies the balance of physical processes, determines unknown constants, and provides a consistency check on the temporal behavior of the solution. The propagation condition (1.5) manifests itself in the asymptotic behavior of the tip when the influence of the toughness is dominant; this depends on the relative scalings of the parameters and the distance from the tip.

We give the main results for w in terms of the dimensional variables; the expansions for p can then be determined from (1.4). For $\mathcal{P}_{ckm} \gg (\ll) (1-\xi)^{1/2}$, respectively,

we find that

$$w \sim \frac{K'}{E'} l^{1/2} \left\{ \left(1 - \frac{x}{l}\right)^{1/2} + \left[\frac{8\pi}{3}\gamma^{3/2}\mathcal{P}_{km}^{-1} + 4\sqrt{2}\pi\gamma\mathcal{P}_{ckm}^{-1}\right]\left(1 - \frac{x}{l}\right)\right\}, \tag{1.9}$$

$$w \sim \left(\frac{C'\mu'}{E'}\right)^{1/4} \frac{l^{3/4}}{t^{1/8}}\gamma^{-3/4}\left\{\tilde{C}_{01}\left(1 - \frac{x}{l}\right)^{5/8} + \mathcal{B}_1\left(1 - \frac{x}{l}\right)^{1/8} + \mathcal{B}_2\mathcal{P}_{cm}^{-1/4}\left(1 - \frac{x}{l}\right)^{3/4} + \mathcal{B}_3\left(1 - \frac{x}{l}\right)^{r}\right\}, \tag{1.10}$$

where $\tilde{C}_{01}$, $\mathcal{B}_1$, $\mathcal{B}_2$, $\mathcal{B}_3$, and r are constants determined in the solution process, and γ is an $O(1)$ quantity introduced in the rescaling below. The three key parameter combinations

$$\mathcal{P}_{km} := \frac{K'^3}{\mu' E'^2}\frac{\gamma^{3/2}t}{l^{3/2}}, \qquad \mathcal{P}_{cm} := \frac{C'^3 E'}{\mu'}\frac{\gamma^3 t^{5/2}}{l^3}, \qquad \mathcal{P}_{ckm} := \frac{K'^4}{C'\mu' E'^3}\frac{\gamma t^{1/2}}{l} \tag{1.11}$$

characterize the different behavior regimes, as shown in the analysis. The leading order term in (1.10) was established by [36] for the stationary solution and then confirmed in [6] for zero toughness. In a preliminary study [9], which considers the infinite limit of a nondimensional parameter for the volume of injected fluid (2.2), the first and last terms in both (1.9) and (1.10) are also determined. However, the other terms are not found there, since certain parameter combinations are fixed (see section 3.2 for further discussion.) These additional terms allow us to analytically construct a uniform solution near the tip, instead of numerically as in [24] for zero leak-off. Near and far-field solutions for semi-infinite approximations of the fracture [23, 35, 34] also use expansions in powers of $(1-\xi)$, which include some of the powers from (1.9)–(1.10) in addition to other terms related to the semi-infinite limit. Power law expansions similar to (1.9)–(1.10) for the zero leak-off case are derived in [44]. Some of these terms are also determined in [6, 24], but the additional terms found in [44] allow the uniform tip behavior to be constructed analytically.

The asymptotic expansions (1.9)–(1.10) explicitly identify the critical parameter combinations (1.11) that dictate transitions between (1.9) and (1.10) in the tip vicinity. These quantities are combinations of the dimensionless parameters that arise in the rescaling below, which quantify the physical processes viscosity, toughness, and leak-off. From the construction of expansions (1.9)–(1.10) we can understand the changes in tip behavior as we scale the quantities (1.11) with a parameter related to the distance from the tip $\xi = 1$. Our method does not use a semi-infinite approximation, and therefore can be extended to study additional time dependencies, transients, and other types of hydraulic fractures, such as finger-like geometries, known as the PKN fracture [45, 48].

In section 2 we describe the new approach and in section 3 obtain expansions when all three processes play a role, in the case that leak-off dictates the leading order behavior. The construction leads to the identification of the parameter combinations (1.11) which are necessary for describing the transition between the near- and intermediate-tip solutions (1.9)–(1.10). Section 4 summarizes our results and briefly outlines extensions of our methodology to situations where time-dependence must be scaled explicitly, or cases where there is more than one transition in the dominant shape of the fracture.

2. Approach of the new method. We introduce the nondimensional quantities, following [6, 26, 24] and others:

$$\xi = x/l, \qquad l = L\gamma, \qquad w = \epsilon L\Omega, \qquad p = \epsilon E'\Pi. \tag{2.1}$$

It is convenient to work with the dimensionless quantities Ω (the opening), Π (the net pressure), and γ (a fracture length), which are all $O(1)$. The parameter ϵ is used in [6, 26, 24] to relate w/l to p/E', so for comparison purposes we include it in our analysis; however, it plays no role here and so could be set to unity. Also, L denotes a length scale and is of the same order as the fracture length l.

We also define four nondimensional quantities,

$$\mathcal{G}_v = \frac{Q_0 t}{\epsilon L^2}, \qquad \mathcal{G}_m = \frac{\mu'}{\epsilon^3 E' t}, \qquad \mathcal{G}_k = \frac{K'}{\epsilon E' L^{1/2}}, \qquad \mathcal{G}_c = \frac{C' t^{1/2}}{\epsilon L}, \tag{2.2}$$

and determine different solutions depending on the size of combinations of these parameters (1.11) without setting any to unity. The governing equations (1.3)–(1.7) are now

$$\frac{t(\epsilon L)_t}{\epsilon L}\Omega + \dot{\Omega} t - \xi \frac{t(L\gamma)_t}{L\gamma}\frac{\partial \Omega}{\partial \xi} + \mathcal{G}_c \Gamma_l = \frac{1}{\mathcal{G}_m}\frac{1}{\gamma^2}\frac{\partial}{\partial \xi}\left[\Omega^3 \frac{\partial \Pi}{\partial \xi}\right], \tag{2.3}$$

$$\Pi = -\frac{1}{4\pi\gamma}\int_{-1}^{1} \frac{\partial \Omega}{\partial \chi}\frac{\mathrm{d}\chi}{\chi - \xi}, \tag{2.4}$$

$$\Omega = \mathcal{G}_k \gamma^{1/2}(1 \mp \xi)^{1/2}, \quad \xi \longrightarrow \pm 1; \qquad \Omega = 0, \quad \Omega^3 \frac{\partial \Pi}{\partial \xi} = 0, \quad \text{at } \xi = 1^{\pm}, \tag{2.5}$$

$$\mathcal{G}_v = \gamma \int_{-1}^{1} \Omega \, \mathrm{d}\chi + \mathcal{G}_c \frac{1}{L}\int_0^1 l(\theta t)\theta^{-1/2} \int_{-1}^{1} \Gamma_l \, \mathrm{d}\chi \, \mathrm{d}\theta, \tag{2.6}$$

where Γ_l is the dimensionless leak-off function discussed below.

The new approach relies on two main ingredients: (i) a scaling parameter $\delta \ll 1$ that relates distance from the tip to the key dimensionless quantities in (2.2), and (ii) a flexible asymptotic expansion which can handle behavior dominated by different physical quantities. Thus we define

$$1 - \delta z = \xi, \tag{2.7}$$

$$\mathcal{G}_v = \delta^{\beta_v}\hat{\mathcal{G}}_v, \quad \mathcal{G}_k = \delta^{\beta_k}\hat{\mathcal{G}}_k, \quad \mathcal{G}_m = \delta^{\beta_m}\hat{\mathcal{G}}_m, \quad \mathcal{G}_c = \delta^{\beta_c}\hat{\mathcal{G}}_c, \tag{2.8}$$

where the $\hat{\mathcal{G}}_{()}$'s are $O(1)$ quantities. The different regimes are then characterized by inequalities between the values of the exponents β_v, β_k, β_m, β_c. Here δ is introduced as a bookkeeping parameter that disappears from the expansion in the end. We assume that z is $O(1)$ and so $\delta \ll 1$ essentially describes the distance from the tip $\xi = 1$. Through this scaling we can explore the dominant behavior of the propagating fracture in a very general way: we have not yet specified the distance from the tip and we do not make a semi-infinite approximation, as used in previous studies such as [6, 26, 24], amongst others.

Because of the symmetry of the solution about $\xi = 0$ (see Figure 1.1), we can restrict our attention to the interval $0 < \xi < 1$ and write the equations in terms of z. The integral in (2.4) is then written as

$$\Pi = -\frac{1}{4\pi\gamma}\int_{-1}^{1} \frac{\mathrm{d}\Omega}{\mathrm{d}\chi}\frac{\mathrm{d}\chi}{\chi - \xi} = -\frac{1}{2\pi\gamma}\int_0^1 \frac{\mathrm{d}\Omega}{\mathrm{d}\chi}\frac{\chi \, \mathrm{d}\chi}{\chi^2 - \xi^2}, \tag{2.9}$$

and similarly for the integrals in (2.6).

Then applying (2.7) and (2.8) in the governing equations (2.3)–(2.6) yields

$$\frac{t(\epsilon L)_t}{\epsilon L}\Omega + (1-\delta z)\frac{t(L\gamma)_t}{L\gamma}\delta^{-1}\frac{\mathrm{d}\Omega}{\mathrm{d}z} + \hat{\mathcal{G}}_c\delta^{\beta_c}\Gamma_l = \frac{1}{\hat{\mathcal{G}}_m\gamma^2}\delta^{-\beta_m-2}\frac{\mathrm{d}}{\mathrm{d}z}\left[\Omega^3\frac{\mathrm{d}\Pi}{\mathrm{d}z}\right], \tag{2.10}$$

$$\Pi = -\frac{1}{2\pi\gamma}\delta^{-1}\int_0^{1/\delta}\frac{\mathrm{d}\Omega}{\mathrm{d}r}\frac{(1-\delta r)}{r(2-\delta r)-z(2-\delta z)}\,\mathrm{d}r, \tag{2.11}$$

$$\Omega = \hat{\mathcal{G}}_k\gamma^{1/2}\delta^{\beta_k+1/2}z^{1/2}, \quad z\longrightarrow 0; \qquad \Omega = 0, \ \ \Omega^3\frac{\mathrm{d}\Pi}{\mathrm{d}z} = 0, \quad \text{at } z=0, \tag{2.12}$$

$$\delta^{\beta_v}\hat{\mathcal{G}}_v = 2\gamma\delta\int_0^{1/\delta}\Omega\,\mathrm{d}r + 2\frac{\hat{\mathcal{G}}_c}{L}\delta^{\beta_c+1}\int_0^1 l(\theta t)\theta^{-1/2}\int_0^{1/\delta}\Gamma_l\,\mathrm{d}r\,\mathrm{d}\theta. \tag{2.13}$$

Here we look for self-similar solutions $\Omega = \Omega(z)$ and $\Pi = \Pi(z)$. We justify the use of these types of solutions in section 3.3. In (2.10)–(2.13) the powers of δ appear explicitly, and they play a central role in understanding the spatial behavior of the solution relative to the key dimensionless quantities.

For $\delta \ll 1$, we expand the solution Ω and Π as follows:

$$\Omega = \delta^{\beta_k+1/2}\big(\Omega_{00} + \delta^{\alpha_1}\Omega_{01} + \delta^{\alpha_2}\Omega_2 + \cdots\big), \tag{2.14}$$

$$\Pi = \delta^{\beta_k}\Pi_{00} + \delta^{\sigma_1}\Pi_{01} + \delta^{\sigma_2}\Pi_2 + \cdots, \tag{2.15}$$

where the exponents α_i and σ_i are determined in terms of the exponents $\beta_{()}$ in (2.8) as part of the method. The prefactor δ^{β_k} corresponds to the dimensionless parameter $\mathcal{G}_k$, and its inclusion in the leading terms is discussed below.

We substitute Ω and Π into the lubrication and elasticity equations (2.10) and (2.11), so that they become, respectively,

$$\begin{aligned}
&\frac{t(\epsilon L)_t}{\epsilon L}\delta^{\beta_k+1/2}\big(\Omega_{00}+\delta^{\alpha_1}\Omega_{01}+\cdots\big)\\
&\quad + (1-\delta z)\frac{t(L\gamma)_t}{L\gamma}\delta^{\beta_k-1/2}\left(\frac{\mathrm{d}\Omega_{00}}{\mathrm{d}z}+\delta^{\alpha_1}\frac{\mathrm{d}\Omega_{01}}{\mathrm{d}z}+\cdots\right) + \hat{\mathcal{G}}_c\delta^{\beta_c}\Gamma_l\\
&\quad = \frac{1}{\hat{\mathcal{G}}_m\gamma^2}\delta^{-\beta_m-1/2+3\beta_k}\frac{\mathrm{d}}{\mathrm{d}z}\left[\big(\Omega_{00}+\delta^{\alpha_1}\Omega_{01}+\big)^3\left(\delta^{\beta_k}\frac{\mathrm{d}\Pi_{00}}{\mathrm{d}z}+\delta^{\sigma_1}\frac{\mathrm{d}\Pi_{01}}{\mathrm{d}z}+\cdots\right)\right],
\end{aligned} \tag{2.16}$$

$$\begin{aligned}
&\delta^{\beta_k}\Pi_{00}+\delta^{\sigma_1}\Pi_{01}+\cdots\\
&\quad = -\frac{1}{2\pi\gamma}\delta^{\beta_k-1/2}\int_0^{1/\delta}\left(\frac{\mathrm{d}\Omega_{00}}{\mathrm{d}r}+\delta^{\alpha_1}\frac{\mathrm{d}\Omega_{01}}{\mathrm{d}r}+\cdots\right)\frac{(1-\delta r)}{r(2-\delta r)-z(2-\delta z)}\,\mathrm{d}r.
\end{aligned} \tag{2.17}$$

The nondimensionalized leak-off function Γ_l is defined as

$$\Gamma_l = \frac{1}{\sqrt{1-t_0(\xi l)/t}} = \frac{1}{\sqrt{1-\xi^{1/\lambda}}} = \frac{1}{\sqrt{1-(1-\delta z)^{1/\lambda}}}, \tag{2.18}$$

where $t_0(\cdot)$ is defined following (1.8), and we have written it in the rescaled coordinates (2.7). This follows from the definition of $t_0(x)$, the time lapsed between the current time t and the time at which the crack tip arrived at the point x, so that $x(t_0(x)) = l(t_0)$. Using $l = \gamma L$, $L = C_L t^\lambda$, and $\xi = x/l$, we find

$$l(t_0) = C_L\gamma\left(\frac{t_0}{t}\right)^\lambda t^\lambda \qquad \text{and} \qquad \frac{t_0}{t} = \xi^{1/\lambda}. \tag{2.19}$$

To motivate the equations which are solved below to leading order in the different regions, we briefly consider the lubrication equation in the form of (2.10). The leading order terms for $\delta \ll 1$ satisfy

$$\lambda\delta^{-1}\frac{\mathrm{d}\Omega}{\mathrm{d}z} + \frac{\hat{\mathcal{G}}_c\delta^{\beta_c-1/2}}{\sqrt{1-(1-\delta z)^{1/\lambda}}} = \frac{1}{\hat{\mathcal{G}}_m\gamma^2}\delta^{-\beta_m-2}\frac{\mathrm{d}}{\mathrm{d}z}\left[\Omega^3\frac{\mathrm{d}\Pi}{\mathrm{d}z}\right]. \tag{2.20}$$

The form of the expansions (2.14) and (2.15) distinguish where toughness dominates, with the sign of α_1 playing a significant role. We consider three cases:

(i) The right-hand side of (2.20) is dominant and set equal to zero. This gives $\Pi = \text{constant}$ to leading order (see Appendix A.1). Then Ω is found using the propagation condition (2.12), giving the leading order square root behavior for Ω_{00} and $\Pi_{00} = \text{constant}$. This case is described by $\alpha_1 > 0$, where toughness dominates the leading order behavior, and thus justifies the use of the prefactor δ^{β_k} in expansions (2.14) and (2.15). The details of this *near-tip* case are given in section 3.1 below.

(ii) The first term on the left-hand side balances with the right-hand side, thus neglecting the term with coefficient $\mathcal{G}_c$ to leading order. This case is described by $\alpha_1 < 0$, where viscosity, not leak-off, dictates the leading order behavior of $z^{2/3}$ [6, 15, 24]. When $\alpha_1 < 0$, the ordering of the terms in (2.14) and (2.15) changes; then Ω_{01} and Π_{01} become leading order and so Ω_{00} and Π_{00} are zero in regions where the toughness is not dominant.

(iii) The second term on the left-hand side matches the right-hand side. We obtain the solution $z^{5/8}$ to leading order, as in [6, 36]. This situation also holds for $\alpha_1 < 0$, with both leak-off and viscosity dictating the leading order behavior. Again, Ω_{00} and Π_{00} are zero, and Ω_{01} and Π_{01} are the leading order terms. The details of this *intermediate-tip* case are given in section 3.1 below. In section 3.3 we show that $\lambda = 1/2$ for sufficiently large time, as in [6], and so we use $\Gamma_l = 1/\sqrt{\delta z(2-\delta z)}$ in (2.18) for this case.

3. The expansion including toughness, leak-off, and viscosity. We consider the case with nonzero leak-off $\mathcal{G}_c \neq 0$ in addition to nonzero toughness and viscosity ($\mathcal{G}_k \neq 0$ and $\mathcal{G}_m \neq 0$.) The expansions (2.14) and (2.15) are used to determine exponents by balancing terms, leading to important combinations of the parameters in (1.11) which characterize the different cases. We focus on scenarios with significant leak-off, leading to the study of transitions between regimes where toughness and leak-off dominate the behavior, namely, cases (i) and (iii). The analysis identifies a critical scaling involving all three processes and gives a parametric characterization for significant leak-off as $\mathcal{G}_c^3/\mathcal{G}_m = O(1)$ or larger. In contrast, case (ii) occurs in regimes where leak-off plays a secondary role, $\mathcal{G}_c^3/\mathcal{G}_m \ll 1$ and $\mathcal{G}_c < \mathcal{G}_k$, and corresponds to the purely viscosity-dominated case with $z^{2/3}$ power law to leading order away from the tip. A straightforward extension of the analysis in [44] of the impermeable case can be used to include higher order corrections involving leak-off in this parameter region, so we do not consider it here. The intermediate case where $\mathcal{G}_c^3/\mathcal{G}_m \ll 1$ and $\mathcal{G}_c < \mathcal{G}_k$ has two transition regions, from the $z^{1/2}$ to $z^{5/8}$ to $z^{2/3}$ behavior. The analyses presented below and in [44] can be extended easily to treat these two transitions, and we outline this situation in section 4.

3.1. Local expansions. *Near-tip behavior* ($\alpha_1 > 0$): The leading order terms in (2.16) and (2.17) for $\delta \ll 1$ satisfy

$$0 = \delta^{-\beta_m - 1/2 + 4\beta_k} \frac{\mathrm{d}}{\mathrm{d}z}\left[\Omega_{00}^3 \frac{\mathrm{d}\Pi_{00}}{\mathrm{d}z}\right], \tag{3.1}$$

$$\delta^{\beta_k}\Pi_{00} = -\frac{1}{2\pi\gamma}\delta^{\beta_k - 1/2}\int_0^{1/\delta} \frac{\mathrm{d}\Omega_{00}}{\mathrm{d}r}\frac{(1-\delta r)}{r(2-\delta r) - z(2-\delta z)}\,\mathrm{d}r. \tag{3.2}$$

From these two equations we deduce that the solution of Ω_{00} is

$$\Omega_{00}(z) = C_{00}\sqrt{z(2-\delta z)}, \tag{3.3}$$

where $C_{00} = 4\pi\Pi_{00}$, and $\Pi_{00} =$ constant, as discussed in Appendix A.1. The expression (3.3) is the eigenfunction solution which, when substituted into (3.2), gives $\Pi_{00} =$ constant exactly; its leading order behavior matches the tip condition (2.12) and it is symmetric about $\xi = 0$. We use the tip condition (2.12) to find $C_{00} = \hat{\mathcal{G}}_k\sqrt{\gamma/2}$. Note that for $\alpha_1 > 0$, the leading order term in the expansion for Ω involves the rescaled toughness parameter $\mathcal{G}_k$.

The next order terms for $\delta \ll 1$ in (2.16) and (2.17) satisfy

$$\lambda\delta^{\beta_k - 1/2}\frac{\mathrm{d}\Omega_{00}}{\mathrm{d}z} + \hat{\mathcal{G}}_c\sqrt{\lambda}\delta^{\beta_c - 1/2}z^{-1/2} = \frac{1}{\hat{\mathcal{G}}_m\gamma^2}\delta^{-\beta_m + 3\beta_k - 1/2 + \sigma_1}\frac{\mathrm{d}}{\mathrm{d}z}\left[\Omega_{00}^3\frac{\mathrm{d}\Pi_{01}}{\mathrm{d}z}\right], \tag{3.4}$$

$$\delta^{\sigma_1}\Pi_{01} = -\frac{1}{2\pi\gamma}\delta^{\beta_k - 1/2 + \alpha_1}\int_0^{1/\delta}\frac{\mathrm{d}\Omega_{01}}{\mathrm{d}r}\frac{(1-\delta r)}{r(2-\delta r) - z(2-\delta z)}\,\mathrm{d}r. \tag{3.5}$$

For the moment we do not balance exponents of δ, but leave them in the expression. Solving (3.4) and (3.5) gives Ω_{01} and Π_{01}: we find $\Omega_{01}(z) = C_{01}z$ and Π_{01} satisfies

$$\Pi_{01} = -\frac{C_{01}}{4\pi\gamma}\ln\left(1 - \frac{1}{2-\delta z}\right) - \frac{C_{01}}{4\pi\gamma}\left[\ln(1-\delta z) - \ln(\delta z)\right] \tag{3.6}$$

for $z = O(1)$. The constant of integration in (3.4) is zero; otherwise the solution for Π_{01} yields an infinite stress intensity factor, as shown in Appendix A.2. The details of the calculation of Π_{01} are similar to the analysis in Appendix B. We now determine C_{01} by substituting Π_{01} into (3.4), and considering the leading order terms only. Hence

$$\delta^{\sigma_1}C_{01} = \hat{\mathcal{G}}_m\gamma^2\left[\delta^{\beta_m - 2\beta_k}\lambda\frac{2\pi\gamma}{C_{00}^2} + \delta^{\beta_c + \beta_m - 3\beta_k}\frac{4\pi\gamma\sqrt{\lambda}\hat{\mathcal{G}}_c}{\sqrt{2}C_{00}^3}\right]. \tag{3.7}$$

Notice that the terms on the right-hand side of (3.7) change order depending on the relative magnitude of $\mathcal{G}_k$ and $\mathcal{G}_c$ (i.e., δ^{β_k} and δ^{β_c}.) If $\mathcal{G}_c \ll \mathcal{G}_k$, then the first term on the right-hand side is dominant and we obtain

$$\alpha_1 = 1/2 + \beta_m - 3\beta_k, \qquad \sigma_1 = \beta_m - 2\beta_k, \qquad C_{01} = 2\pi\lambda\frac{\hat{\mathcal{G}}_m\gamma^3}{C_{00}^2}. \tag{3.8}$$

However, if $\mathcal{G}_c \gg \mathcal{G}_k$, then the second term on the right-hand side of (3.7) is dominant. The exponents α_1 and σ_1 and coefficient C_{01} are now

$$\alpha_1 = 1/2 + \beta_c + \beta_m - 4\beta_k, \qquad \sigma_1 = \beta_c + \beta_m - 3\beta_k, \qquad C_{01} = \frac{4\pi\sqrt{\lambda}\hat{\mathcal{G}}_c\hat{\mathcal{G}}_m\gamma^3}{\sqrt{2}C_{00}^3}. \tag{3.9}$$

Since both of these cases represent particular solutions to the linear equations (3.4)–(3.5), we can treat them simultaneously. Thus if $\alpha_1 > 0$, the first three terms in the expansion (2.14) for Ω are

$$\Omega \sim \delta^{\beta_k+1/2}\left[C_{00}\sqrt{z(2-\delta z)} + \delta^{1/2+\beta_m-3\beta_k}C_{01}z + \delta^{1/2+\beta_c+\beta_m-4\beta_k}C_{02}z\right], \tag{3.10}$$

where C_{02} is the redefined C_{01} coefficient from (3.9). Then the second term dominates over the third term when $\mathcal{G}_c \ll \mathcal{G}_k$, and vice versa when $\mathcal{G}_c \gg \mathcal{G}_k$.

Intermediate-tip behavior ($\alpha_1 < 0$): The leading order terms in (2.16) are

$$\lambda\delta^{\beta_k-1/2+\alpha_1}\frac{\mathrm{d}\Omega_{01}}{\mathrm{d}z} + \frac{\hat{\mathcal{G}}_c\delta^{\beta_c-1/2}}{\sqrt{z(2-\delta z)}} = \frac{\delta^{-\beta_m+3\beta_k-1/2+3\alpha_1+\sigma_1}}{\hat{\mathcal{G}}_m\gamma^2}\frac{\mathrm{d}}{\mathrm{d}z}\left[\Omega_{01}^3\frac{\mathrm{d}\Pi_{01}}{\mathrm{d}z}\right], \tag{3.11}$$

coupled with the elasticity equation (3.5). Note that we now use the form of Γ_l with $\lambda = 1/2$, as mentioned in case (iii) above, since we are in the leak-off–dominated regime. In the case that $\mathcal{G}_c^3/\mathcal{G}_m \gg (1-\xi)^{1/2}$, the second term on the left-hand side of (3.11) is dominant over the term with $\Omega_{01}'(z)$. Then balancing powers of δ in (3.11) and the elasticity equation (3.5) gives

$$\alpha_1 = 1/8 + (\beta_c+\beta_m)/4 - \beta_k, \qquad \sigma_1 = -3/8 + (\beta_c+\beta_m)/4. \tag{3.12}$$

Note that if the first term in (3.11) is dominant, then the leading order terms for Ω_{01} are the same as in the case of zero leak-off studied in [6, 15, 24, 44]. This corresponds to the case $\delta^{1/8+(\beta_m+\beta_c)/4} > \delta^{\beta_c}$ or, equivalently, $\mathcal{G}_c^3/\mathcal{G}_m \ll (1-\xi)^{1/2}$. Ignoring this term to leading order is consistent with the fact that leak-off is dominant, i.e., $\mathcal{G}_c^3/\mathcal{G}_m = O(1)$ or larger, as discussed in section 3.

The form of Ω_{01} is written as a combination of powers in z, namely,

$$\Omega_{01} = \tilde{C}_{01}z^q + \hat{\mathcal{B}}_1 z^g + \hat{\mathcal{B}}_2 z^p + \hat{\mathcal{B}}_3 z^r. \tag{3.13}$$

This is equivalent to a perturbation expansion for the solution of (3.11) when $\hat{\mathcal{B}}_i \ll 1$, which we verify below. The elasticity equation (3.5) is then solved to give

$$\begin{aligned}\Pi_{01} = {} & \cot\pi q\frac{\tilde{C}_{01}q}{4\gamma}z^{q-1} + \cot\pi g\frac{\hat{\mathcal{B}}_1 g}{4\gamma}z^{g-1} + \cot\pi p\frac{\hat{\mathcal{B}}_2 p}{4\gamma}z^{p-1} + \cot\pi r\frac{\hat{\mathcal{B}}_3 r}{4\gamma}z^{r-1} \\ & + \frac{\tilde{C}_{01}q\delta^{1-q}(2-\delta z)^{-1}}{4\pi\gamma} + \frac{\hat{\mathcal{B}}_1 g\delta^{1-g}(2-\delta z)^{-1}}{4\pi\gamma} + \frac{\hat{\mathcal{B}}_2 p\delta^{1-p}(2-\delta z)^{-1}}{4\pi\gamma} \\ & + \frac{\hat{\mathcal{B}}_3 r\delta^{1-r}(2-\delta z)^{-1}}{4\pi\gamma} + O(\delta^{1-q},\delta^{1-g},\delta^{1-p},\delta^{1-r}).\end{aligned} \tag{3.14}$$

By using the definitions of α_1 and σ_1 from (3.12) we integrate (3.11) to obtain

$$\lambda\delta^{1/8+\beta_m/4-3\beta_c/4}\Omega_{01} + \hat{\mathcal{G}}_c\sqrt{2}\left(z^{1/2} + \frac{1}{12}\delta z^{3/2}\right) + k = \frac{1}{\hat{\mathcal{G}}_m\gamma^2}\Omega_{01}^3\frac{\mathrm{d}\Pi_{01}}{\mathrm{d}z}, \tag{3.15}$$

where k is an arbitrary constant. We then have four expressions enabling us to determine q, g, p, and r. The leading order terms are

$$\sqrt{2}\hat{\mathcal{G}}_c z^{1/2} = \cot\pi q\frac{\tilde{C}_{01}^4 q(q-1)}{4\hat{\mathcal{G}}_m\gamma^3}z^{4q-2}, \tag{3.16}$$

which yields $q = 5/8$ and gives the coefficient $\tilde{C}_{01}$,

$$\tilde{C}_{01} = \left\{ \frac{4\sqrt{2}\hat{\mathcal{G}}_c\hat{\mathcal{G}}_m\gamma^3}{q(q-1)\cot\pi q} \right\}^{1/4}. \tag{3.17}$$

Then the next order terms satisfy

$$k = \frac{\tilde{C}_{01}^3\hat{\mathcal{B}}_1}{4\hat{\mathcal{G}}_m\gamma^3}\big[g(g-1)\cot\pi g + 3q(q-1)\cot\pi q\big] z^{3q+g-2}, \tag{3.18}$$

$$\lambda\delta^{1/8+(\beta_m-3\beta_c)/4}\tilde{C}_{01}z^q = \frac{\tilde{C}_{01}^3\hat{\mathcal{B}}_2}{4\hat{\mathcal{G}}_m\gamma^3}\big[p(p-1)\cot\pi p + 3q(q-1)\cot\pi q\big] z^{3q+p-2}, \tag{3.19}$$

$$\lambda\delta^{1/8+(\beta_m-3\beta_c)/4}\hat{\mathcal{B}}_1z^g = \frac{\tilde{C}_{01}^3\hat{\mathcal{B}}_3}{4\hat{\mathcal{G}}_m\gamma^3}\big[r(r-1)\cot\pi r + 3q(q-1)\cot\pi q\big] z^{3q+r-2}. \tag{3.20}$$

For simplicity of presentation we assume $\delta^{1/8+(\beta_m-3\beta_c)/4} > \delta$, which means we can neglect the $z^{3/2}$ term in (3.15). This corresponds to the case $\mathcal{G}_c^3/\mathcal{G}_m \ll (1-\xi)^{-7/2}$. Including this term results in an additional contribution to Ω_{01} in (3.13) with power law $z^{13/8}$. This term can be shown to be higher order in the matching below and could be included in a straightforward manner if $\mathcal{G}_c^3/\mathcal{G}_m = O((1-\xi)^{-7/2})$ or larger.

Matching exponents of z in (3.18)–(3.19) gives $g = 1/8$ and $p = 3/4$, respectively. Since $\hat{\mathcal{B}}_1$ is assumed small, the left-hand side of (3.20) is of higher order; r must be approximated by solving the following nonlinear algebraic equation corresponding to the right-hand side of (3.20) vanishing:

$$r(r-1)\cot\pi r + 3q(q-1)\cot\pi q = 0, \tag{3.21}$$

and we find that $r \approx 0.0699928$. Then

$$\hat{\mathcal{B}}_2 = \delta^{1/8+(\beta_m-3\beta_c)/4}\mathcal{B}_2, \qquad \mathcal{B}_2 = \lambda\frac{4\hat{\mathcal{G}}_m\gamma^3}{\tilde{C}_{01}^2\big[p(p-1)cot\pi p + 3q(q-1)cot\pi q\big]}, \tag{3.22}$$

and the coefficients $\hat{\mathcal{B}}_1$ and $\hat{\mathcal{B}}_3$ are determined by matching with the near-tip expansion (3.10) in the next section. Since $q,\ g,\ p,\ r \neq 1/2$, the solution (3.13) cannot satisfy the $\sqrt{z}$ behavior from the propagation condition (2.12) for $\hat{\mathcal{G}}_k > 0$. Observe that for $\alpha_1 < 0$, toughness does not dominate the leading order behavior in this regime and the terms $\Omega_{00} = \Pi_{00} = 0$. Then the first term in the expansion (2.14) for Ω is

$$\Omega \sim \delta^{(\beta_c+\beta_m)/4}\left[\tilde{C}_{01}(\delta z)^{5/8} + \mathcal{B}_1(\delta z)^{1/8} + \mathcal{B}_2\delta^{(\beta_m-3\beta_c)/4}\frac{\hat{\mathcal{G}}_m^{1/4}}{\hat{\mathcal{G}}_c^{3/4}}(\delta z)^{3/4} + \mathcal{B}_3(\delta z)^r\right] \tag{3.23}$$

for $\alpha_1 < 0$. Note that we have redefined $\tilde{C}_{01}$ and $\mathcal{B}_2$ without the $\hat{\mathcal{G}}_()$ terms, which allows us to highlight explicitly the dependence of Ω on the key dimensionless quantities $\mathcal{G}_()$ in (2.2). The coefficients $\hat{\mathcal{B}}_1$ and $\hat{\mathcal{B}}_3$ are redefined as $\mathcal{B}_1$ and $\mathcal{B}_3$ and they are found in the following section.

3.2. Transition in spatial behavior and matching. Now we compare the two local expansions, (3.10) and (3.23), in terms of the parameter combinations

$$\mathcal{P}_{km} = \frac{\mathcal{G}_k^3}{\mathcal{G}_m}, \qquad \mathcal{P}_{cm} = \frac{\mathcal{G}_c^3}{\mathcal{G}_m}, \qquad \mathcal{P}_{ckm} = \frac{\mathcal{G}_k^4}{\mathcal{G}_c\mathcal{G}_m}, \tag{3.24}$$

which appear explicitly in both expansions and were introduced in (1.11). We consider the range of parameters for which either $\mathcal{P}_{cm} \gg (1-\xi)^{1/2}$ or $\mathcal{P}_{km} \gg (1-\xi)^{1/2}$, i.e., parameter values away from the viscosity-dominated regime. For $\mathcal{G}_c = O(1)$ this corresponds to $\mathcal{G}_m \ll 1$ or $\beta_m > 1$. Other situations are discussed in the next section.

Then the expansions for Ω in terms of $\mathcal{P}_{km}$, $\mathcal{P}_{cm}$, and $\mathcal{P}_{ckm}$ are

$$\Omega \sim \mathcal{G}_k \left[C_{00}\sqrt{1-\xi^2} + C_{01}\mathcal{P}_{km}^{-1}(1-\xi) + C_{02}\mathcal{P}_{ckm}^{-1}(1-\xi) \right] \quad \text{for } \alpha_1 > 0, \tag{3.25}$$

$$\Omega \sim (\mathcal{G}_c\mathcal{G}_m)^{1/4} \left[\tilde{C}_{01}(1-\xi)^{5/8} + \mathcal{B}_1(\mathcal{P}_{ckm})(1-\xi)^{1/8} + \mathcal{P}_{cm}^{-1/4}\mathcal{B}_2(1-\xi)^{3/4} + \mathcal{B}_3(\mathcal{P}_{ckm})(1-\xi)^{r} \right] \quad \text{for } \alpha_1 < 0. \tag{3.26}$$

We have redefined C_{00}, C_{01}, and C_{02} in (3.25) without the $\hat{\mathcal{G}}_{()}$ terms which are incorporated into the dimensionless $\mathcal{G}_{()}$ terms. Also, observe that the δ's have disappeared from the expressions. The expansions (3.25)–(3.26) give a transition in behavior of Ω for $1-\xi = O(\mathcal{P}_{ckm}^2)$: the unknown coefficients $\mathcal{B}_1(\mathcal{P}_{ckm})$ and $\mathcal{B}_3(\mathcal{P}_{ckm})$ are determined by matching the expansions in this transition region. The leading order term in (3.26) was given in [36] and later in [6] for vanishing toughness. In a preliminary study [9] some of the terms in (3.26) are obtained. There, both the global balance and lubrication equations are scaled by ${\mathcal{G}_v}^{-1}$, and they consider the limit of small toughness, with $\mathcal{G}_v \to \infty$, for fixed nondimensional parameters $\mathcal{G}_c/\mathcal{G}_v = \mathcal{G}_m\mathcal{G}_v = 1$. Then some of the terms in (3.26) are excluded for large $\mathcal{G}_v$.

The motivation for defining the parameter $\mathcal{P}_{ckm}$ follows directly from the expression for α_1 in both cases, i.e., (3.9) and (3.12). Since the $\hat{\mathcal{G}}_{()}$ quantities are $O(1)$, the condition $\alpha_1 > (<)\, 0$ can be rewritten as $\mathcal{P}_{ckm} \gg (\ll)\,(1-\xi)^{1/2}$. The solution for $\alpha_1 > 0$ is physically significant in the toughness dominated regime ($\mathcal{G}_k \gg \mathcal{G}_c$), which is close to the tip and corresponds to $\mathcal{P}_{ckm} \gg (1-\xi)^{1/2}$. As $(1-\xi)^{7/2}$ approaches $\mathcal{P}_{ckm}$, the first and third terms in (3.25) and the first term in (3.26) are the same order of magnitude. A transition occurs in the region $(1-\xi) = O(\mathcal{P}_{ckm}^{1/2})$ and the solution in the intermediate-tip region is found by considering $\alpha_1 < 0$ in (3.26). This corresponds to the leak-off dominated regime, which is away from the tip for $\mathcal{P}_{ckm} \ll (1-\xi)^{1/2}$. Hence the expansion in (3.25) holds for $1-\xi < \mathcal{P}_{ckm}^2$ and the expansion in (3.26) holds for $1-\xi = O(\mathcal{P}_{ckm}^s)$ (see Figure 3.1), with $0 < s < 2$, and $\mathcal{P}_{ckm} \ll 1$.

To construct a uniform asymptotic approximation by matching (3.25) and (3.26), we note that they are obtained by solving (2.20) in different asymptotic limits. The matching is therefore straightforward in the transition region where $\mathcal{G}_k^4/(\mathcal{G}_c\mathcal{G}_m) = O((1-\xi)^{1/2})$; to leading order the solution satisfies (2.20) together with the propagation condition (2.12). While there is no closed form solution in this region, it can be constructed numerically where $1-\xi = O(\mathcal{P}_{ckm}^2)$ for $0 < \mathcal{P}_{ckm} \ll 1$.

Alternatively, one can give an analytical expression for the matching of (3.25) and (3.26), obtained from solving for the remaining unknown coefficients $\mathcal{B}_1(\mathcal{P}_{ckm})$ and $\mathcal{B}_3(\mathcal{P}_{ckm})$. Writing these expressions in terms of the critical scaling $1-\xi = \mathcal{P}_{ckm}^2\zeta$ for $\zeta = 0(1)$, (3.25) and (3.26) are, respectively,

$$\Omega \sim \mathcal{G}_k \left[C_{00}\mathcal{P}_{ckm}\zeta^{1/2}\sqrt{2-\mathcal{P}_{ckm}^2\zeta} + \left(C_{01}\mathcal{P}_{km}^{-1} + C_{02}\mathcal{P}_{ckm}^{-1}\right)\mathcal{P}_{ckm}^2\zeta \right], \tag{3.27}$$

$$\Omega \sim (\mathcal{G}_c\mathcal{G}_m)^{1/4} \left[\tilde{C}_{01}\mathcal{P}_{ckm}^{5/4}\zeta^{5/8} + \mathcal{B}_1(\mathcal{P}_{ckm})\mathcal{P}_{ckm}^{1/4}\zeta^{1/8} + \mathcal{P}_{cm}^{-1/4}\mathcal{B}_2\mathcal{P}_{ckm}^{3/2}\zeta^{3/4} + \mathcal{B}_3(\mathcal{P}_{ckm})\mathcal{P}_{ckm}^{2r}\zeta^{r} \right]. \tag{3.28}$$

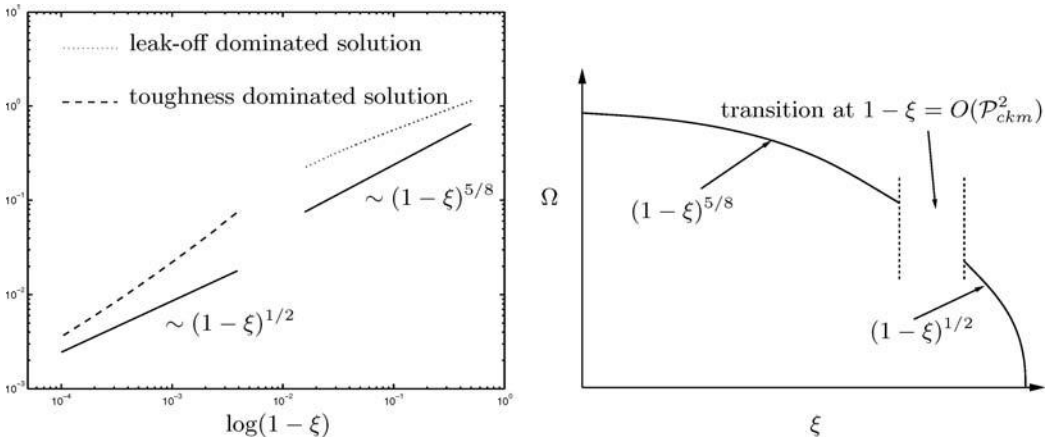


FIG. 3.1. *The left plot shows* $\log\Omega$ *vs.* $\log(1-\xi)$ *with* $\mathcal{G}_c = 1$, $\mathcal{G}_m = 0.2$, *and* $\mathcal{G}_k = 0.35$ *(and so* $\mathcal{P}_{ckm} = 0.075$*.) The solid lines denote the leading order power law solutions, as indicated above. The right plot shows a diagram of the solution* Ω *vs.* ξ *near the fracture tip for the leak-off–dominated regime. The transition region is* $1-\xi = O(\mathcal{P}_{ckm}^2)$.

Equating (3.27) and (3.28) and their first derivatives yields $\mathcal{B}_1(\mathcal{P}_{ckm})$ and $\mathcal{B}_3(\mathcal{P}_{ckm})$, which is equivalent to matching the first two terms in a Taylor series expansion about $1-\xi = O(\mathcal{P}_{ckm}^2)$ where Ω is regular. Figure 3.2 shows solution profiles of Ω, matched at $1-\xi = \mathcal{P}_{ckm}^2$. In these parameter regimes all three processes contribute to the transition between the near- and intermediate-tip behavior, described by (3.25) and (3.26). There the coefficients $\mathcal{B}_1(\mathcal{P}_{ckm})$ and $\mathcal{B}_3(\mathcal{P}_{ckm})$ are

$$\mathcal{B}_1(\mathcal{P}_{ckm}) = \frac{\mathcal{P}_{ckm}}{8r-1}\Big[(8r-4)\sqrt{2}C_{00} + (8r-8)\big(C_{01}\mathcal{P}_{km}^{-1}\mathcal{P}_{ckm} + C_{02}\big) - (8r-5)\tilde{C}_{01} - (8r-6)\mathcal{B}_2\mathcal{P}_{cm}^{-1/4}\mathcal{P}_{ckm}^{1/4}\Big], \tag{3.29}$$

$$\mathcal{B}_3(\mathcal{P}_{ckm}) = \frac{\mathcal{P}_{ckm}^{-2r+5/4}}{8r-1}\Big[3\sqrt{2}C_{00} + 7C_{01}\mathcal{P}_{km}^{-1}\mathcal{P}_{ckm} + 7C_{02} - 4\tilde{C}_{01} - 5\mathcal{B}_2\mathcal{P}_{cm}^{-1/4}\mathcal{P}_{ckm}^{1/4}\Big], \tag{3.30}$$

which are small since $\mathcal{P}_{ckm} \ll 1$. As $\mathcal{P}_{ckm}$ increases we observe that the transition region moves away from the tip. Figure 3.2 also shows two solution profiles when (3.25) holds to leading order for both near- and intermediate-tip behavior for $\mathcal{P}_{ckm} < \mathcal{P}_{km}$ and $\mathcal{P}_{ckm} > \mathcal{P}_{km}$. The shape of Ω depends on whether the second or third term in (3.25) plays a larger role in the correction to the leading order behavior. Also, Figure 3.1 shows a log-log plot of Ω for both the leak-off and toughness-dominated regimes. In the near- and intermediate-tip regions we obtain the asymptotic 1/2 and 5/8 power law solutions, respectively, but in the transition region the correction terms are important, so that the behavior cannot be described by a purely power law solution, also observed in [6, 24] for zero leak-off.

For completeness we write down the near- and intermediate-tip expansions for Π, which are determined using the elasticity equation (2.11). Hence we obtain, for

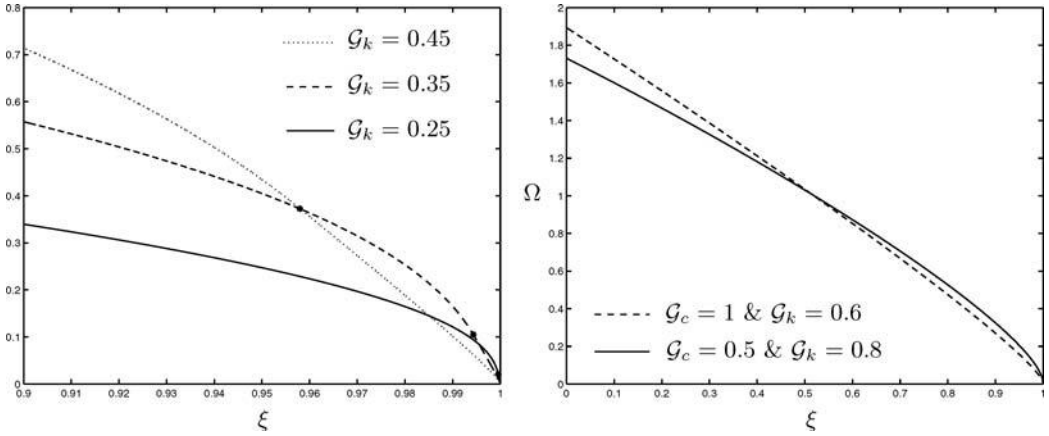


FIG. 3.2. *Solution profiles of* Ω *vs.* ξ. *On the left,* $\mathcal{G}_c = 1$, $\mathcal{G}_m = 0.2$, *and* $\mathcal{G}_k = 0.45, 0.35, 0.25$ *(and* $\mathcal{P}_{ckm} = 0.205, 0.075, 0.020$, *respectively). The transition region is* $1 - \xi = O(\mathcal{P}_{ckm}^2)$, *indicated by* $*$*'s at* $\xi = 1 - \mathcal{P}_{ckm}^2$ *on the graphs. On the right,* $\mathcal{G}_m = 0.1$; *then, for the toughness dominated regime (solid line),* $\mathcal{P}_{ckm} = 8.192$ *and* $\mathcal{P}_{km} = 5.12$, *and for the leak-off dominated regime (dashed line),* $\mathcal{P}_{ckm} = 1.296$ *and* $\mathcal{P}_{km} = 2.16$.

$\mathcal{P}_{ckm} \gg (\ll) (1-\xi)^{1/2}$, respectively,

$$\Pi \sim \mathcal{G}_k \left[\Pi_{00} - \frac{1}{4\pi\gamma} \left(C_{01} \mathcal{P}_{km}^{-1} + C_{02} \mathcal{P}_{ckm}^{-1} \right) \left\{ \ln \left| 1 - \frac{1}{1+\xi} \right| + \ln \left| \frac{1}{1-\xi} \right| + \ln \xi \right\} \right], \tag{3.31}$$

$$\Pi \sim (\mathcal{G}_c \mathcal{G}_m)^{1/4} \left[q \cot \pi q \frac{\tilde{C}_{01}}{4\gamma} (1-\xi)^{-3/8} + g \cot \pi g \frac{\mathcal{B}_1(\mathcal{P}_{ckm})}{4\gamma} (1-\xi)^{-7/8} \right.$$

$$\left. + p \cot \pi p \frac{\mathcal{P}_{cm}^{-1/4} \mathcal{B}_2}{4\gamma} (1-\xi)^{-1/4} + r \cot \pi r \frac{\mathcal{B}_3(\mathcal{P}_{ckm})}{4\gamma} (1-\xi)^{r-1} \right]. \tag{3.32}$$

The details of this calculation are given in Appendix B. In Figure 3.3 we graph Π for different values of $\mathcal{G}_k \ll 1$ with $\mathcal{G}_m = 0.2$ and $\mathcal{G}_c = 1$. We observe that as $\mathcal{P}_{ckm}$ increases, which in this case corresponds to increasing the toughness parameter $\mathcal{G}_k$ since $\mathcal{G}_m$ and $\mathcal{G}_c$ are fixed, the transition point between the two regimes moves away from the tip and Π drops off at a faster rate. This is due to the power law behavior in (3.32) being matched with the near-tip behavior in (3.31), which becomes more dominant through the logarithmic correction for increasing $\mathcal{G}_k$.

3.3. The global volume balance condition. The constants (i.e., the coefficients C_{00}, C_{01}, C_{02}, $\tilde{C}_{01}$, and γ) are determined by applying the global volume balance condition (2.13) and balancing terms according to the size of the parameters. This condition also checks the consistency of the expansion and shows when we need to consider additional time-dependencies which we discuss below.

The global volume balance equation in terms of the ξ scaling is given by (2.6). Since $l = \gamma L$ and $L = C_L t^{\lambda}$, where C_L is an undetermined constant, we use Γ_l defined in (2.18) to simplify the double integral. Hence (2.6) reduces to

$$\mathcal{G}_v = 2\gamma \int_0^1 \Omega \, \mathrm{d}\chi + \frac{2\lambda\sqrt{\pi}\gamma\Gamma(\lambda)\mathcal{G}_c}{(\lambda + 1/2)\Gamma(\lambda + 1/2)}, \tag{3.33}$$

where $\Gamma(\cdot)$ represents the Gamma function. We must analyze the different situations that arise for $\mathcal{P}_{ckm} \gg 1$ and $\mathcal{P}_{ckm} \ll 1$. The former case corresponds to toughness

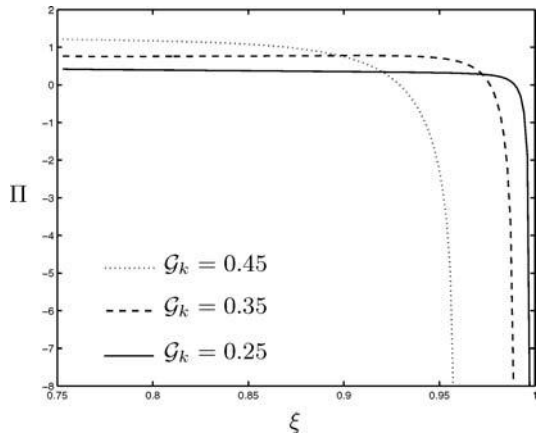


FIG. 3.3. *Solution profiles of* Π *against* ξ *with fixed* $\mathcal{G}_c = 1$, $\mathcal{G}_m = 0.2$, *and* $\mathcal{G}_k = 0.45, 0.35, 0.25$ *(and* $\mathcal{P}_{ckm} = 0.205, 0.075, 0.020$, *respectively).*

dominating the behavior over the whole fracture, with $\alpha_1 > 0$, and is discussed in detail in [44]. Then the left-hand side balances with the integral on the right-hand side to leading order; it follows that $\lambda = 2/3$, that is, $L = C_L t^{2/3}$. Substitution of (3.25) then leads to an expression for γ in terms of $\mathcal{G}_v$ and $\mathcal{G}_k$. The integral is evaluated using the asymptotic expansion near the tip and using numerical evaluation away from the tip where the behavior is regular [6, 24].

In contrast, for the case of $\mathcal{P}_{ckm} \ll 1$, the expansion (3.26) with $\alpha_1 < 0$ must be used for $\mathcal{G}_k^4/(\mathcal{G}_c\mathcal{G}_m) < (1-\xi)^{1/2}$. For this range of ξ the leading order terms in the lubrication equation are those with coefficients $\mathcal{G}_c$ and $\mathcal{G}_m^{-1}$. Together with the elasticity equation, these terms indicate that Ω and Π must scale with $(\mathcal{G}_c\mathcal{G}_m)^{1/4}$, as in (3.26). Then the global balance condition has the form

$$\mathcal{G}_v/\mathcal{G}_c = \text{const} \cdot \mathcal{P}_{cm}^{-1/4} + \text{const}. \tag{3.34}$$

As discussed in section 3.1 following (3.11), $\mathcal{P}_{cm} \gg 1$ in this case, and so the leading order terms are the first and third, and the contribution from the integral is higher order. Then we equate the leading order terms to obtain γ analytically. The global balance equation verifies that the self-similar solution is appropriate for sufficiently large $\mathcal{P}_{cm}$. Writing the t dependence explicitly in (3.34) and using the definitions in (2.2) with $L = C_L t^\lambda$ gives

$$\frac{Q_0}{\epsilon C_L^2} t^{1-2\lambda} = \text{const} \cdot \left(\frac{C'\mu'}{\epsilon^4 C_L E'}\right)^{1/4} t^{-(2\lambda+1)/8} + \text{const} \cdot \frac{C'}{\epsilon C_L} t^{1/2-\lambda}. \tag{3.35}$$

Comparing exponents and balancing the first and third terms gives $\lambda = 1/2$ and the expression for γ simplifies to $\gamma = \mathcal{G}_v/\pi\mathcal{G}_c$. Balancing the first and second terms leads to a contradiction unless $t < 1$. The second term in (3.35) can be neglected for $t^{-1/4} \ll 1$, corresponding to $\mathcal{P}_{cm}^{-1/4} \ll 1$ as in (3.34). This verifies that it is appropriate to use $\lambda = 1/2$ for the leak-off–dominated intermediate-tip behavior in section 3.1 for sufficiently large t. If this condition is violated, for example, for short times, we can no longer conclude that γ is constant, and additional time-dependence must be included in the expansion.

4. Discussion and future work. In conclusion, we have introduced a new approach for studying the system of integrodifferential equations that are found in hydraulic fracturing problems. Our method enables us to simultaneously consider the three primary physical mechanisms, namely, viscosity, toughness, and leak-off, and we have obtained a continuous solution for the fracture opening w when one or more of these processes are in balance. The technique determines critical relationships between the nondimensional distance from the tip $1-\xi$ and the key nondimensional quantities $\mathcal{G}_k$, $\mathcal{G}_c$, $\mathcal{G}_m$, and $\mathcal{G}_v$, representing toughness, leak-off, viscosity, and injected fluid volume, respectively.

For small toughness and $O(1)$ leak-off, the behavior of Ω follows from combining (3.25) for values of ξ in the near-tip region, and (3.26) for ξ in the intermediate-tip region. The critical parameter combination in this case is $\mathcal{P}_{ckm} = \mathcal{G}_k^4/\mathcal{G}_c\mathcal{G}_m$, with the transition layer occurring for values of $\mathcal{P}_{ckm} = O((1-\xi)^{1/2})$. Additional higher order corrections depend on the relative magnitude of extra parameter combinations, $\mathcal{P}_{km}$ and $\mathcal{G}_c^3/\mathcal{G}_m = \mathcal{P}_{cm}$. These results are obtained by simultaneously solving the elasticity and lubrication equations (2.10)–(2.11). The physical process of leak-off has often been ignored in previous studies, and the new terms in expansions (3.25) and (3.26) allow us to match expansions in different regions analytically, in contrast to previous work [24] for zero leak-off. This analysis is new and the flexibility is invaluable in extending the technique to other fracture geometries, namely, the PKN fracture [43], and adding different effects, such as stress-jumps and fluid lag. It is important to determine analytic solutions such as those derived in this paper, which can be used to test the validity of the model against more complete numerical models as well as data from laboratory and field experiments.

The application of the global volume balance equation is used to determine the remaining constants in the solution, and it also provides valuable information related to time dependence. In the regimes considered in this paper, we find that it provides a consistency check for the use of a self-similar solution (2.14)–(2.15) with the coefficient γ constant to leading order. We also determine the time-dependence of L, which gives the power law scaling in time of the length of the fracture.

The global volume balance equation can also be used to determine regimes where additional time-dependence must be included in the solution. For example, using this equation, for finite leak-off we can deduce that for large but finite time the coefficient γ varies on a slow time scale, $T = \mathcal{P}_{cm}^{-1/4}t$ for $\mathcal{P}_{cm} \gg 1$. For small time this analysis is not sufficient: then a multiple-scale analysis of the resulting equations is necessary to describe additional time dependence and transient behavior.

As noted in section 1.1 of the introduction, the results presented here rely on the use of Carter's leak-off model (1.8). However, our framework does not depend on a specific form for the leak-off term, so its implementation is not restricted to the use of (1.8). These results are valid for the balance of physical processes corresponding to $O(1)$ (or smaller) leak-off, which is typical of many fracturing treatments due to the cake-building properties of the fracturing fluid [51]. Specifically, we require that the combined parameter $\mathcal{P}_{cm} := \mathcal{G}_c^3/\mathcal{G}_m \gg 1$. In situations where $\mathcal{G}_c \gg 1$, such as in waterflood fractures, the leak-off term is not a higher order correction at the tip, and therefore it would require a different modeling approach in that region.

Typical values of Carter's leak-off coefficient C' range from 4–64×10^{-5} m/s$^{-1/2}$, as discussed in [6] and the references therein. Suppose we consider the case $\mathcal{P}_{ckm} \ll 1$ and set $\mathcal{G}_v = \mathcal{G}_c = 1$. Then the expression for γ, as found in section 3.3, is simply $\gamma = 1/\pi$. Typical values of the other parameters are found in [6, 19, 27]: $Q_0 = 4$–40×10^{-4} m^2/s, $E = 10000$–25000 MPa, $\nu = 0.15$, $\mu = 1 \times 10^{-7}$ MPa$\cdot$s, and

$K_{Ic} = 1$ MPa $\cdot$ m$^{1/2}$. For example, using $Q_0 = 4 \times 10^{-4}$ m^2/s, $C' = 16 \times 10^{-5}$, and $E = 10000$ MPa, the combined parameters can be shown to satisfy $\mathcal{P}_{cm} \approx 0.0022t$, $\mathcal{P}_{ckm} \approx 0.20$, and $\mathcal{P}_{km} \approx 0.066t^{1/4}$. Then the transition region is at $\xi = O(1-\mathcal{P}_{ckm}^2) \approx 0.96$, which supports the values used in the figures in section 3. This also shows that our analysis is applicable for large time, since we require $\mathcal{P}_{cm} \gg 1$. If we increase E and Q_0, say to 14000 MPa and 7×10^{-4} m^2/s, respectively, then $\mathcal{P}_{cm} \approx 0.00058t$, $\mathcal{P}_{ckm} \approx 0.042$, $\mathcal{P}_{km} \approx 0.014t^{1/4}$, and the transition region is at $\xi \approx 0.998$, which is now much closer to the tip.

For the situation when the rock is impermeable, that is, for zero leak-off with $\mathcal{G}_c = 0$, we have used the same procedure to obtain an analytically matched asymptotic solution in the tip region [44]. The expansions for Ω in terms of the key parameter $\mathcal{P}_{km} := \mathcal{G}_k^3/\mathcal{G}_m$ are, for $\mathcal{P}_{km} \gg (\ll) (1-\xi)^{1/2}$, respectively,

$$\Omega \sim \mathcal{G}_k \left[C_{00}\sqrt{1-\xi^2} + C_{01}\mathcal{P}_{km}^{-1}(1-\xi) \right], \tag{4.1}$$

$$\Omega \sim \mathcal{G}_m^{1/3} \left[\bar{C}_{01}(1-\xi)^{2/3} + \mathcal{A}_1(\mathcal{P}_{km})(1-\xi)^h + \mathcal{A}_2(\mathcal{P}_{km}) \right]. \tag{4.2}$$

Hence the expansion (4.1) holds for values of ξ in the near-tip region, the expansion (4.2) holds for ξ in the intermediate-tip region, and the analysis yields

$$C_{00} = \hat{\mathcal{G}}_k \sqrt{\frac{\gamma}{2}}, \qquad C_{01} = 2\pi\lambda \frac{\hat{\mathcal{G}}_m \gamma^3}{C_{00}^2}, \qquad \bar{C}_{01} = \left\{ \frac{4\hat{\mathcal{G}}_m \lambda \gamma^3}{m(m-1)\cot \pi m} \right\}^{1/3}, \tag{4.3}$$

where $m = 2/3$. The motivation for defining the parameter $\mathcal{P}_{km}$ again follows directly from the dominant behavior expressed by balancing the exponents of δ, which becomes $\mathcal{P}_{km} \gg (\ll) (1-\xi)^{1/2}$. The expansion (4.1) is physically significant in the toughness dominated regime or valid close to the tip when $\mathcal{G}_k^3/\mathcal{G}_m \gg (1-\xi)^{1/2}$, and the expansion (4.2) corresponds to the viscosity dominated regime in which $\mathcal{G}_k^3/\mathcal{G}_m \ll (1-\xi)^{1/2}$. Previous work [24] give some of the terms in (4.1)–(4.2), but the matching is done numerically. The additional terms allow us to give an analytical expression for the matching of (4.1) and (4.2), obtained by solving for the remaining unknown coefficients $\mathcal{A}_1(\mathcal{P}_{km})$ and $\mathcal{A}_2(\mathcal{P}_{km})$, as shown in [44].

In other asymptotic limits, for example, for $\mathcal{P}_{cm} \ll (1-\xi)^{1/2}$ and $\mathcal{G}_c > \mathcal{G}_k$, one can obtain solutions which involve more than one transition region, as discussed at the start of section 3. In particular, for sufficiently large leak-off and viscosity, the solution for Ω consists of a leading order behavior with power law $(1-\xi)^{1/2}$ for ξ in the near-tip region, $(1-\xi)^{5/8}$ for an intermediate-tip region, and $(1-\xi)^{2/3}$ for values of ξ farther from the tip. In the case $\mathcal{G}_c > \mathcal{G}_k$, a transition must occur at $1-\xi = O(\mathcal{P}_{ckm}^2)$ between the near-tip square root behavior and the $(1-\xi)^{5/8}$ behavior. If, in addition, $\mathcal{P}_{cm}$ is such that $\mathcal{P}_{cm} \ll (1-\xi)^{1/2}$, then there is another transition farther from the tip at $1-\xi = O(\mathcal{P}_{cm}^2)$ between the $(1-\xi)^{5/8}$ and $(1-\xi)^{2/3}$ behavior. The construction of the nondimensionalized width Ω proceeds as in the previous sections by identifying the appropriate balance of $(1-\xi)$ with combinations of the parameters $\mathcal{G}_c$, $\mathcal{G}_k$, and $\mathcal{G}_m$. Finally, in the same way as the analysis described here in section 3.2 and in [44], we can match across the two regions to determine the unknown coefficients which are now in terms of both $\mathcal{P}_{cm}$ and $\mathcal{P}_{ckm}$.

Appendix A. Results on Π in the toughness-dominated regime.

A.1. Π_{00} is constant. Consider (3.1). We integrate to obtain

$$\Omega_{00}^3 \frac{\mathrm{d}\Pi_{00}}{\mathrm{d}z} = k_1 \tag{A.1}$$

for some constant $k_1 \neq 0$. Using the propagation condition (2.12) for Ω_{00} in (A.1) and integrating with respect to z yields

$$\Pi_{00} = -\frac{2k_1}{\hat{\mathcal{G}}_k^3 \gamma^{3/2}} z^{-1/2} + \text{ const} \tag{A.2}$$

for $z \ll 1$. Now we compare this with the result from the elasticity equation (3.2), again using the propagation condition (2.12) for Ω_{00}, to get

$$\delta^{\beta_k}\Pi_{00} = \text{const} * \int_0^{1/\delta} \frac{\mathrm{d}\Omega_{00}}{\mathrm{d}r} \frac{(1-\delta r)}{r(2-\delta r) - z(2-\delta z)}\,\mathrm{d}r = \text{ const}$$

to leading order. This contradicts (A.2) so that $k_1 = 0$. Thus, from (A.1) it follows that Π_{00} is constant.

A.2. Integration constant from the lubrication equation is zero. Integrating (3.4) with respect to z, using the leading order behavior $(2z)^{-1/2}$ for the leak-off term, gives

$$\lambda\delta^{\beta_k - 1/2}\Omega_{00} + \hat{\mathcal{G}}_c \delta^{\beta_c - 1/2}\sqrt{2} z^{1/2} + k_2 = \frac{1}{\hat{\mathcal{G}}_m \gamma^2}\Omega_{00}^3 \frac{\mathrm{d}\Pi_{01}}{\mathrm{d}z} \tag{A.3}$$

for some constant $k_2 \neq 0$. Using the propagation condition (2.12) for Ω_{00} in (A.3) and integrating with respect to z yields

$$\delta^{\sigma_1}\Pi_{01} = \frac{\hat{\mathcal{G}}_m}{\hat{\mathcal{G}}_k^3}\delta^{\beta_m - 3\beta_k}\left[\lambda\gamma\hat{\mathcal{G}}_k\delta^{\beta_k}\ln z + \sqrt{2\gamma}\hat{\mathcal{G}}_c\delta^{\beta_c}\ln z - 2\sqrt{\gamma}k_2\delta^{1/2}z^{-1/2}\right] + k_3. \tag{A.4}$$

Without loss of generality, we set the integrating constant $k_3 = 0$, as it can be incorporated into Π_{00}. We now consider the stress intensity factor K_I given in (1.2). This can be rewritten in the ξ scaling as

$$\mathcal{G}_k = \frac{8\sqrt{2}}{\pi}\gamma^{1/2}\int_0^1 \frac{\Pi}{\sqrt{1-\xi^2}}\,\mathrm{d}\xi. \tag{A.5}$$

For $\xi = 1 - \delta z$, we consider the contribution to (A.5) obtained from the term in (A.4) with coefficient k_2, namely,

$$\text{const} * \frac{8\sqrt{2}}{\pi}\gamma^{1/2}\delta^{1/2}\int_0^{1/\delta} \frac{k_2}{z\sqrt{2-\delta z}}\,\mathrm{d}z. \tag{A.6}$$

For $k_2 \neq 0$, this term is infinite at $z = 0$. So $k_2 = 0$ to maintain a finite energy.

Appendix B. Calculation of the expression for Π. We summarize the calculation of the asymptotic behavior of Π from (2.9) to find the leading order behavior. We introduce a parameter ξ^* which is in the transition region $1 - \xi = O(\mathcal{P}_{ckm}^2)$. Then the expansion (3.25) holds near the tip and (3.26) holds away from the tip, i.e.,

$$\Omega \sim \mathcal{G}_k\left[C_{00}\sqrt{1-\xi^2} + C_{01}\mathcal{P}_{km}^{-1}(1-\xi) + C_{02}\mathcal{P}_{ckm}^{-1}(1-\xi)\right] \tag{B.1}$$

for $\xi^* < \xi < 1$, and

$$\begin{aligned}\Omega \sim (\mathcal{G}_c\mathcal{G}_m)^{1/4}\Big[&\tilde{C}_{01}(1-\xi)^{5/8} + \mathcal{B}_1(\mathcal{P}_{ckm})(1-\xi)^{1/8} \\ &+ \mathcal{P}_{cm}^{-1/4}\mathcal{B}_2(1-\xi)^{3/4} + \mathcal{B}_3(\mathcal{P}_{ckm})(1-\xi)^r\Big]\end{aligned} \tag{B.2}$$

for $\xi < \xi^*$, and (2.9) becomes

(B.3)

$$\begin{aligned}\Pi =& -\frac{\mathcal{G}_k C_{00}}{4\pi\gamma}\int_{\xi^*}^1 \frac{\left[\sqrt{1-\chi^2}\right]' 2\chi \mathrm{d}\chi}{\chi^2-\xi^2} - \frac{\mathcal{G}_k\left[C_{01}\mathcal{P}_{km}^{-1}+C_{02}\mathcal{P}_{ckm}^{-1}\right]}{4\pi\gamma}\int_{\xi^*}^1 \frac{\left[(1-\chi)\right]' 2\chi \mathrm{d}\chi}{\chi^2-\xi^2}\\ &-(\mathcal{G}_c\mathcal{G}_m)^{1/4}\left\{\frac{\tilde{C}_{01}}{4\pi\gamma}\int_0^{\xi^*}\frac{\left[(1-\chi)^{5/8}\right]' 2\chi \mathrm{d}\chi}{\chi^2-\xi^2} + \frac{\mathcal{B}_1(\mathcal{P}_{ckm})}{4\pi\gamma}\int_0^{\xi^*}\frac{\left[(1-\chi)^{1/8}\right]' 2\chi \mathrm{d}\chi}{\chi^2-\xi^2}\right\}\\ &-(\mathcal{G}_c\mathcal{G}_m)^{1/4}\left\{\frac{\mathcal{P}_{cm}^{-1/4}\mathcal{B}_2}{4\pi\gamma}\int_0^{\xi^*}\frac{\left[(1-\chi)^{3/4}\right]' 2\chi \mathrm{d}\chi}{\chi^2-\xi^2} + \frac{\mathcal{B}_3(\mathcal{P}_{ckm})}{4\pi\gamma}\int_0^{\xi^*}\frac{\left[(1-\chi)^{r}\right]' 2\chi \mathrm{d}\chi}{\chi^2-\xi^2}\right\}\\ =:& \ I_1+I_2+I_3+I_4+I_5+I_6.\end{aligned}$$

In the intermediate region $1-\xi = O(\mathcal{P}_{ckm}^s)$ for $0 < s < 2$, the integrals I_1 and I_2 can be evaluated to give

$$\begin{aligned}I_1 =& \frac{\mathcal{G}_k C_{00}}{4\pi\gamma}(1+\xi)^{-1}\sqrt{2}(1-\xi^*)^{1/2}\left\{\left(1+O\big(1-\xi^*\big)\right)+O\left(\frac{1-\xi^*}{1+\xi}\right)\right\}\\ &+\frac{\mathcal{G}_k C_{00}}{4\pi\gamma}(1-\xi)^{-1}\sqrt{2}(1-\xi^*)^{1/2}\left\{\left(1+O\big(1-\xi^*\big)\right)+O\left(\frac{1-\xi^*}{1-\xi}\right)\right\},\\ I_2 =& -\frac{\mathcal{G}_k\left[C_{01}\mathcal{P}_{km}^{-1}+C_{02}\mathcal{P}_{ckm}^{-1}\right]}{4\pi\gamma}\left\{\ln\left|1-\frac{1-\xi^*}{1+\xi}\right|+\ln\left|1-\frac{1-\xi^*}{1-\xi}\right|\right\},\end{aligned} \tag{B.4}$$

while in the near-tip region $1-\xi = O(\mathcal{P}_{ckm}^s)$ for $s > 2$ the integrals are of the form

(B.5)

$$\begin{aligned}I_1 =& \frac{\mathcal{G}_k C_{00}}{4\pi\gamma}(1+\xi)^{-1}\sqrt{2}(1-\xi^*)^{1/2}\left\{\left(1+O\big(1-\xi^*\big)\right)+O\left(\frac{1-\xi^*}{1+\xi}\right)\right\}\\ &+\frac{\mathcal{G}_k C_{00}}{4\pi\gamma}\left\{\pi-\arctan\left(\frac{\xi^*}{\sqrt{1-\xi^{*2}}}\right)+\frac{\xi}{\sqrt{1-\xi^2}}\ln\left|\frac{1-\xi^*\xi+\sqrt{1-\xi^2}\sqrt{1-\xi^{*2}}}{\xi-\xi^*}\right|\right\},\\ I_2 =& -\frac{\mathcal{G}_k\left[C_{01}\mathcal{P}_{km}^{-1}+C_{02}\mathcal{P}_{ckm}^{-1}\right]}{4\pi\gamma}\left\{\ln\left|1-\frac{1-\xi^*}{1+\xi}\right|+\ln\left|\frac{1-\xi^*}{1-\xi}\right|+\ln\left|1-\frac{1-\xi}{1-\xi^*}\right|\right\}.\end{aligned}$$

We briefly outline the calculation of I_3, I_4, I_5, and I_6 for a general integral of that form with parameter $0 < a < 1$: then the results follow from setting $a = 5/8$, $1/8$, $3/4$, r, respectively. Thus the integral is

$$J_3 = \int_0^{\xi^*}(1-\chi)^{a-1}\frac{2\chi \mathrm{d}\chi}{\chi^2-\xi^2}, \tag{B.6}$$

which is now split as

$$J_3 = \int_0^{\xi^*}\frac{(1-\chi)^{a-1}}{\chi-\xi}\,\mathrm{d}\chi - \int_{-\xi^*}^0\frac{(1+\chi')^{a-1}}{\chi'-\xi}\,\mathrm{d}\chi' =: J_3^A + J_3^B. \tag{B.7}$$

Note that ξ can vary over the whole interval, i.e., $-1 < \xi < 1$. Then asymptotic expansions for the integrals are used, depending on whether ξ is inside or outside the interval of integration. It is convenient to use a change of variables which captures the asymptotic behavior of Π near the tip. It is also convenient to use different variables on different intervals, such as $\delta Z = 1-\xi$, $\delta R = 1-\chi$ in J_3^A and, $\delta Z = 1+\xi$,

$\delta R = 1 + \chi$ in J_3^B. We describe the procedure for the integral J_3^A (then J_3^B follows from an analogous calculation). This is split into three parts as

$$J_3^A = -\delta^{a-1}\left(\int_0^\infty - \int_{1/\delta}^\infty - \int_0^{(1-\xi^*)/\delta}\right)\frac{R^{a-1}}{R-Z}\,\mathrm{d}R. \tag{B.8}$$

In the intermediate region $1-\xi = O(\mathcal{P}_{ckm}^s)$ for $0 < s < 2$, the leading order behavior is determined by the first integral for $\delta = \mathcal{P}_{ckm}^2 \ll 1$ (as given in [40]), and so

$$J_3^A = (\delta Z)^{a-1}\pi\cot\pi a + O(1).$$

Then, for intermediate values of $\xi < \xi^*$, the integrals I_1, I_2, and J_3^B all give $O(1)$ contributions, which are lower order compared to the leading order term in J_3^A. The integral for J_4 is calculated in the same way. Hence the expression for Π in (B.3) is

$$\begin{aligned}\Pi = (\mathcal{G}_c\mathcal{G}_m)^{1/4}&\left\{\frac{q\tilde{C}_{01}}{4\pi\gamma}(1-\xi)^{q-1}\pi\cot\pi q + \frac{g\mathcal{B}_1(\mathcal{P}_{ckm})}{4\pi\gamma}(1-\xi)^{g-1}\pi\cot\pi g\right\}\\ &+(\mathcal{G}_c\mathcal{G}_m)^{1/4}\left\{\frac{p\mathcal{P}_{cm}^{-1/4}\mathcal{B}_2}{4\pi\gamma}(1-\xi)^{p-1}\pi\cot\pi p\right.\\ &\left.+\frac{r\mathcal{B}_3(\mathcal{P}_{ckm})}{4\pi\gamma}(1-\xi)^{r-1}\pi\cot\pi r\right\} + O(1),\end{aligned}$$

where $q = 5/8$, $g = 1/8$, $p = 3/4$ and the $O(1)$ terms are higher order with respect to $1-\xi \ll 1$.

Similarly, for values of ξ in the near-tip region $1-\xi = O(\mathcal{P}_{ckm}^s)$ for $s > 2$, the integral I_3 gives $O(1)$ contributions, and the leading order term is the singularity in I_2 defined in (B.5). Hence the expression for Π is now

$$\begin{aligned}\Pi = \mathcal{G}_k&\left[\frac{C_{00}}{4\gamma} - \frac{C_{01}\mathcal{P}_{km}^{-1} + C_{02}\mathcal{P}_{ckm}^{-1}}{4\pi\gamma}\left\{\ln\left|1-\frac{1-\xi^*}{1+\xi}\right| + \ln\left|\frac{1-\xi^*}{1-\xi}\right| + \ln\left|1-\frac{1-\xi}{1-\xi^*}\right|\right\}\right]\\ &+O(1).\end{aligned}$$

Here we have explicitly included the leading order term from I_1 for comparison with (3.31). Additional error terms not shown here also result from the fact that higher order derivatives for $\Omega(\xi)$ are not matched in the transition region: these can be shown to be higher order for $\mathcal{P}_{ckm} \ll 1$.

Acknowledgments. We would like to thank Emmanuel Detournay, José Adachi, and Dmitry Garagash for their very helpful comments and for sharing a number of preprints with us.

REFERENCES

[1] H. Abé, T. Mura, and L. M. Keer, *Growth rate of a penny-shaped crack in hydraulic fracturing of rocks*, J. Geophys. Res., 81 (1976), pp. 5335–5340.

[2] J. I. Adachi and E. Detournay, *Self-similar solution of a plane-strain fracture driven by a power-law fluid*, Int. J. Numer. Anal. Meth. Geomech., 26 (2002), pp. 579–604.

[3] J. I. Adachi and E. Detournay, *Plane-strain propagation of a fluid-driven fracture: Finite toughness self-similar solution*, in preparation, 2006.

[4] J. I. Adachi and E. Detournay, *Propagation of a fluid-driven fracture in a permeable medium*, J. Eng. Fracture Mech., submitted, 2006.

[5] J. I. Adachi, E. Siebrits, A. Peirce, and J. Desroches, *Computer simulation of hydraulic fractures*, Int. J. Rock Mech. and Min. Sci., submitted, 2006.
[6] J. I. Adachi, *Fluid-Driven Fracture in Permeable Rock*, Ph.D. Thesis, University of Minnesota, 2001; available at www.umi.com.
[7] S. H. Advani, T. S. Lee, and J. K. Lee, *Three-dimensional modeling of hydraulic fractures in layered media: Finite element formulations*, ASME J. Energy Res. Technol., 112 (1990), pp. 1–18.
[8] G. Barenblatt, *The mathematical theory of equilibrium cracks in brittle fracture*, Adv. Appl. Mech., 7 (1962), pp. 55–129.
[9] A. P. Bunger, E. Detournay, and D. I. Garagash, *Toughness-dominated regime with leak-off*, Int. J. Fracture, 134 (2005), pp. 175–190.
[10] R. S. Carbonell, J. Desroches, and E. Detournay, *A comparison between a semi-analytical and a numerical solution of a two-dimensional hydraulic fracture*, Internat. J. Solids Structures, 36 (1999), pp. 4869–4888.
[11] E. Carter, *Optimum fluid characteristics for fracture extension*, in Drilling & Production Practices, G. Howard and C. Fast, eds., American Petroleum Institute, Tulsa, OK, 1957, pp. 261–270.
[12] R. J. Clifton and A. S. Abou-Sayed, *A variational approach to the prediction of the three-dimensional geometry of hydraulic fractures*, in Proceedings of the SPE/DOE Low Permeability Reservoir Symposium, Denver, CO, 1981.
[13] B. Cotterell and J. R. Rice, *Slightly curved or kinked cracks*, Internat. J. Fracture, 16 (1980), pp. 155–169.
[14] B. C. Crittendon, *The mechanics of design and interpretation of hydraulic fracture treatments*, J. Pet. Tech. 11, October 1959, pp. 21–29.
[15] J. Desroches, E. Detournay, B. Lenoach, P. Papanastasiou, J. R. A. Pearson, M. Thiercelin, and A. H.-D. Cheng, *The crack tip region in hydraulic fracturing*, Proc. R. Soc. London Ser. A, 447 (1994), pp. 39–48.
[16] E. Detournay, J. I. Adachi, and D. I. Garagash, *Asymptotic and intermediate asymptotic behavior near the tip of a fluid-driven fracture propagating in a permeable elastic medium*, in Structural Integrity and Fracture, A. V. Dyskin, X. Hu, and E. Sahouryeh, eds., Swets & Zeitlinger, Lisse, Zuid-Holland, The Netherlands, 2002.
[17] E. Detournay and A. H.-D. Cheng, *Plane strain analysis of a stationary hydraulic fracture in a poroelastic medium*, Internat. J. Solids Structures, 27 (1991), pp. 1645–1662.
[18] E. Detournay and A. H.-D. Cheng, *Fundamentals of Poroelasticity*, in Comprehensive Rock Engineering, Principles, Practice & Projects, Vol. 2, J. A. Hudson, ed., Pergamon Press, Oxford, UK, 1993, Chap 5, pp. 113–169.
[19] E. Detournay and D. I. Garagash, *The near-tip region of a fluid-driven fracture propagating in a permeable elastic solid*, J. Fluid Mech., 494 (2003), pp. 1–32.
[20] E. Detournay and D. I. Garagash, *General scaling laws for fluid-driven fractures*, preprint, 2004.
[21] E. Detournay, *Propagation regimes of of fluid-driven fractures in impermeable rocks*, Int. J. Geomech., 4 (2004), pp. 1–11.
[22] A. V. Dyskin, L. N. Germanovich, and K. B. Ustinov, *Asymptotic analysis of crack interaction with free boundary*, Internat. J. Solids Structures, 37 (2000), pp. 857–886.
[23] D. I. Garagash, E. Detournay, and J. I. Adachi, *Tip solution of a fluid-driven fracture in a permeable rock*, in preparation, 2006.
[24] D. I. Garagash and E. Detournay, *Plane-strain propagation of a fluid-driven fracture: Small toughness solution*, J. Appl. Mech., 72 (2005), pp. 916–928.
[25] D. I. Garagash, *Hydraulic fracture propagation in elastic rock with large toughness*, in Proceedings of the 4th North American Rock Mechanics Symposium, J. Girard, M. Liebman, C. Breeds, and T. Doe, eds., 2000, pp. 221–228.
[26] D. I. Garagash, *Plane-strain propagation of a hydraulic fracture during injection and shut-in: Asymptotics of large toughness*, Engrg. Fracture Mech., 73 (2006), pp. 456–481.
[27] D. Garagash and E. Detournay, *The tip region of a fluid-driven fracture propagating in an elastic medium*, ASME J. Appl. Mech., 67 (2000), pp. 183–192.
[28] J. Geertsma and F. de Klerk, *A rapid method of predicting width and extent of hydraulically induced fractures*, J. Pet. Technol., 246 (1969), pp. 1571–1581.
[29] Y. N. Gordeyev and V. M. Entov, *The pressure distribution around a growing crack*, J. Appl. Math. Mech., 61 (1997), pp. 1025–1029.
[30] E. Harrison, W. F. Kieschnick, and W. J. McGuire, *The mechanics of fracture induction and extension*, Petroleum Trans. AIME, 201 (1954), pp. 252–263.
[31] G. C. Howard and C. R. Fast, *Optimum fluid characteristics for fracture extension*, Drilling and Production Practice, 24 (1957), pp. 261–270.

[32] M. K. Hubbert and D. G. Willis, *Mechanics of hydraulic fracturing*, Pet. Trans. (AIME), 210 (1957), pp. 153–166.
[33] S. Khristianovic and Y. Zheltov, *Formation of vertical fractures by means of highly viscous fluids*, in Proceedings of the 4th World Petroleum Congress, Rome, Italy, 1955, pp. 579–586.
[34] O. Kresse, E. Detournay, and D. I. Garagash, *Universal tip solution for a fracture driven by a power law fluid*, preprint (to be submitted to J. Non-Newtonian Fluid Mech.), 2005.
[35] O. Kresse and E. Detournay, *Tip solution for a fracture driven by a perfectly plastic fluid*, preprint (to be submitted to J. Elasticity), 2005.
[36] B. Lenoach, *The crack tip solution for hydraulic fracturing in a permeable solid*, J. Mech. Phys. Solids, 43 (1995), pp. 1025–1043.
[37] J. R. Lister, *Buoyancy-driven fluid fracture: similarity solutions for the horizontal and vertical propagation of fluid-filled cracks*, J. Fluid Mech., 217 (1990), pp. 213–239.
[38] J. R. Lister, *Buoyancy-driven fluid fracture: The effects of material toughness and of low-viscosity precursors*, J. Fluid Mech., 210 (1990), pp. 263–280.
[39] M. G. Mack and N. R. Warpinski, *Mechanics of hydraulic fracturing*, in Reservoir Stimulation, 3rd ed., M. Economides and K. Nolte, eds., John Wiley & Sons, New York, Chap. 6, 2000.
[40] P. A. Martin, *End-point behaviour of solutions to hypersingular integral equations*, Proc. Roy. Soc. London Ser. A, 432 (1991), pp. 301–320.
[41] P. A. Martin, *Perturbed cracks in two dimensions: An integral-equation approach*, Internat. J. Fracture, 104 (2000), pp. 317–327.
[42] P. A. Martin, *On wrinkled penny-shaped cracks*, J. Mech. Phys. Solids, 49 (2001), pp. 1481–1495.
[43] S. L. Mitchell, R. Kuske, A. P. Peirce, and J. I. Adachi, *An asymptotic analysis of a finger-like fluid-driven fracture*, in preparation, 2006.
[44] S. L. Mitchell, R. Kuske, and A. P. Peirce, *An asymptotic framework for the analysis of hydraulic fractures: The impermeable case*, J. Appl. Mech., in press (preprint available at www.iam.ubc.ca/∼sarah/), 2007.
[45] R. Nordgren, *Propagation of vertical hydraulic fractures*, SPE J., 12 (1972), pp. 306–314.
[46] A. P. Peirce and E. Siebrits, *The scaled flexibility matrix method for the efficient solution of boundary value problems in* $2d$ *and* $3d$ *layered elastic media*, Comput. Methods Appl. Mech. Engrg., 1990 (2001), pp. 5935–5956.
[47] A. P. Peirce and E. Siebrits, *An efficient multilayer planar* $3d$ *fracture growth algorithm using a fixed mesh approach*, Internat. J. Numer. Methods Engrg., 53 (2002), pp. 691–717.
[48] T. K. Perkins and L. R. Kern, *Widths of hydraulic fractures*, J. Pet. Tech., 222 (1961), pp. 937–949.
[49] J. R. Rice, *Mathematical analysis in the mechanics of fracture*, in Fracture, an Advanced Treatise, Vol. 2, Academic Press, New York, 1968, Chap. 3, pp. 191–311.
[50] A. A. Savitski and E. Detournay, *Propagation of a penny-shaped fluid-driven fracture in an impermeable rock: Asymptotic solutions*, Internat. J. Solids Structures, 39 (2002), pp. 6311–6337.
[51] A. Settari, *A new general model of fluid loss in hydraulic fracturing*, Soc. Pet. Engrg. J., 4 (1985), pp. 491–501.
[52] I. Sneddon and Lowengrub M, *Crack Problems in the Classical Theory of Elasticity*, John Wiley & Sons, New York, 1969.
[53] J. L. S. Sousa, B. J. Carter, and A. R. Ingraffea, *Numerical simulation of* $3d$ *hydraulic fracturing using Newtonian and power-law fluids*, Int. J. Rock Mech. Min. Sci., 30 (1993), pp. 1265–1271.
[54] D. A. Spence and P. Sharp, *Self-similar solutions for elastohydrodynamic cavity flow*, Proc. Roy. Soc. London Ser. A, 400 (1985), pp. 289–313.
[55] D. A. Spence and D. L. Turcotte, *Magma-driven propagation of cracks*, J. Geophys. Res., 90 (1985), pp. 575–580.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 387–407

STOCHASTIC DIFFERENTIAL DELAY EQUATION, MOMENT STABILITY, AND APPLICATION TO HEMATOPOIETIC STEM CELL REGULATION SYSTEM*

JINZHI LEI† AND MICHAEL C. MACKEY‡

Abstract. We study the moment stability of the trivial solution of a linear differential delay equation in the presence of additive and multiplicative white noise. The results established here are applied to examining the local stability of the hematopoietic stem cell (HSC) regulation system in the presence of noise. The stability of the first moment for the solutions of a linear differential delay equation under stochastic perturbation is identical to that of the unperturbed system. However, the stability of the second moment is altered by the perturbation. We obtain, using Laplace transform techniques, necessary and sufficient conditions for the second moment to be bounded. In applying the results to the HSC system, we find that the system stability is sensitive to perturbations in the stem cell differentiation and death rates, but insensitive to perturbations in the proliferation rate.

Key words. stochastic differential delay equation, moment stability, hematopoietic disease

AMS subject classifications. 34K50, 92C99

DOI. 10.1137/060650234

1. Introduction.

Delays in feedback regulation are ubiquitous in biological control systems, where the retardation usually originates from maturing processes or finite signaling velocities [4, 15, 16, 17, 21, 30, 34, 36, 37, 39, 45, 46]. Differential delay equation model systems with retarded arguments have been extensively developed in the past several decades (see [3, 10, 11, 18, 19, 20] and the references therein). However, in applied areas, deterministic systems fail to capture the essence of the fluctuations in the real situation, and one must instead consider models with stochastic processes that take into account the perturbations present in the real world. In situations where delays are important, models with stochastic perturbations are framed by stochastic differential delay equations.

The current study is motivated by an investigation of the stability of the hematopoietic regulatory system and its connection with several hematological diseases [5, 7, 8, 9, 21, 30, 39]. All blood cells originate from the hematopoietic stem cells (HSC) in the bone marrow. These stem cells differentiate and proliferate, giving rise to the three major cell lines: the leukocytes (white blood cells), the platelets, and the erythrocytes (red blood cells). The three peripheral regulatory loops are all of a negative feedback nature, and are mediated by a variety of cytokines including erythropoietin (EPO), thrombopoietin (TPO), and granulocyte colony-stimulating factor (G-CSF) [1, 50, 53, 55, 58]. These cytokines are synthesized and released by cells of the hematopoietic system. They control the hematopoietic system by regulating the growth, differentiation, and survival of cells.

*Received by the editors January 18, 2006; accepted for publication (in revised form) September 18, 2006; published electronically January 12, 2007. This work was supported by MITACS (Canada) and the Natural Sciences and Engineering Research Council of Canada (NSERC grant OGP-0036920).

http://www.siam.org/journals/siap/67-2/65023.html

†Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing, P. R. China, 100084 (jin_zhi_lei@yahoo.com; jzlei@mail.tsinghua.edu.cn). This author's work was supported by MCME (P. R. China) and MITACS (Canada).

‡Departments of Physiology, Physics, and Mathematics, and Centre for Nonlinear Dynamics, McGill University, 3655 Drummond, Montréal, QC, Canada H3G 1Y6 (mackey@cnd.mcgill.ca).

A mathematical model of the hematopoietic regulation system that combines the delay for cell maturation and negative feedback of the differentiated cells has been studied in [7, 8]. The numbers of circulating cells in a healthy person usually fluctuate with small amplitude around their normal levels. However, there are several hematological diseases that display a highly dynamic nature characterized by statistically significant oscillations in one or more of the circulating progeny of the HSC [21]. These diseases include, but are not limited to, cyclical neutropenia [8, 21, 22, 23], periodic chronic myelogenous leukemia [7, 12, 52], cyclical thrombocytopenia [49, 54, 59], and periodic hemolytic anemia [33, 45]. For example, cyclical neutropenia is a rare genetic blood disease in which the patient's neutrophil level drops to an extremely low level for six to eight days every three weeks. Neutrophils are a type of white blood cell important in the defense of the body against infection. Since stem cell oscillations are thought to drive oscillations in several periodic hematological diseases [21], understanding the HSC dynamics is important.

The differential delay equations that model the HSC dynamics have been developed for a G_0 cell cycle model in [13, 31, 32, 33, 39, 51]. The delays in these models reflect the nonzero time that it takes the cells to complete the proliferative phase of the cell cycle. For example, the HSC takes about 2.8 days to complete one cell cycle. Previous studies suggested that the HSC population becomes unstable and develops oscillations when the steady state corresponding to the healthy state is destabilized, for example by increasing the apoptosis (death) rate or the differentiation rate in the stem cells. However, in these studies, the stochastic perturbations that occur in the real world, and which might lead to instability and oscillation, were not taken into account. In this paper, we will investigate the effects of random perturbation and answer the following two questions:

1. If the steady state of the system without noise is unstable, is it possible to stabilize the steady state by noise perturbation?

2. If the steady state of the system without noise is stable, is it possible to destabilize the steady state by noise perturbation? If the answer is "YES," such perturbation usually originates from the perturbation in the system parameters. Therefore, there are thresholds for each of the parameters such that the steady state is stable when the perturbation is smaller than this threshold and unstable otherwise. The quantitative estimation of these thresholds will also be considered in this paper.

The answers to these two questions offer insight into the stability of the hematopoietic system in the face of stochastic perturbations.

The HSC dynamics with stochastic perturbation is modeled by a nonlinear stochastic differential delay equation. To answer the questions posed above, we linearize the equation around the steady state and study the stability of the resulting equation.

Consider the process described by the differential delay equation

$$\frac{dz}{dt} = f(z, z_\tau), \tag{1.1}$$

where $z_\tau = z(t-\tau)$. It may be the case that the function $f(z, z_\tau)$ is subject to some random effect (noise), so that we have

$$\frac{dz}{dt} = f(z, z_\tau) + \sigma(t, z, z_\tau) \cdot \xi_t(\omega), \tag{1.2}$$

where $\xi_t(\omega)$ is a stochastic process that represents the noise term. In our study of the hematopoietic system, this noise is internal to the system because of random

fluctuations in the system parameters, e.g., fluctuations in the differentiation rate, death rate, or proliferation rate of stem cells. However, the precise properties of the noise are not known. To gain insight into the effect of noise on the system, we assume the noise to be Gaussian distributed white noise with zero mean and a delta function autocorrelation $\langle \xi_t \xi_s \rangle = \delta(t-s)$. We assume further that the function σ does not depend on t explicitly. Using the definition of Gaussian white noise ξ_t as the derivative of the Wiener process $W(t)$, equation (1.2) can be written as

$$dz = f(z, z_\tau)dt + \sigma(z, z_\tau)dW(t). \tag{1.3}$$

From a formal point of view, we can solve (1.3) and write the stochastic process $z(t) = z(t;\omega)$ as

$$z(t) = z(0) + \int_0^t f(z(s), z_\tau(s))ds + \text{“}\int_0^t \sigma(z(s), z_\tau(s))dW(s).\text{”} \tag{1.4}$$

There are two interpretations for the stochastic integral

$$\text{“}\int_0^t \sigma(z(s), z_\tau(s))dW(s),\text{”}$$

the Itô interpretation and the Stratonovich interpretation. The Itô interpretation is usually used when the noise is white, but when the noise is colored (i.e., does not have a delta function autocorrelation), the Stratonovich interpretation is preferable. This issue has been discussed by many people (see, for example, [26, pp. 232–237], [28, pp. 346–351], [29, pp. 152–155], and [48, pp. 35–37]), and it is safe to say that the debate over the issue is far from settled. In this study we adopt the Itô interpretation for two reasons. First, the Itô approach is mathematically preferable [29], and second it is relatively straightforward to pass from results obtained using the Itô interpretation to one appropriate for the Stratonovich interpretation.

Indeed, assuming that the stochastic integral is to be interpreted as an Itô integral, (1.4) can be written as

$$z(t) = z(0) + \int_0^t f(z(s), z_\tau(s))ds + \int_0^t \sigma(z(s), z_\tau(s))dW(s). \tag{1.5}$$

There is a simple relation between the Itô interpretation and the Stratonovich interpretation [14, 48, 57]. Thus, the solution of (1.3) using the Stratonovich interpretation of the stochastic integral

$$z(t) = z(0) + \int_0^t f(z(s), z_\tau(s))ds + \int_0^t \sigma(z(s), z_\tau(s)) \circ dW(s)$$

is equivalent to the solution of the modified Itô equation

$$\begin{aligned} z(t) = z(0) &+ \int_0^t f(z(s), z_\tau(s))ds + \frac{1}{2}\int_0^t \sigma'_z(z(s), z_\tau(s))\sigma(z(s), z_\tau(s))ds \\ &+ \int_0^t \sigma(z(s), z_\tau(s))dW(s). \end{aligned}$$

Thus, the results in this paper obtained from the Itô approach are also applicable to a Stratonovich interpretation after replacing $f(z, z_\tau)$ in (1.3) by

$$f(z, z_\tau) + \frac{1}{2}\sigma'_z(z, z_\tau)\sigma(z, z_\tau). \tag{1.6}$$

We will see below that these two different interpretations can lead to significant changes in the predicted stability of the system.

Assume that $z = z_*$ is a steady state of (1.1); i.e., $f(z_*, z_*) = 0$. What we are interested to know is the effect of the noise perturbation on the steady state. In general, we do not have $\sigma(z_*, z_*) = 0$. Hence, $z(t) \equiv z_*$ is not a solution of the perturbed equation (1.3). We will address the question of under what condition the stochastic process $z(t)$ satisfying the perturbed equation (1.3) remains close to the steady state $z = z_*$, i.e., when the solution $z = z_*$ is "stable" under stochastic perturbation.

Linearizing (1.3) around the steady state yields the linear stochastic differential delay equation

$$dx = (ax + bx_\tau)dt + (\sigma_0 x + \sigma_1 x_\tau + \sigma_2)dW(t), \tag{1.7}$$

where $x(t) = z(t) - z_*$ and a, b, σ_i are constants given by

$$a = f'_z(z_*, z_*), \quad b = f'_{z_\tau}(z_*, z_*),$$

$$\sigma_0 = \sigma'_z(z_*, z_*), \quad \sigma_0 = \sigma'_{z_\tau}(z_*, z_*), \quad \sigma_2 = \sigma(z_*, z_*).$$

At this point, we will study the moment stability of (1.7) to answer the following questions:

1. Under what conditions does the ensemble mean of the solutions of (1.7) approach 0 when $t \to \infty$?

2. Under what condition is the variance of the solutions bounded (or unbounded) for all $t > 0$?

3. When the variance is bounded, then the upper limit of the variance, when $t \to \infty$, provides the estimation of its upper bound when t is large. Therefore, the estimation of the variance when $t \to \infty$ is interesting and will be studied in this paper.

Despite the apparently simple form of (1.7), the stability problem is not trivial, because of the combination of delay and stochastic terms.

Stochastic differential delay equations were introduced by Itô and Nisio in the 1960s [24]. Those authors also discussed the existence and uniqueness of the solution. However, progress in this area has been slow, and most of the results including stochastic stability, numerical approximation, etc., have been developed in the last decade [2, 27, 40, 41, 42, 43, 44, 47]; see [25] for a recent survey of these results. Despite the efforts of many researchers, this field is still in its infancy. For example, conditions for the stability of (1.7), a linear stochastic differential delay equation with constant coefficients, are not known. In the case of a stochastic ordinary differential equation ($b = \sigma_1 = 0$) and a delay differential equation ($\sigma_i = 0$), the stability conditions of the equation have been well established [20, 41]. However, when trying to extend these results to stochastic differential delay equations, one encounters serious difficulties because of the combination of delay and stochastic processes, and the explicit solution of (1.7) is not known.

The Lyapunov function method is useful for studying the stability of differential equations and has been developed for both differential delay equations and stochastic differential equations. In the 1990s, Mao extended this method to stochastic functional differential equations [41, Chapter 5]. Because of the results of Mao, we have some results for the stability of stochastic differential delay equations (see [41, section 5.6] for details). However, when applying these results to (1.7), we find that they are applicable only when $a < 0$. In our study of the hematopoietic system, the case $a > 0$

is the most interesting, and we therefore need to develop new results for the moment stability of (1.7).

In this paper, we will first develop the mathematical theory for the moment stability of the linear stochastic differential delay equation (1.7), and then apply the result to studying the stability of the hematopoietic system under stochastic perturbation. The paper is organized as follows. In section 2 we briefly present the mathematical preliminaries for linear differential delay equations needed for the rest of the paper. Section 3 examines the effect of stochastic perturbation on the behavior of the first and second moments of (1.7). This section contains the main mathematical results for the moment stability. The first moment is discussed in section 3.1. Section 3.2 considers the second moment and is divided into two parts according to the type of stochastic perturbation, namely, additive white noise and general cases. Section 4 studies the stability of the hematopoietic regulation system under stochastic perturbations. The paper concludes with a brief discussion in section 5.

In what follows, we will take $\tau = 1$ by normalizing the time through

$$(x, t, a, b, \sigma_i, \tau) \to (x, t/\tau, a/\tau, b/\tau, \sigma_i/\tau, 1).$$

Thus, we will study the equation

$$dx = (ax + bx_1)dt + (\sigma_0 x + \sigma_1 x_1 + \sigma_2)dW(t). \tag{1.8}$$

2. Mathematical preliminaries: The system without noise. When the σ_i in (1.8) are zero, we have the linear differential delay equation

$$\frac{dx}{dt} = ax + bx_1. \tag{2.1}$$

The differential delay equation (2.1) has been studied extensively, and [20] can be consulted for a detailed exposition.

The characteristic function of (2.1) is

$$h(\lambda) = \lambda - a - be^{-\lambda}. \tag{2.2}$$

The fundamental solution of (2.1), denoted by $X(t)$, has a Laplace transform given by $h^{-1}(\lambda)$ [20, Chapter 1]. This fundamental solution of (2.1) will be essential in following study.

Let $C([-1, 0], \mathbb{R})$ be the family of continuous functions ϕ from $[-1, 0]$ to $\mathbb{R}$ with the norm $\|\phi\| = \sup_{-1\le\theta\le 0} |\phi(\theta)|$. Using the fundamental solution, the solution of (2.1) with initial condition $x(\theta) = \phi(\theta) \in C([-1, 0], \mathbb{R})$ is given by

$$x_\phi(t) = X(t)\phi(0) + \int_{-1}^{0} X(t - 1 - s)\phi(s)ds. \tag{2.3}$$

From (2.3), the asymptotic behavior of $x_\phi(t)$ is determined by the fundamental solution $X(t)$. We have following result.

THEOREM 2.1 (see [20, Chapter 1, Theorem 5.2]). *If $\alpha_0 = \max\{\Re(\lambda) : h(\lambda) = 0\}$, then, for any $\alpha > \alpha_0$, there is a constant $K = K(\alpha)$ such that the fundamental solution X satisfies the inequality*

$$|X(t)| \le Ke^{\alpha t} \qquad (t \ge 0). \tag{2.4}$$

From Theorem 2.1, the solutions (2.3) with any $\phi(\theta) \in C([-1,0],\mathbb{R})$ approach 0 as $t \to \infty$ if and only if $\alpha_0 < 0$. The region in the (a,b)-plane such that $\alpha_0 < 0$ is given by [20]

$$S = \{(a,b) \in \mathbb{R}^2 \mid -a\sec\xi < b < a, \text{ where } \xi = a\tan\xi,\ a < 1,\ \xi \in (0,\pi)\}. \tag{2.5}$$

Here, the values of α_0 and $K(\alpha)$ are significant for understanding the stability of the system. The estimation of α_0 and $K(\alpha)$ are given below.

The number α_0 is given by the maximum real solution of the equation

$$(\alpha_0 - a)^2 - b^2 e^{-2\alpha_0} + \left[\arccos\frac{\alpha_0 - a}{be^{-\alpha_0}}\right]^2 = 0. \tag{2.6}$$

When $b \neq 0$, for any $\alpha > \alpha_0$ define

$$K(\alpha) = 1 + \xi(\alpha) + \frac{(|a-\alpha_0|e^{\alpha} + |b|)\log 2}{|b|\pi}, \tag{2.7}$$

where

$$\xi(\alpha) = \frac{1}{2\pi}\left|\int_{-2|b|e^{-\alpha}}^{2|b|e^{-\alpha}} \frac{a + be^{-(\alpha+iz)} - \alpha_0}{(\alpha - \alpha_0 + iz)h(\alpha+iz)}dz\right|.$$

Then (2.4) is satisfied. When $b = 0$, it is obvious that (2.4) holds with $K(\alpha) = 1$ whenever $\alpha \geq a$.

When $|b| < -a$, it is not difficult to prove that

$$|X(t)| \leq e^{(a+\mu)t} \quad (\forall t > 0), \tag{2.8}$$

where $|b| < \mu < -a$ is such that $\mu e^{a+\mu} - |b| = 0$. Thus, we can specify $K(\alpha) = 1$ with $\alpha = a + \mu$ when $|b| < -a$.

These considerations provide a framework for computing α_0 and $K(\alpha)$ with $\alpha > \alpha_0$ that satisfies (2.4).

3. Moment stability: The system with noise perturbation. We now turn to a study of the system with noise; i.e., the parameters σ_i in (1.8) are not all zero.

From the fundamental solution $X(t)$ in the previous section, the solution of (1.8) with the initial function $x(\theta) = \phi(\theta) \in C([-1,0],\mathbb{R})$ is a stochastic process given by

$$x(t;\phi) = x_\phi(t) + \int_0^t X(t-s)(\sigma_0 x(s;\phi) + \sigma_1 x_1(s;\phi) + \sigma_2)dW(s), \tag{3.1}$$

where $x_1(s;\phi) = x(s-1;\phi)$ and $x_\phi(t)$, the solution of the deterministic equation (2.1), is defined by (2.3). The existence and uniqueness theorem for the stochastic differential delay equation has been established in [24] (see also [41, Chapter 5]). The solution $x(t;\phi)$ is a stochastic process with distribution at any time t determined by the initial function $\phi(\theta)$. From the Chebyshev inequality, the possible range of x at time t is "almost" determined by its mean and variance at time t. Thus, the first and second moments of the solution are important for investigating the solution behavior and will be studied in this section. We first define pth moment exponential stability and pth moment boundedness.

DEFINITION 3.1. *The solution of* (1.8) *is said to be first moment exponentially stable if there is a pair of positive constants* λ *and* C *such that*

$$|Ex(t;\phi)| \le C\|\phi\|e^{-\lambda t} \quad (\forall t > 0)$$

for all $\phi \in C([-1,0],\mathbb{R})$. *When* $p \ge 2$, *the solution of* (1.8) *is said to be pth moment exponentially stable if there is a pair of positive constants* λ *and* C *such that*

$$E\left(|x(t;\phi) - E(x(t;\phi))|^p\right) \le C\|\phi\|^p e^{-\lambda t} \quad (\forall t \ge 0)$$

for all $\phi \in C([-1,0],\mathbb{R})$.

DEFINITION 3.2. *For* $p \ge 2$, *the solution of* (1.8) *is said to be pth moment bounded if there is a constant* A *such that*

$$E\left(|x(t;\phi) - E(x(t;\phi))|^p\right) \le A \quad (\forall t \ge 0)$$

for all $\phi \in C([-1,0],\mathbb{R})$. *Otherwise, the pth moment is said to be unbounded.*

We have used E to denote the mathematical expectation. In this paper, we will study the exponential stability of the first moment and the boundedness of the second moment. Hereinafter, we denote $x(t;\phi)$ simply by $x(t)$.

3.1. The first moment. Taking the mathematical expectation of both sides of (1.8), we have, with the Itô interpretation,

$$\frac{dEx(t)}{dt} = aEx(t) + bEx(t-1). \tag{3.2}$$

Thus, we obtain a differential delay equation for the first moment $Ex(t)$. From the discussion in the previous section, the first moment $Ex(t)$ approaches 0 as $t \to \infty$ if and only if the parameter α_0 defined in Theorem 2.1 is less than 0. In fact, by (3.1) and the properties of Itô integral, we have

$$Ex(t) = X(t)\phi(0) + \int_{-1}^{0} X(t-1-s)\phi(s)ds. \tag{3.3}$$

THEOREM 3.3. *If* $\alpha_0 = \max\{\Re(\lambda) : h(\lambda) = 0\}$, *then for any* $\alpha > \alpha_0$ *there is a constant* $K_1 = K_1(\alpha)$ *such that*

$$|Ex(t)| \le K_1\|\phi\|e^{\alpha t} \qquad (t \ge 0). \tag{3.4}$$

Therefore, if $\alpha_0 < 0$, *then* (1.8) *is first moment exponentially stable.*

3.2. The second moment. We now turn to the behavior of the second moment of the solution $x(t)$. From Theorem 3.3, the stability condition of the first moment is identical to that of the unperturbed system and is determined exclusively by a and b. Thus the stability of the first moment is independent of the parameters σ_i. However, the situation of second moment is more complicated and depends on σ_i. When $\sigma_2 \ne 0$, we cannot expect the second moment to be exponentially stable. Let $M(t)$ be the second moment of the solution at a time t. Then the Chebyshev inequality yields

$$P\left[|x(t) - Ex(t)| \ge k\sqrt{M(t)}\right] \le \frac{1}{k^2} \tag{3.5}$$

for any $k > 0$. Thus, when the second moment is bounded, the solutions of (1.8) are also bounded in some sense. We will answer in this section when the second moment is bounded for all $t > 0$.

The following notation will be used. Let $x(t)$ be a solution of (1.8), and define

$$\tilde{x}(t) = x(t) - Ex(t), \tag{3.6}$$

$$M(t) = E(\tilde{x}(t)^2), \quad M_1(t) = M(t-1), \quad N(t) = E(\tilde{x}(t)\tilde{x}(t-1)), \tag{3.7}$$

$$F(t) = \int_0^t X^2(t-s)(\sigma_0 Ex(s) + \sigma_1 Ex_1(s) + \sigma_2)^2 ds. \tag{3.8}$$

$M(t)$ is the second moment studied below. Applying the Itô isometry to $M(t)$, a simple computation yields that

$$M(t) = F(t) + \int_0^t X^2(t-s)(\sigma_0^2 M(s) + \sigma_1^2 M_1(s) + 2\sigma_0\sigma_1 N(s))ds. \tag{3.9}$$

3.2.1. Additive noise. When $\sigma_0 = \sigma_1 = 0$, we have the additive noise case, and the second moment is given explicitly by

$$M(t) = \sigma_2^2 \int_0^t X^2(t-s)ds. \tag{3.10}$$

By Theorem 2.1, we have the following result in the case of additive noise.

THEOREM 3.4. *Let* $\alpha_0 = \{\Re(\lambda) : h(\lambda) = 0\}$. *If* $\sigma_0 = \sigma_1 = 0$, *the second moment of* (1.8) *is bounded if and only if* $\alpha_0 < 0$. *Furthermore, for any* $\alpha_0 < \alpha < 0$, *there exists* $K = K(\alpha)$ *such that*

$$\left| M(t) - \sigma_2^2 \int_0^\infty X^2(s)ds \right| \le -\frac{\sigma_2^2 K^2}{2\alpha} e^{2\alpha t}. \tag{3.11}$$

From Theorem 3.4, the boundedness of the second moment is characterized by α_0, which is in turn determined by a and b of the unperturbed equation. This result was presented in [38], but the proof in [38] is in error. We reprove this result here and the estimation of the second moment $M(t)$ when $t \to \infty$ is given by

$$\lim_{t\to\infty} M(t) = \sigma_2^2 \int_0^\infty X^2(s)ds \le -\frac{\sigma_2^2 K^2}{2\alpha}. \tag{3.12}$$

3.2.2. General cases ($\sigma_0 \neq 0$ or $\sigma_1 \neq 0$). When $\sigma_0 \neq 0$ or $\sigma_1 \neq 0$, the noise at time t depends on x at time t or time $(t-1)$. In this general case, there is no simple form for the second moment. First, we have by (3.9) that

$$M(t) \ge F(t).$$

When $\alpha_0 \ge 0$, we have $X(t) = O(e^{\alpha_0 t})$ as $t \to \infty$. Thus, we can always take an initial function $\phi(\theta)$ such that $F(t)$ tends to infinity when $t \to \infty$, for example, the initial functions $\phi(\theta)$ such that

$$Ex(t) = \sum_i c_i e^{\lambda_i t},$$

where $h(\lambda_i) = 0$ and c_i are nonzero constants.

Therefore, we have the following necessary condition for the boundedness of the second moment.

LEMMA 3.5. *If the second moment of* (1.8) *is bounded, then* $\alpha_0 < 0$*; i.e., the unperturbed equation is exponentially stable.*

From now on, we will always assume that $\alpha_0 < 0$. In this situation, we have

$$\lim_{t\to\infty} F(t) = \sigma_2^2 \int_0^\infty X^2(s)ds. \tag{3.13}$$

We next study the second moment using the Laplace transform. We denote by $\mathcal{L}(p)(s)$ the Laplace transform of $p(t)$ when

$$p(t) < Pe^{at} \quad (t > 0)$$

for constants P and a.

Let $X_1(t) = X(t-1)$. It is easy to check that the functions $X^2(t)$, $X(t)X_1(t)$, $M(t)$, and $N(t)$ have Laplace transforms.

The following theorem establishes the condition for the second moment of the solution of (1.8) to be bounded.

THEOREM 3.6. *Let*

$$f(s) = \frac{\mathcal{L}(XX_1)(s)}{\mathcal{L}(X^2)(s)}, \qquad g(s) = \frac{\mathcal{L}(N)(s)}{\mathcal{L}(M)(s)}, \tag{3.14}$$

and

$$H(s) = s - (2a + \sigma_0^2) - (2bf(s) + 2\sigma_0\sigma_1 g(s)) - \sigma_1^2 e^{-s}. \tag{3.15}$$

The second moment of the solution of (1.8) *is bounded if and only if all solutions of the characteristic equation* $H(s) = 0$ *have negative real part. Furthermore, when the second moment is bounded, it approaches a constant exponentially when* $t \to \infty$.

Proof. We will divide the proof into several steps.

(1) By (3.9), we have

$$M(t) = F(t) + X^2 * (\sigma_0^2 M + \sigma_1^2 M_1 + 2\sigma_0\sigma_1 N),$$

where $*$ denotes convolution. Taking the Laplace transform of both sides and solving the resulting equation for $\mathcal{L}(M)(s)$, we have

$$\mathcal{L}(M)(s) = \frac{\mathcal{L}(F)(s)}{1 - \mathcal{L}(X^2)(s)(\sigma_0^2 + \sigma_1^2 e^{-s} + 2\sigma_0\sigma_1 g(s))}, \tag{3.16}$$

where $\mathcal{L}(X^2)(s)$ is given by

$$\mathcal{L}(X^2)(s) = \frac{1}{s - 2a - 2bf(s)}.$$

Thus, by (3.16), we have

$$\mathcal{L}(M)(s) = \frac{\mathcal{L}(F)(s)}{\mathcal{L}(X^2)(s)} H^{-1}(s).$$

Let

$$G(t) = \mathcal{L}^{-1}\left[\frac{\mathcal{L}(F)}{\mathcal{L}(X^2)}\right](t) = (\sigma_0 Ex(t) + \sigma_1 Ex_1(t) + \sigma_2)^2$$

and $Y(t) = \mathcal{L}^{-1}(H^{-1})(t)$. Then we have

$$M(t) = G * Y = \int_0^t G(t-s)Y(s)ds. \tag{3.17}$$

(2) Let $\beta_0 = \max\{\Re(s) : H(s) = 0\}$. We will prove that for any $\beta > \beta_0$ there is a constant $K_2 = K_2(\beta)$ such that

$$|Y(t)| \le K_2 e^{\beta t}. \tag{3.18}$$

To start, we will show that $\beta_0 < \infty$ is well defined. To do this, noting that there exist A_1 and A_2 such that if $\Re(s)$ is large enough,

$$|f(s)| \le A_1 e^{-\Re(s)/2} \quad \text{and} \quad |g(s)| \le A_2 e^{-\Re(s)/2}. \tag{3.19}$$

We omit the proof of (3.19) due to space constraints. Thus, when $\Re(s)$ is large enough, $H(s) > 0$, and therefore the value $\beta_0 < \infty$ is well defined.

Now, (3.18) follows from the inverse Laplace transform

$$Y(t) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{\beta-iT}^{\beta+iT} H^{-1}(s)e^{st}ds \le K_2 e^{\beta t}$$

for any $\beta \ge \beta_0$ and a constant $K_2 = K_2(\beta)$. The details of the proof are exactly the same as for the proof given in [20, pp. 20–21], and we omit the details here.

(3) Now we have

$$M(t) = \int_0^t G(t-s)Y(s)ds,$$

with $|Y(t)| \le K_2 e^{\beta t}$ for any $\beta > \beta_0$. Furthermore, for any $\alpha_0 < \alpha < 0$ there exists $K_3 = K_3(\alpha, \phi)$ such that

$$|G(t)| \le \sigma_2^2 + K_3 e^{\alpha t}.$$

If $\beta_0 < 0$, we choose $\beta_0 < \beta < 0$ and K_2 as above. Then

$$|M(t)| \le \int_0^t (\sigma_2^2 + K_3 e^{\alpha(t-s)})K_2 e^{\beta s}ds,$$

and thus the second moment $M(t)$ is bounded for any initial function $\phi(\theta)$. In this situation, let

$$M_\infty = \sigma_2^2 \int_0^\infty Y(s)ds,$$

so that

$$|M(t) - M_\infty| \le K_2\sigma_2^2 e^{\beta t} + \frac{K_2K_3}{\beta - \alpha}(e^{\beta t} - e^{\alpha t}).$$

Thus, there exists a positive constant $K_4 = K_4(\alpha, \beta, \phi)$ such that

$$|M(t) - M_\infty| \le K_4 e^{t\max\{\alpha,\beta\}};$$

i.e., $M(t)$ approaches to M_∞ exponentially when $t \to \infty$.

If $\beta_0 \geq 0$, by the inverse Laplace transform, we have $Y(t) = O(e^{\beta_0 t})$ when t is large enough. We can choose an initial function $\phi(\theta)$ such that

$$Ex(t) = \sum_i c_i e^{\lambda_i t},$$

where $h(\lambda_i) = 0$ and c_i are nonzero constants. For this particular initial function, we have either $G(t) = O(1)$ as $t \to \infty$ when $\sigma_2 \neq 0$, or $G(t) = O(e^{2\alpha t})$ as $t \to \infty$ for some $\alpha \leq \alpha_0 < 0$ when $\sigma_2 = 0$. In either case,

$$M(t) = \int_0^t G(t-s)Y(s)ds = O(e^{\beta_0 t})$$

when $t \to \infty$, and hence the second moment is unbounded. ∎

Theorem 3.6 establishes a criterion for the second moment of the linear stochastic delay differential equation to be bounded. However, this criterion is not particularly useful. The function $g(s)$ in (3.15) involves the Laplace transforms of $M(t)$ and $N(t)$ that are unknown. In many applications, perturbations for system parameters affect only the right-hand side of the equation that involves either the current state or the retarded state, and thus either $\sigma_1 = 0$ or $\sigma_0 = 0$. In this situation, the function $H(s)$ reads

$$H(s) = s - (2a + \sigma_0^2) - 2bf(s) - \sigma_1^2 e^{-s}$$

and is determined by the system coefficients and by $f(s)$, which depends on the Laplace transforms of $X^2(t)$ and $X(t)X_1(t)$. Nevertheless, it is not trivial to obtain the explicit form of $f(s)$ for a given system. In the rest of this section, we will develop some estimates for $f(s)$ and $g(s)$ and present direct criteria for the second moment stability.

THEOREM 3.7. *If $b < 0$, $\sigma_0\sigma_1 \leq 0$, and either*

$$(\sigma_0 + \sigma_1)^2 \geq -2(a+b) \tag{3.20}$$

or

$$u = \begin{cases} \dfrac{-(b+\sigma_0\sigma_1) - \sqrt{(b+\sigma_0\sigma_1)^2 - 4\sigma_1^2}}{2\sigma_1^2}, & \sigma_1 \neq 0, \\ -\dfrac{1}{b}, & \sigma_1 = 0, \end{cases}$$

such that $0 < u < 1$ and

$$-2\log u - (2a + \sigma_0^2) - (2b + 2\sigma_0\sigma_1)u - \sigma_1^2 u^2 \leq 0, \tag{3.21}$$

then the second moment is unbounded.

Proof. Let

$$H_0(s) = s - (2a + \sigma_0^2) - (2b + 2\sigma_0\sigma_1)e^{-s/2} - \sigma_1^2 e^{-s}.$$

Then when $b < 0$ and $\sigma_0\sigma_1 \leq 0$, we have $H(s) \leq H_0(s)$ for all $s \in \mathbb{R}$. Therefore, either (3.20) or (3.21) implies that there exists $s_* > 0$ such that $H(s) \leq H_0(s_*) \leq 0$. However, $H(s) > 0$ when s is large enough. Therefore the equation $H(s) = 0$ has a nonnegative solution. Thus, the theorem follows from Theorem 3.6. ∎

Theorem 3.7 tells us when the second moment is unbounded. The following result will tell us when the second moment is bounded.

THEOREM 3.8. *If there exists* $\alpha < 0$ *and* $K = K(\alpha) > 0$ *such that*

$$|X(t)| \le Ke^{\alpha t} \quad (t \ge 0) \tag{3.22}$$

and

$$(|\sigma_0| + |\sigma_1|)^2 < -\frac{2\alpha}{K^2}, \tag{3.23}$$

then the second moment $M(t)$ *is bounded when* $t > 0$.

Theorem 3.8 will be proved using the following delay-type Gronwall inequality, the proof of which (which we omit) is similar to that of the Gronwall inequality.

LEMMA 3.9. *If* $y(t)$ *is a nonnegative continuous function on* $[-1, \infty)$ *and there are positive constants* p *and* q *such that*

$$y(t) \le p\int_0^t y(s)ds + q\int_0^t y(s-1)ds + r(t), \tag{3.24}$$

then for any $\beta > 0$ *such that*

$$\beta - p - qe^{-\beta} > 0$$

and

$$\sup_{t\ge 0} |r(t)e^{-\beta t}| < \infty$$

there exists $A = A(\beta)$ *such that*

$$y(t) \le Ae^{\beta t} \quad (t \ge 0). \tag{3.25}$$

We can now turn to the proof of Theorem 3.8.

Proof of Theorem 3.8. Note that

$$M(t) \le (|\sigma_0| + |\sigma_1|)\int_0^t X^2(t-s)(|\sigma_0|M(s) + |\sigma_1|M_1(s))ds + F(t). \tag{3.26}$$

For α such that (3.22) is satisfied, we have $K_5 = K_5(\alpha, \phi)$ such that

$$0 \le F(t) \le K_5(1 - e^{2\alpha t}).$$

Thus from (3.26) it follows that

$$M(t) \le K^2\,(|\sigma_0| + |\sigma_1|)e^{2\alpha t}\int_0^t (|\sigma_0|e^{-2\alpha s}M(s) + |\sigma_1|e^{-2\alpha s}M_1(s))ds + K_5(1 - e^{2\alpha t}).$$

Let

$$y(t) = e^{-2\alpha t}M(t), \qquad r(t) = K_5(e^{-2\alpha t} - 1),$$

and

$$p = K^2|\sigma_0|(|\sigma_0| + |\sigma_1|), \qquad q = K^2|\sigma_1|(|\sigma_0| + |\sigma_1|)e^{-2\alpha}.$$

Then

$$y(t) \le p\int_0^t y(s)ds + q\int_0^t y_1(s)ds + r(t).$$

The inequality (3.23) implies

$$-2\alpha - p - qe^{2\alpha} > 0.$$

Thus, by Lemma 3.9 there is a constant A such that

$$M(t)e^{-2\alpha t} = y(t) \le Ae^{-2\alpha t},$$

i.e., $M(t) \le A$ for all $t > 0$. □

From Theorem 3.6, if the second moment $M(t)$ is bounded, it exponentially approaches a constant M_∞. Let $t \to \infty$ in (3.26) and apply (3.13) so that we have

$$M_\infty \le \frac{\sigma_2^2 \int_0^\infty X^2(s)ds}{1 - (|\sigma_0| + |\sigma_1|)^2 \int_0^\infty X^2(s)ds} \le -\frac{\sigma_2^2 K^2(\alpha)}{2\alpha + (|\sigma_0| + |\sigma_1|)^2 K^2(\alpha)} \tag{3.27}$$

for any α and $K(\alpha)$ given in Theorem 3.8. It follows from (3.27) that when $\sigma_2 = 0$, boundedness of the second moment implies exponential stability.

From Theorems 3.3 and 3.8, for any parameter pair (a, b) in the region S defined in (2.5), the first moment of the solution of the stochastic differential delay equation (1.8) approaches 0 as $t \to \infty$. Furthermore, there exists $P(a, b) > 0$ such that if

$$(|\sigma_0| + |\sigma_1|)^2 < P(a, b), \tag{3.28}$$

the second moment of the solution is bounded with an upper bound, as $t \to \infty$, given by (3.27).

From the estimates of α_0 and $K(\alpha)$ given in section 2, the function $P(a, b)$ can be computed, and its graph is shown in Figure 3.1.

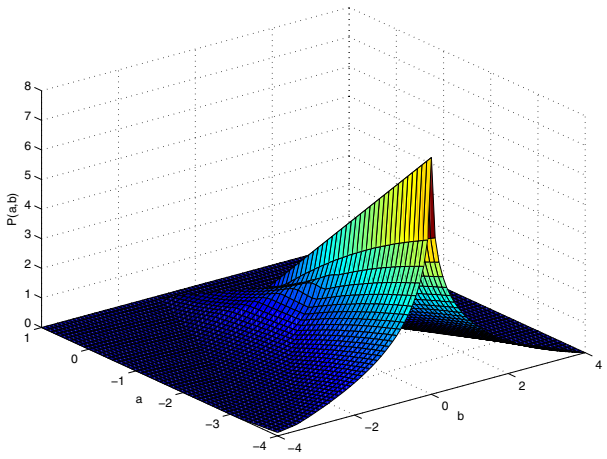


FIG. 3.1. *The function* $P(a, b)$.

4. Stability of the hematopoietic regulation system under stochastic perturbation. In this section, we will study the stability of the HSC in the face of stochastic perturbation, using the results of the previous sections. The HSC regulation system is modeled by a classical G_0 model [6, 35, 56]. Blood cells differentiate from HSC in the resting (G_0) phase. The HSC has high self-renewal capacity with the re-entry rate dependent on the number of HSC through a negative feedback loop. The proliferating phase cells include those cells in S phase (DNA synthesis), M phase (mitosis), and two segments known as the G_1 and G_2 phases (the G stands for "gap"). In addition, there is a loss of proliferating phase cells due to apoptosis (see Figure 4.1).

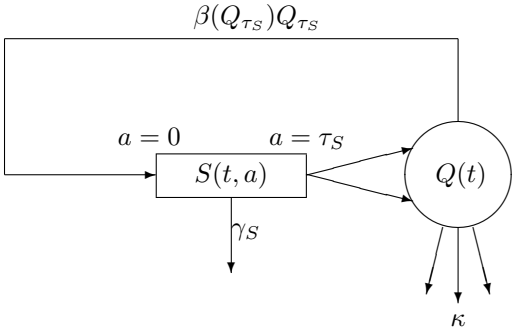


FIG. 4.1. *A cartoon representation of the HSC model.*

The HSC dynamics is modeled by a differential delay equation [7, 8, 9, 13, 51]

$$\frac{dQ}{dt} = -\beta(Q)Q - \kappa(N,R,P)Q + 2e^{-\gamma_S\tau_S}\beta(Q_{\tau_S})Q_{\tau_S}, \tag{4.1}$$

where Q, N, R, P are the quiescent stem cells, leukocytes, erythrocytes, and platelets, respectively, and $Q_{\tau_S} = Q(t-\tau_S)$. The model parameters are the apoptosis (death) rate γ_S, the maturation delay τ_S, the HSC self-renewal (proliferation) rate β, and the differentiation rate κ at which the HSC forms the three peripheral cell lines. The proliferation rate β and differentiation rate κ involve negative feedback loops that take the form of a Hill function [5, 7, 8],

$$\beta(Q) = k_0\frac{\theta_2^s}{\theta_2^s + Q^s},$$

$$\kappa(N,P,R) = f_0\frac{\theta_1^n}{\theta_1^n + N^n} + \frac{\bar{\kappa}_p}{1+K_pP^p} + \frac{\bar{\kappa}_r}{1+K_rR^r}.$$

Note that the rate κ depends on the state of three cell lines (leukocytes, erythrocytes, platelets), and thus (4.1) does not constitute a closed system. In this study, since we are interested only in the situation close to the steady state, we take the total differentiation out of the stem cell compartment to be a single constant. This decouples the model for the stem cell compartment from the full system.

Let us introduce nondimensional variables as follows:

$$q = \frac{Q}{\theta_2}, \quad \hat{t} = \frac{t}{\tau_S},$$

$$b_1 = \tau_S k_0, \quad \mu_1 = 2e^{-\gamma_S\tau_S}, \quad \delta = \tau_S\kappa.$$

We have the nondimensional form of (4.1) (see [9]),

$$\frac{dq}{dt} = -\frac{b_1}{1+q^s}q + \mu_1 \frac{b_1}{1+q_1^s}q_1 - \delta q, \tag{4.2}$$

where $q_1 = q(t-1)$ and $\hat{t}$ has been simply replaced by t. Typical values of the dimensionless parameters are $b_1 = 22.4$ and $\mu_1 = 1.64$. The parameter s, which denotes the number of cytokine molecules needed to trigger HSC proliferation in vitro, is chosen as $s = 4$ (see [9]).

When $\delta < b_1(\mu_1 - 1)$, equation (4.2) has a unique positive steady state

$$q^* = \left(\frac{b_1(\mu_1 - 1)}{\delta} - 1\right)^{1/4},$$

corresponding to the normal level of the stem cells. Linearizing (4.2) around this steady state, we obtain the variational equation

$$\frac{dx}{dt} = ax + bx_1,$$

where $x = q - q^*$,

$$a = -\frac{\delta}{b_1(\mu_1 - 1)^2}(-3b_1(\mu_1 - 1) + b_1(\mu_1 - 1)^2 + 4\delta),$$

and

$$b = \frac{\delta\mu_1}{b_1(\mu_1 - 1)^2}(-3b_1(\mu_1 - 1) + 4\delta).$$

From the discussion in section 2, there exists a critical value δ_c (≈ 0.16) such that the steady state is stable when $0 < \delta < \delta_c$.

We will now study the stability of the steady state when there are stochastic perturbations in the system parameters δ, b_1, or μ_1. We have the following equations for the perturbed system:

1. perturbation in δ:

$$dq = \left[-\frac{b_1 q}{1+q^4} - \delta q + \frac{b_1\mu_1 q_1}{1+q_1^4}\right]dt - \sigma q dW(t), \tag{4.3}$$

2. perturbation in b_1:

$$dq = \left[-\frac{b_1 q}{1+q^4} - \delta q + \frac{b_1\mu_1 q_1}{1+q_1^4}\right]dt + \sigma\left[-\frac{q}{1+q^4} + \frac{\mu_1 q_1}{1+q_1^4}\right]dW(t), \tag{4.4}$$

3. and perturbation in μ_1:

$$dq = \left[-\frac{b_1 q}{1+q^4} - \delta q + \frac{b_1\mu_1 q_1}{1+q_1^4}\right]dt + \sigma\frac{b_1 q_1}{1+q_1^4}dW(t), \tag{4.5}$$

where $W(t)$ is the standard Wiener process and σ is the noise amplitude. The linearized versions of (4.3)–(4.5) around the steady state $q = q^*$ are

$$dx = (ax + bx_1)dt - \sigma(x + q^*)dW(t), \tag{4.6}$$

$$dx = (ax + bx_1)dt + \left(\frac{\sigma}{b_1}\right)((a+\delta)x + bx_1 + \delta q^*)dW(t), \tag{4.7}$$

and

$$dx = (ax + bx_1)dt + \left(\frac{\sigma}{\mu_1}\right)\left(bx_1 + \delta q^* + \frac{b_1 q^*}{1 + q^{*4}}\right) dW(t), \tag{4.8}$$

respectively.

Applying the results from the previous section, when $0 < \delta < \delta_c$, the first moment of solutions of the perturbed system is locally stable. Furthermore, from Theorem 3.8, for any $0 < \delta < \delta_c$ (and b_1, μ_1 held at their typical values), there exists $\sigma_b(\delta)$ such that when $\sigma < \sigma_b(\delta)$ the second moment is bounded. Further, from Theorem 3.7, for any $0 < \delta < \delta_c$, there exists $\sigma_u(\delta)$ such that when $\sigma > \sigma_u(\delta)$ the second moment is unbounded. When $\sigma_b < \sigma < \sigma_u$, however, the previous results fail to delineate the stability of the second moment. A more accurate estimation for the characteristic function $H(s)$ in Theorem 3.6 is required to fill this gap. Graphs of the curves $\sigma = \sigma_b(\delta)$ and $\sigma = \sigma_u(\delta)$ are given in Figure 4.2.

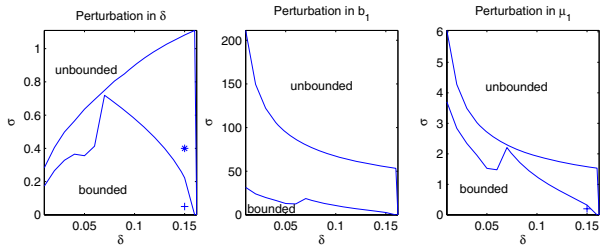


FIG. 4.2. *Parameter dependence for the second moment of the solution of the stochastic HSC system to be bounded.*

In Figure 4.2, the thresholds of the three parameter perturbations to ensure the stability of a steady state under noise perturbation are sorted from low to high as the threshold for δ, for μ_1, and for b_1. Thus, the HSC system is more easily destabilized by noise in the differentiation rate (δ) than in the death rate (μ_1), and least likely to be destabilized by perturbations in the proliferation rate (b_1).

Note that the solutions of (4.3)–(4.5) are always bounded because of the negative feedback. Thus, destabilization of the steady state may lead to fluctuating solutions characteristic of dynamic hematological disease (see Figure 4.3(b)). When the second moment is bounded, the range of the solution at time t can be estimated by the Chebyshev inequality (3.5). However, this cannot exclude the possibility of obtaining an oscillating solution (see Figure 4.3(c),(d)). In this situation, the amplitude of the oscillating solution is determined by the second moment, which is estimated by (3.27) when t is large. The graphs of M_∞ as a function of δ and σ are shown in Figure 4.4 for each of the cases. From Figure 4.4, for given values of b_1, μ_1, δ, and the perturbation amplitude σ, the second moment of the solutions of the HSC model with random perturbation in δ is larger than when there are perturbations in μ_1 and in b_1. Thus, small fluctuations in δ are able to produce large amplitude fluctuations in HSC numbers. Larger perturbations in μ_1 are required to produce a fluctuating HSC solution with the same amplitude (Figure 4.3(c),(d)). The second moment of solutions of the HSC system with perturbations in b_1 is small and not likely to produce a large amplitude fluctuation in HSC numbers.

These numerical results suggest that the dynamic hematological diseases [16] characterized by oscillations in blood cell numbers could originate from the stochastic

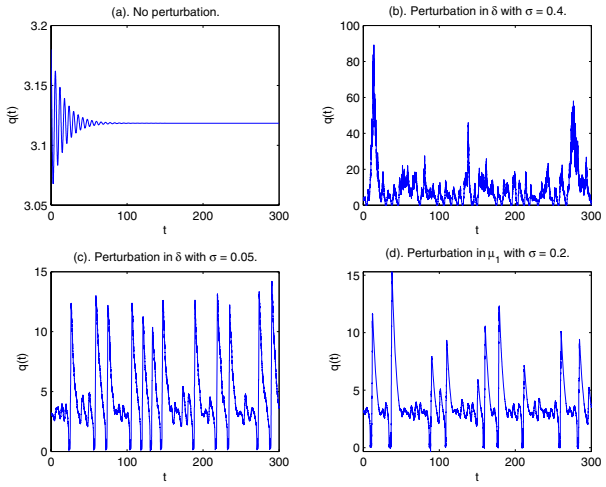


FIG. 4.3. *Sample solutions of the HSC system. In all these solutions, we choose* $b_1 = 22.4$, $\mu_1 = 1.64$, $\delta = 0.15$. *The perturbation is added to* δ *with* $\sigma = 0.4$ *and* $\sigma = 0.05$ *(cf. Figure* 4.2, *left-hand panel, marked by the "*" and "+," respectively), and to* μ_1 *with* $\sigma = 0.2$ *(cf. Figure* 4.2, *middle panel, marked with a "+"). The solution of the system without perturbation is also shown.*

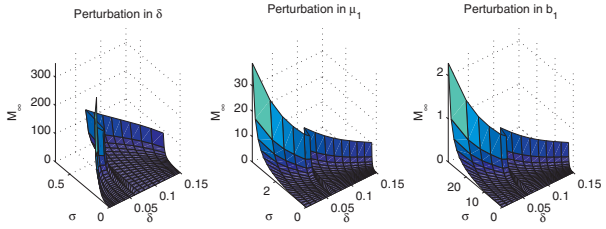


FIG. 4.4. *The function* M_∞ *as a function of* δ *and* σ.

perturbation of the differentiation rate and/or the death rate of HSCs. On the other hand, the system is relatively insensitive to perturbations in the proliferation rate.

5. Discussion. We have investigated the effects of white noise on the stability of the trivial solution of a linear differential delay equation by deriving the solutions for the first and second order moments and examining the exponential estimation by the Laplace transform method.

We have shown that the stability domain of the first moment is identical to that of the unperturbed system (Theorem 3.3). This result is also true for the second moment when the perturbation is simple additive noise (Theorem 3.4). However, when there is multiplicative white noise, there are no simple results on the stability (bounded nature) of the second moment. From our study, when the trivial solution of the unperturbed equation is unstable, the second moment of solutions of the perturbed equation is unbounded. The condition for the second moment to be bounded has been shown to be related to the solutions of a characteristic equation given in Theorem 3.6. Nevertheless, the explicit expression of the characteristic equation is not available in terms of the system parameters. We have presented several direct criteria for the second moment to be bounded (Theorems 3.7 and 3.8).

Significant oscillations in one or more of the circulating progeny of the HSC

are often characteristic of dynamic hematological diseases like cyclical neutropenia, cyclical thrombocytopenia, and periodic chronic myelogenous leukemia. The steady state of the HSC system can be destabilized by increasing the differentiation rate, and this has been implicated in the genesis of the hematological disorder cyclical neutropenia [21].

We have applied these results to examining the stability of HSC dynamics in the presence of stochastic perturbation. Our results indicate that stochastic perturbation cannot stabilize a large amplitude oscillation solution. When random perturbations are introduced in parameters characterizing the HSCs when the steady state is locally stable, we found that as the amplitude of the noise perturbation is increased, the system can be destabilized in the sense that the second moment becomes unbounded. In this situation, the system can display a large amplitude fluctuating solution.

When the second moment is large and bounded, however, we cannot exclude the possibility of an oscillatory solution, since the HSC system may have a large amplitude oscillatory solution in this circumstance.

We have obtained estimates of the second moment for three different types of perturbation (see Figure 4.4). These results suggest that small perturbations in the HSC differentiation or apoptosis (death) rate are able to generate large amplitude fluctuations in HSC numbers, but a much larger perturbation of the proliferation rate is needed to generate comparable fluctuations in HSC numbers. These results suggest that the HSC model system is more sensitive to random perturbations in the differentiation or death rate than in the proliferation rate.

The results in this paper were obtained under the Itô interpretation of stochastic integrals. Analogous results can be obtained for the Stratonovich interpretation. When Stratonovich interpretation is used, the solution of (1.8) can be expressed as

$$x(t) = x(0) + \int_0^t (\tilde{a}x(s) + \tilde{b}x_1(s) + \tilde{c})ds + \int_0^t (\sigma_0 x(s) + \sigma_1 x_1(s) + \sigma_2)dW(s), \tag{5.1}$$

in terms of the Itô integral, where

$$\tilde{a} = a + \frac{1}{2}\sigma_0^2, \quad \tilde{b} = b + \frac{1}{2}\sigma_0\sigma_1, \quad \tilde{c} = \frac{1}{2}\sigma_0\sigma_2. \tag{5.2}$$

Unlike the situation when the Itô interpretation is used, namely that the first moment stability is determined merely by the unperturbed system, the first moment stability is changed when we use a Stratonovich interpretation. Let

$$\tilde{h}(\lambda) = \lambda - \tilde{a} - \tilde{b}e^{-\lambda}, \qquad x_* = -\frac{\tilde{c}}{\tilde{a} + \tilde{b}},$$

so that we have the following theorem.

THEOREM 5.1. *If $\tilde{\alpha}_0 = \max\{\Re(\lambda) : \tilde{h}(\lambda) = 0\}$, then, for any $\alpha > \alpha_0$, there is a constant $\tilde{K} = \tilde{K}(\alpha)$ such that*

$$|Ex(t;\phi) - x_*| \le \tilde{K}\|\phi\|e^{\alpha t} \tag{5.3}$$

for any $\phi \in C([-1, 0], \mathbb{R})$.

When $\tilde{\alpha}_0 < 0$, Theorem 5.1 implies that $Ex(t;\phi)$ approaches x_* exponentially when $t \to +\infty$. Thus, the expectation of the solutions drifts from zero to x_* due to the stochastic perturbation. Note that $\tilde{a} \ge a$; it is easy to show that the stochastic perturbation is able to destabilize the first moment of (1.8). On the other hand, the

first moment cannot be stabilized by stochastic perturbation when the zero solution of the system without noise is unstable.

To study the second moment, let $y = x - x_*$, so that $y(t)$ satisfies

$$y(t) = y(0) + \int_0^t (\tilde{a}y(s) + \tilde{b}y_1(s))ds + \int_0^t (\sigma_0 y(s) + \sigma_1 y_1(s) + \tilde{\sigma}_2)dW(s), \tag{5.4}$$

where

$$\tilde{\sigma}_2 = \sigma_2 + (\sigma_0 + \sigma_1)x_*.$$

Applying the results in section 3 to the Itô equation (5.4), we can obtain the corresponding results for second moment stability of (1.8) in terms of the Stratonovich interpretation. The statement of these results is straightforward by replacing a, b, and σ_2 with $\tilde{a}$, $\tilde{b}$, and $\tilde{\sigma}_2$, respectively, and we omit them here.

Acknowledgments. J. Lei would like to thank Dr. Caroline Colijn for helpful discussions. The authors would like to thank the referees for their helpful comments.

REFERENCES

[1] J. W. Adamson, *The relationship of erythropoietin and iron metabolism to red blood cell production in humans*, Sem. Oncol., 21 (1994), pp. 9–15.
[2] J. Appleby and X. Mao, *Stochastic stabilisation of functional differential equations*, Systems Control Lett., 54 (2005), pp. 1069–1081.
[3] R. Bellman and K. L. Cooke, *Differential-Difference Equations*, Academic, New York, 1963.
[4] D. L. Bennett and S. A. Gourley, *Asymptotic properties of a delay differential equation model for the interaction of glucose with plasma and interstitial insulin*, Appl. Math. Comput., 151 (2004), pp. 189–207.
[5] S. Bernard, J. Bélair, and M. C. Mackey, *Oscillations in cyclical neutropenia: New evidence based on mathematical modeling*, J. Theoret. Biol., 223 (2003), pp. 283–298.
[6] F. Burns and I. Tannock, *On the existence of a G_0 phase in the cell cycle*, Cell Tissue Kinet., 3 (1970), pp. 321–334.
[7] C. Colijn and M. C. Mackey, *A mathematical model of hematopoiesis:* I. *Periodic chronic myelogenous leukemia*, J. Theoret. Biol., 237 (2005), pp. 117–132.
[8] C. Colijn and M. C. Mackey, *A mathematical model of hematopoiesis:* II. *Cyclical neutropenia*, J. Theoret. Biol., 237 (2005), pp. 133–146.
[9] C. Colijn and M. C. Mackey, *Bifurcation and bistability in a model of hematopoietic regulation*, SIAM J. Appl. Dyn. Syst., submitted.
[10] R. D. Driver, *Ordinary and Delay Differential Equations*, Appl. Math. Sci. 20, Springer-Verlag, New York, 1977.
[11] L. E. El'sgol'ts and S. B. Norkin, *Introduction to the Theory and Application of Differential Equations with Deviating Argument*, Academic, New York, 1973.
[12] P. Fortin and M. C. Mackey, *Periodic chronic myelogenous leukaemia: Spectral analysis of blood cell counts and aetiological implications*, Br. J. Haematol., 104 (1999), pp. 336–345.
[13] A. C. Fowler and M. C. Mackey, *Relaxation oscillations in a class of delay differential equations*, SIAM J. Appl. Math., 63 (2002), pp. 299–323.
[14] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, 3rd ed., Springer-Verlag, New York, 2004.
[15] L. Glass and M. C. Mackey, *Pathological conditions resulting from instabilities in physiological control systems*, Ann. New York Acad. Sci., 316 (1979), pp. 214–235.
[16] L. Glass and M. C. Mackey, *From Clocks to Chaos: The Rhythms of Life*, Princeton University Press, Princeton, NJ, 1988.
[17] S. A. Gourley and Y. Kuang, *A delay reaction-diffusion model of the spread of bacteriophage infection*, SIAM J. Appl. Math., 65 (2005), pp. 550–566.
[18] A. Halanay, *Differential Equations: Stability, Oscillations, Time Lags*, Academic, New York, 1966.
[19] J. K. Hale, *Functional Differential Equations*, Springer-Verlag, New York, 1971.
[20] J. K. Hale, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[21] C. Haurie, D. C. Dale, and M. C. Mackey, *Cyclical neutropenia and other periodic hematological diseases: A review of mechanisms and mathematical models*, Blood, 92 (1998), pp. 2629–2640.
[22] C. Haurie, M. C. Mackey, and D. C. Dale, *Occurrence of periodic oscillations in the differential blood counts of congenital, idiopathic and cyclical neutropenic patients before and during treatment with G-CSF*, Exper. Hematol., 27 (1997), pp. 401–409.
[23] C. Haurie, R. Person, D. C. Dale, and M. C. Mackey, *Haematopoietic dynamics in grey collies*, Exper. Hematol., 27 (1999), pp. 1139–1148.
[24] K. Ito and M. Nisio, *On stationary solutions of a stochastic differential equations*, J. Math. Kyoto Univ., 4 (1964), pp. 1–75.
[25] A. F. Ivanov, Y. I. Kazmerchuk, and A. V. Swishchuk, *Theory, stochastic stability and applications of stochastic delay differential equations: A survey of results*, Differential Equations Dynam. Systems, 11 (2003), pp. 55–115.
[26] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, North–Holland, Amsterdam, 1992.
[27] U. Küchler and B. Mensch, *Langevin's stochastic differential equation extended by a time-delayed term*, Stochastics Stochastics Rep., 40 (1992), pp. 23–42.
[28] A. Lasota and M. C. Mackey, *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*, 2nd ed., Springer-Verlag, New York, 1994.
[29] A. Longtin, *Effects of noise on nonlinear dynamics*, in Nonlinear Dynamics in Physiology and Medicine, A. Beuter, L. Glass, M. C. Mackey, and M. S. Titcombe, eds., Springer-Verlag, New York, 2003, pp. 149–189.
[30] M. C. Mackey, *Mathematical models of hematopoietic cell replication and control*, in The Art of Mathematical Modelling: Case Studies in Ecology, Physiology and Biofluids, H. G. Othmer, F. R. Adler, M. A. Lewis, and J. C. Dallon, eds., Prentice–Hall, New York, 1997, pp. 149–178.
[31] M. C. Mackey, *Unified hypothesis for the origin of aplastic anemia and periodic haematopoiesis*, Blood, 51 (1978), pp. 941–956.
[32] M. C. Mackey, *Dynamic haematological disorders of stem cell origin*, in Biophysical and Biochemical Information Transfer in Recognition, J. G. Vassileva-Popova and E. V. Jensen, eds., Plenum Publishing, New York, 1979, pp. 373–409.
[33] M. C. Mackey, *Periodic auto-immune hemolytic anemia: An induced dynamical disease*, Bull. Math. Biol., 41 (1979), pp. 829–834.
[34] M. C. Mackey and U. van der Heiden, *Dynamic diseases and bifurcations in physiological control systems*, Funk. Biol. Med., 1 (1982), pp. 156–164.
[35] M. C. Mackey and P. Dörmer, *Continuous maturation of proliferating erythroid precursors*, Cell Tissue Kinet., 15 (1982), pp. 381–392.
[36] M. C. Mackey and J. G. Milton, *Dynamical diseases*, Ann. New York Acad. Sci., 504 (1987), pp. 16–32.
[37] M. C. Mackey and J. G. Milton, *Feedback, delays, and the origins of blood cell dynamics*, Comm. Theoret. Biol., 1 (1990), pp. 299–237.
[38] M. C. Mackey and I. G. Nechaeva, *Solution moment stability in stochastic differential delay equations*, Phys. Rev. E (3), 52 (1995), pp. 3366–3376.
[39] M. C. Mackey, *Cell kinetic status of haematopoietic stem cells*, Cell Prolif., 43 (2001), pp. 71–83.
[40] X. Mao, *Exponential Stability of Stochastic Differential Equations*, Marcel Dekker, New York, 1994.
[41] X. Mao, *Stochastic Differential Equations and Their Applications*, Horwood Publishing, Chichester, UK, 1997.
[42] X. Mao, *Almost sure exponential stability of delay equations with damped stochastic perturbation*, Stoch. Anal. Appl., 19 (2001), pp. 67–84.
[43] X. Mao and S. Sabanis, *Numerical solutions of stochastic differential delay equations under local Lipschitz condition*, J. Comput. Appl. Math., 151 (2003), pp. 215–227.
[44] X. Mao and M. J. Rassias, *Khasminskii-type theorems for stochastic differential delay equations*, Stoch. Anal. Appl., 23 (2005), pp. 1045–1069.
[45] J. G. Milton and M. C. Mackey, *Periodic haematological diseases: Mystical entities or dynamical disorders?*, J. Roy. Coll. Phys. (Lond.), 23 (1989), pp. 236–241.
[46] J. G. Milton, U. van der Heiden, A. Longtin, and M. C. Mackey, *Complex dynamics and noise in simple neural networks with delayed mixed feedback*, Biomed. Biochim. Acta, 49 (1990), pp. 697–707.
[47] S. E. A. Mohammed, *Stochastic Functional Differential Equations*, Res. Notes in Math. 99, Pitman, Boston, 1984.

[48] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 6th ed., Springer-Verlag, New York, 2003.

[49] S. Pavord, M. Sivakumaran, P. Furber, and V. Mitchell, *Cyclical thrombocytopenia as a rare manifestation of myelodysplastic syndrome*, Clin. Lab. Haematol., 18 (1996), pp. 221–223.

[50] T. H. Price, G. S. Chatta, and D. C. Dale, *Effect of recombinant granulocyte colony-stimulating factor on neutrophil kinetics in normal young and elderly humans*, Blood, 88 (1996), pp. 335–340.

[51] L. Pujo-Menjouet, S. Bernard, and M. C. Mackey, *Long period oscillations in a G_0 model of hematopoietic stem cells*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 312–332.

[52] L. Pujo-Menjouet and M. C. Mackey, *Contribution to the study of periodic chronic myelogenous leukemia*, C. R. Biol., 327 (2004), pp. 235–244.

[53] M. Z. Ratajczak, J. Ratajczak, W. Marlicz, C. H. Pletcher, Jr., B. Machalinshi, J. Moore, H. Hung, and A. M. Gewirtz, *Recombinant human thrombopoietin (TPO) stimulates erythropoiesis by inhibiting erythroid progenitor cell apoptosis*, Br. J. Haematol., 98 (1997), pp. 8–17.

[54] M. Santillan, J. M. Mahaffy, J. Belair, and M. C. Mackey, *Regulation of platelet production: The normal response to perturbation and cyclical platelet disease*, J. Theoret. Biol., 206 (2000), pp. 585–603.

[55] M. Silva, D. Grillot, A. Benito, C. Richard, G. Nunez, and J. Fernandez-Luna, *Erythropoietin can promote erythroid progenitor survival by repressing apoptosis through bcl-1 and bcl-2*, Blood, 88 (1996), pp. 1576–1582.

[56] J. A. Smith and L. Martin, *Do cells cycle?*, Proc. Natl. Acad. Sci. USA, 70 (1973), pp. 1263–1267.

[57] R. L. Stratonovich, *A new representation for stochastic integrals and equations*, SIAM J. Control, 4 (1966), pp. 362–371.

[58] S. Tanimukai, T. Kimura, H. Sakabe, Y. Ohmizono, T. Kato, H. Miyazaki, H. Yamagishi, and Y. Sonoda, *Recombinant human c-Mpl ligand (thrombopoietin) not only acts on megakaryocyte progenitors, but also on erythroid and multipotential progenitors in vitro*, Exp. Hematol., 25 (1997), pp. 1025–1033.

[59] P. Wahlberg, D. Nyman, P. Ekelund, S. A. Carlsson, and H. Granlund, *Cyclical thrombocytopenia with remission during lynestrenol treatment in a woman*, Ann. Clin. Res., 9 (1977), pp. 356–358.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 408–433

SOME VECTOR BORNE DISEASES WITH STRUCTURED HOST POPULATIONS: EXTINCTION AND SPATIAL SPREAD*

STEPHEN A. GOURLEY†, RONGSONG LIU‡, AND JIANHONG WU§

Abstract. We derive from a structured population model a system of delay differential equations describing the interaction of five subpopulations, namely susceptible and infected adult and juvenile reservoirs and infected adult vectors, for a vector borne disease with particular reference to West Nile virus, and we also incorporate spatial movements by considering the analogue reaction-diffusion systems with nonlocal delayed terms. Specific conditions for the disease eradication and sharp conditions for the local stability of the disease-free equilibrium are obtained using comparison techniques coupled with the spectral theory of monotone linear semiflows. A formal calculation of the minimal wave speed for the traveling waves is given and compared with field observation data.

Key words. stage-structure, epidemic, delay, traveling front, vector borne disease

AMS subject classifications. 34K20, 34K60, 35K57, 92D25

DOI. 10.1137/050648717

1. Introduction.

Vector borne diseases are infectious diseases that are carried by insects from one host to another. Examples include malaria, West Nile virus, yellow fever, dengue fever, lyme disease, and plague. In many of these diseases it is the mosquito that carries the virus, but ticks and fleas can also be responsible. The diseases can be spread to humans, birds, and other animals.

Much has been done in terms of modeling and analysis of the transmission dynamics and spatial spread of vector borne diseases; see Anderson and May [1], Murray [20], Brauer and Castillo-Chavez [4], Edelstein-Keshet [6], Hethcote [10], Kot [13], Jones and Sleeman [12], Wonham and coworkers [26, 27], etc. However, one important biological aspect of the hosts—the stage structure—seems to have received little attention, although structured population models have been intensively studied (see Diekmann and Heesterbeek [5]) in the context of population dynamics and spatial ecology, and the interaction of stage-structure with spatial dispersal has been receiving considerable attention in association with the theoretical development of the so-called reaction-diffusion equations with nonlocal delayed feedback (see the papers by Gourley, So, and Wu [7] and Gourley and Wu [8] and the references therein).

The developmental stages of hosts have a profound impact on the transmission dynamics of vector borne diseases. In the case of West Nile virus the transmission cycle involves both mosquitoes and birds, the crow species being particularly important. *Nestling* crows are crows that have hatched but are helpless and stay in the nest, receiving more-or-less continuous care from the mother for up to two weeks and less

*Received by the editors December 30, 2005; accepted for publication (in revised form) September 20, 2006; published electronically January 12, 2007. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada, by the Canada Research Chairs Program, by Public Health Agency of Canada, and by the Network of Centers of Excellence Program: Mathematics for Information Technology and Complex Systems.

http://www.siam.org/journals/siap/67-2/64871.html

†Department of Mathematics, University of Surrey, Guildford, Surrey, GU2 7XH, UK (s.gourley@surrey.ac.uk).

‡Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067 (rliu@math.purdue.edu).

§Laboratory for Industrial and Applied Mathematics, Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada, M3J 1P3 (wujh@mathstat.yorku.ca).

continuous care thereafter. *Fledgling* crows are old enough to have left the nest (they leave it after about five weeks), but they cannot fly very well. After three or four months these fledglings will be old enough to obtain all of their food by themselves. As these facts demonstrate, the maturation stages of adult birds, fledglings, and nestlings are all very different from a biological and an epidemiological perspective, and a realistic model needs to take these different stages into account. For example, in comparison with grown birds, the nestlings and fledglings have much higher disease-induced death rate, much poorer ability to avoid being bitten by mosquitoes, and much less spatial mobility [18, 2, 22]. In this paper we shall, in fact, assume that there is only one preadult stage for the host population, which in the West Nile virus context could be thought of as the fledgling stage of crows.

The aim of this paper is to formulate a model for the evolution of some vector borne diseases whose transmission dynamics and patterns are similar to those of West Nile virus. Other recent mathematical models for this disease include the works of Bowman et al. [3], Lewis, Renclawowicz, and van den Driessche [16], and Wonham et al. [26, 27], some of which use a different incidence function normalized by bird density. We start with the classical McKendrick von-Foerster equations for an age-structured reservoir population divided into two epidemiological compartments of susceptible and infected (and infectious), coupled with a scalar delay differential equation for the adult vector population under the assumption that the total vector population is maintained at a constant level. We then use the standard technique of integration along characteristics to reduce the model to a system of five coupled delay differential equations for the susceptible and infected juvenile and adult reservoir populations and the adult infected vector. If spatial diffusion is allowed, a similar derivation leads to a reaction-diffusion system with nonlocal and highly nonlinear delayed interactions. The model derivation is carried out in detail in sections 2 (for ODE models) and 3 (for PDE models), together with some detailed biological and epidemiological explanations of all terms involved.

We consider the qualitative behaviors of the reduced ordinary delay differential system in subsections 2.1–2.4. We establish the positiveness and boundedness of the reduced system, and we emphasize the need to restrict the initial data to the subset which is biologically and epidemiologically realistic. We then establish a concrete criterion, expressed in terms of the model parameters, for disease eradication. This is achieved using some comparison techniques and differential inequalities. We also obtain a necessary and sufficient condition for the disease-free equilibrium to be locally asymptotically stable—this is done using an application of the spectral properties of a linear delay differential system due to Smith [23]. The sharpness of the disease eradication condition is then tested using the available data and parameters for West Nile virus, and our simulations show that sustained oscillation can occur, should this disease eradication condition be violated.

In section 3, consider the issue of spatial spread of the disease in a one-dimensional setting. We provide a detailed formal calculation of the so-called minimal wave speed that is expected to coincide with the propagation speed of the disease, and we compare the predicted wave speed with data in the literature relating to the observed speed of spread of West Nile virus across North America. Finally, in section 4, we discuss our findings together with some of the corresponding results for a modified model with a different incidence function.

2. Model derivation. We shall think of the disease as mosquito borne, since mosquitoes are responsible for transmitting many of the vector borne diseases that

currently constitute significant public health issues in various parts of the world.

We will also refer to the reservoir as the host, and assume that the host population is age-structured. We start with a simple division of the host population into susceptible hosts $s(t,a)$ and infected hosts $i(t,a)$ at time t and age a. These host populations are assumed to evolve according to the McKendrick von-Foerster equations for an age-structured population:

$$\frac{\partial s}{\partial t} + \frac{\partial s}{\partial a} = -d_s(a)s(t,a) - \beta(a)s(t,a)m_i(t) \tag{2.1}$$

and

$$\frac{\partial i}{\partial t} + \frac{\partial i}{\partial a} = -d_i(a)i(t,a) + \beta(a)s(t,a)m_i(t), \tag{2.2}$$

where $m_i(t)$ is the number of infected adult mosquitoes satisfying another equation below, and $\beta(a)$ is the age-dependent transmission coefficient, and it is assumed that conversion of hosts from susceptible to infected occurs through interaction of susceptible hosts with infected mosquitoes, and at this stage we assume that the rate of conversion is given by mass action. We shall discuss the limitations of the model involving mass action and shall indicate how our work can be extended to include a more standard incidence term that includes dividing by the density of the host population. The functions $d_s(a)$ and $d_i(a)$ are the age-dependent death rates of susceptible and infected hosts.

We shall further split the host population into juveniles and adults, defined respectively as those of age less than some number τ and those of age greater than τ. We will work with the following choices for the death rates and the transmission coefficient $\beta(a)$:

$$d_s(a) = \begin{cases} d_{sj}, & a < \tau, \\ d_{sa}, & a > \tau, \end{cases} \qquad d_i(a) = \begin{cases} d_{ij}, & a < \tau, \\ d_{ia}, & a > \tau, \end{cases} \tag{2.3}$$

and

$$\beta(a) = \begin{cases} \beta_j, & a < \tau, \\ \beta_a, & a > \tau. \end{cases} \tag{2.4}$$

The subscripts in these quantities refer to disease and juvenile/adult status; thus, for example, the per capita death rates for susceptible juveniles and infected adults would be d_{sj} and d_{ia}, respectively. The above choices enable us to formulate a closed system of delay differential equations involving only the total numbers of hosts classified as adult susceptibles, adult infected, juvenile susceptibles, and juvenile infected. These total numbers are given respectively, using self-explanatory notations, by

$$\begin{aligned} A_s(t) &= \int_\tau^\infty s(t,a)\,da, \qquad A_i(t) = \int_\tau^\infty i(t,a)\,da, \qquad J_s(t) = \int_0^\tau s(t,a)\,da, \\ &\qquad\qquad J_i(t) = \int_0^\tau i(t,a)\,da. \end{aligned} \tag{2.5}$$

We assume no vertical transmission in the system (both from host and vector). On the further assumption that the birth rate is a function of the total number of susceptible adult hosts, we have the following expressions for the birth rates $s(t,0)$ and $i(t,0)$:

$$s(t,0) = b(A_s(t)), \qquad i(t,0) = 0, \tag{2.6}$$

where $b(\cdot)$ is the birth rate function for hosts (we shall later introduce $B(\cdot)$ as the birth rate function for mosquitoes). Equations (2.1) and (2.2) are solved subject to (2.6).

Let us now find a differential equation for $A_s(t)$. Differentiating the expression for $A_s(t)$ in (2.5), making use of (2.1), (2.3), and (2.4), and assuming (reasonably) that $s(t,\infty)=0$, we quickly find that

$$\frac{dA_s}{dt} = s(t,\tau) - d_{sa}A_s(t) - \beta_a m_i(t)A_s(t). \tag{2.7}$$

Next we need to find $s(t,\tau)$. This will be achieved by solving (2.1) for $0<a<\tau$. Set

$$s_\xi(a) = s(\xi+a,a).$$

Then

$$\begin{aligned}\frac{ds_\xi}{da} &= \left[\frac{\partial s}{\partial t} + \frac{\partial s}{\partial a}\right]_{t=\xi+a}\\ &= -s_\xi(a)[d_s(a) + \beta(a)m_i(\xi+a)],\end{aligned}$$

so that

$$\begin{aligned}s(\xi+a,a) = s_\xi(a) &= s_\xi(0)\exp\left(-\int_0^a (d_s(\eta)+\beta(\eta)m_i(\xi+\eta))\,d\eta\right)\\ &= b(A_s(\xi))\exp\left(-\int_0^a (d_s(\eta)+\beta(\eta)m_i(\xi+\eta))\,d\eta\right).\end{aligned} \tag{2.8}$$

Setting $a=\tau$ and $\xi = t-\tau$ and using (2.3), (2.4) gives

$$s(t,\tau) = b(A_s(t-\tau))\exp\left(-\int_0^\tau (d_{sj}+\beta_j m_i(t-\tau+\eta))\,d\eta\right).$$

Substituting this into (2.7) gives, after a change of variables in the integral,

$$\frac{dA_s}{dt} = b(A_s(t-\tau))\exp\left(-\int_{t-\tau}^t (d_{sj}+\beta_j m_i(u))\,du\right) - d_{sa}A_s(t) - \beta_a m_i(t)A_s(t). \tag{2.9}$$

In much the same way, we obtain the following equation for $J_s(t)$:

$$\begin{aligned}\frac{dJ_s}{dt} &= b(A_s(t)) - b(A_s(t-\tau))\exp\left(-\int_{t-\tau}^t (d_{sj}+\beta_j m_i(u))\,du\right)\\ &\quad - d_{sj}J_s(t) - \beta_j m_i(t)J_s(t).\end{aligned} \tag{2.10}$$

The differential equation for $A_i(t)$ turns out to be more complicated. Differentiating the expression for $A_i(t)$ in (2.5), assuming $i(t,\infty)=0$, and using (2.3) and (2.4) gives

$$\frac{dA_i(t)}{dt} = i(t,\tau) - d_{ia}A_i(t) + \beta_a m_i(t)A_s(t), \tag{2.11}$$

and we need to find $i(t,\tau)$, by solving (2.2) for $0<a<\tau$. Setting $i_\xi(a) = i(\xi+a,a)$ and differentiating with respect to a, we find from (2.2) that

$$\frac{di_\xi(a)}{da} + d_{ij}i_\xi(a) = \beta_j m_i(\xi+a)s(\xi+a,a).$$

Integrating this from 0 to a and recalling that $i_\xi(0) = i(\xi, 0) = 0$ by (2.6), we find that

$$i(\xi + a, a) = i_\xi(a) = \beta_j \int_0^a e^{-d_{ij}(a-\eta)} m_i(\xi + \eta) s(\xi + \eta, \eta)\, d\eta.$$

Therefore,

$$\begin{aligned} i(t, \tau) &= \beta_j \int_0^\tau e^{-d_{ij}(\tau-\eta)} m_i(t - \tau + \eta) s(t - \tau + \eta, \eta)\, d\eta \\ &= \beta_j \int_{t-\tau}^t e^{-d_{ij}(t-\xi)} m_i(\xi) s(\xi, \xi + \tau - t)\, d\xi. \end{aligned} \tag{2.12}$$

In this integral the second argument of $s(\xi, \xi + \tau - t)$ goes from 0 to τ, and therefore an expression for $s(\xi, \xi + \tau - t)$ can be obtained from the earlier analysis. From (2.8),

$$s(\xi, \xi + \tau - t) = b(A_s(t - \tau)) \exp\left(- \int_{t-\tau}^\xi [d_{sj} + \beta_j m_i(v)]\, dv \right).$$

Insertion of this expression into (2.12) yields an expression for $i(t, \tau)$ that involves only the state variables in (2.5) and $m_i(t)$, and insertion of this expression for $i(t, \tau)$ into (2.11) finally gives the differential equation for $A_i(t)$ to be

$$\begin{aligned} \frac{dA_i(t)}{dt} &= -d_{ia} A_i(t) + \beta_a m_i(t) A_s(t) \\ &\quad + \beta_j b(A_s(t - \tau)) \int_{t-\tau}^t m_i(\xi) e^{-d_{ij}(t-\xi)} \exp\left(- \int_{t-\tau}^\xi (d_{sj} + \beta_j m_i(v))\, dv \right) d\xi. \end{aligned} \tag{2.13}$$

Similarly, the differential equation for $J_i(t)$ can be shown to be

$$\begin{aligned} \frac{dJ_i(t)}{dt} &= -d_{ij} J_i(t) + \beta_j m_i(t) J_s(t) \\ &\quad - \beta_j b(A_s(t - \tau)) \int_{t-\tau}^t m_i(\xi) e^{-d_{ij}(t-\xi)} \exp\left(- \int_{t-\tau}^\xi (d_{sj} + \beta_j m_i(v))\, dv \right) d\xi. \end{aligned} \tag{2.14}$$

To close the system we still need a differential equation for the variable $m_i(t)$, but first we would like to discuss the ecological interpretation of the complicated integral term appearing in (2.13) and (2.14). The first two terms in the right-hand side of (2.13) are easy to interpret. They are, respectively, the death rate of infected adults and conversion of susceptible adults to infected adults via contact with infected mosquitoes. The last term in (2.13) tells us the rate at which infected immatures become infected adults having contracted the disease in childhood. This term is the rate at which infected individuals pass through age τ. Now, an individual that is of age τ at time t will have been born at time $t - \tau$. Recall, however, that all individuals are born as susceptibles. This is why the birth rate $b(A_s(t-\tau))$ is involved. The individuals we are presently discussing have each acquired the infection at some stage during childhood, so assume that a particular individual acquires it at a time $\xi \in (t - \tau, t)$. This particular individual remained susceptible from its birth at time $t - \tau$ until time ξ, and the probability of this happening is

$$\exp\left(- \int_{t-\tau}^\xi (d_{sj} + \beta_j m_i(v))\, dv \right).$$

The probability that the individual will survive from becoming infected at time ξ until becoming an adult at time t is

$$e^{-d_{ij}(t-\xi)}.$$

These two exponentials both feature in the last term in (2.13). The product $\beta_j m_i(\xi)$ is the per capita conversion rate of susceptible juveniles to infected juveniles at time ξ, and ξ running from $t-\tau$ to t totals up the contributions from all possible times at which infected individuals passing into adulthood might have acquired the infection.

Finally, we need differential equations for the mosquitoes. Let $m_T(t)$ be the total number of (adult) mosquitoes, divided into infected mosquitoes $m_i(t)$ and susceptible mosquitoes $m_T(t) - m_i(t)$. Death and reproductive activity for mosquitoes are assumed not to depend on whether they are carrying the disease or not, and so the total number of adult mosquitoes is assumed to obey

$$\frac{dm_T(t)}{dt} = e^{-d_l\sigma}B(m_T(t-\sigma)) - d_m m_T(t), \tag{2.15}$$

where d_l and d_m denote the death rates of larval and adult mosquitoes, respectively, and σ is the length of the larval phase from egg to adult. The function $B(\cdot)$ is the birth rate function for mosquitoes. It is possible but unnecessary to write down a differential equation for larval mosquitoes. Infected adult mosquitoes $m_i(t)$ are assumed to obey

$$\frac{dm_i(t)}{dt} = -d_m m_i(t) + \beta_m(m_T(t) - m_i(t))(J_i(t) + \alpha A_i(t)). \tag{2.16}$$

Thus, the rate at which mosquitoes become infected is given by mass action as the product of susceptible mosquitoes $m_T(t) - m_i(t)$ and infected birds which may be either juvenile or adult. The presence of the factor α is to account for the possibility that juvenile and adult birds might not be equally vulnerable to being bitten. Again, we defer the discussion of a more standard incidence term to the final section.

Certain assumptions will be made concerning the birth function $B(\cdot)$ for the mosquitoes. These assumptions, which are ecologically reasonable, are geared towards ensuring that the total number $m_T(t)$ of mosquitoes stabilizes and does not tend to zero (otherwise the disease is automatically eradicated and the model is not interesting). These assumptions are

$$\left.\begin{array}{l} B(0) = 0,\ B(\cdot) \text{ is strictly monotonically increasing, there exists } m_T^* > 0 \\ \text{such that } e^{-d_l\sigma}B(m) > d_m m \text{ when } m < m_T^* \text{ and } e^{-d_l\sigma}B(m) < d_m m \text{ when} \\ m > m_T^*. \end{array}\right\} \tag{2.17}$$

The quantity $m_T^* > 0$ in (2.17) is an equilibrium of (2.15), and $m_T(t) \to m_T^*$ as $t \to \infty$, provided $m_T(\theta) \geq 0$ and $m_T(\theta) \not\equiv 0$ on $\theta \in [-\sigma, 0]$ (see Kuang [14]). Accordingly, (2.16) is asymptotically autonomous, and we may replace $m_T(t)$ by m_T^* in (2.16), thereby lowering the order of the system to be studied, which we now note consists of (2.9), (2.10), (2.13), and (2.14) together with

$$\frac{dm_i(t)}{dt} = -d_m m_i(t) + \beta_m(m_T^* - m_i(t))(J_i(t) + \alpha A_i(t)), \tag{2.18}$$

which is the asymptotically autonomous limiting form of (2.16). Note that this system does not explicitly involve the delay σ, but this delay is still involved via the quantity m_T^*.

2.1. Positivity of solutions. We will prove that the system consisting of (2.9), (2.10), (2.13), (2.14), and (2.18) has a positivity preserving property. It is easy to appreciate that this system cannot have a positivity preserving property for completely arbitrary nonnegative initial data (a glance at the terms in the right-hand side of (2.14) makes this clear). However, positivity preservation does hold when some components of the initial data satisfy certain relations. These relations are easily seen to be the only ones that make sense ecologically and therefore are easily admitted. We therefore now append to the above-mentioned system the following initial data:

$$
\begin{aligned}
&A_s(\theta) = A_s^0(\theta) \geq 0, \qquad \theta \in [-\tau, 0],\\
&m_i(\theta) = m_i^0(\theta) \in [0, m_T^*], \qquad \theta \in [-\tau, 0],\\
&A_i(0) = A_i^0(0) \geq 0,\\
&J_s(0) = \int_{-\tau}^{0} b(A_s^0(\xi)) \exp\left(- \int_{\xi}^{0} [d_{sj} + \beta_j m_i^0(u)]\, du \right) d\xi,\\
&J_i(0) = \int_{-\tau}^{0} b(A_s^0(\xi)) \left\{ \int_{\xi}^{0} \beta_j m_i^0(\eta) e^{d_{ij}\eta} e^{-\int_{\xi}^{\eta} [d_{sj} + \beta_j m_i^0(v)]\, dv} d\eta \right\} d\xi,
\end{aligned}
\tag{2.19}
$$

where $A_s^0(\theta)$ and $m_i^0(\theta)$ are prescribed continuous functions of the variable $\theta \in [-\tau, 0]$, and $A_i^0(0)$ is also a given value. Note that $J_s(0)$ and $J_i(0)$ have to be calculated from the initial data for A_s and m_i. This is ecologically reasonable; after all, $J_s(0)$ is the number of juvenile susceptibles at time $t = 0$. The integral on the right in the expression for $J_s(0)$ is simply accounting for all these juvenile susceptibles at $t = 0$. Each one was born at some time $\xi \in [-\tau, 0]$—hence the presence of the birth rate $b(A_s^0(\xi))$—and each has to have survived and remained susceptible until time 0, hence the exponential term which represents the probability of this actually happening. The interpretation of the expression for $J_i(0)$ is similar but more complicated. Of the infected juveniles $J_i(0)$ at time 0, each one was born at some time $\xi \in [-\tau, 0]$ as a susceptible, and each of these newborns at time ξ then became infected at some subsequent time $\eta \in [\xi, 0]$.

We will now prove the following positivity preservation result.

PROPOSITION 2.1. *Let* (2.17) *hold. Then each component of the solution of the system consisting of* (2.9), (2.10), (2.13), (2.14), *and* (2.18) *for* $t > 0$, *subject to the initial conditions* (2.19), *remains nonnegative for all* $t > 0$. *Also,* $m_i(t) \leq m_T^*$ *for all* $t > 0$. *If, furthermore, the function* b *is bounded, then each component of the above solution is also bounded for all* $t > 0$.

Proof. First we will show that $m_i(t) \leq m_T^*$ for all $t > 0$. Suppose the contrary; then there must exist a time t_1 such that $m_i(t_1) = m_T^*$ and $dm_i(t_1)/dt \geq 0$. Evaluating (2.18) at time t_1 immediately gives a contradiction.

Next we prove nonnegativity of $A_s(t)$, for $t \in (0, \tau]$ in the first instance. On this interval,

$$
\frac{dA_s(t)}{dt} \geq -d_{sa} A_s(t) - \beta_a m_i(t) A_s(t).
$$

By comparison, $A_s(t)$ is bounded below by the solution of the corresponding differential equation obtained by replacing $\geq$ by $=$, and this differential equation contains a factor of $A_s(t)$ in its right-hand side. Since $A_s(0) \geq 0$, it follows that $A_s(t) \geq 0$ for all $t \in (0, \tau]$. This argument can be continued using the method of steps, and we conclude that $A_s(t) \geq 0$ for all $t > 0$.

Nonnegativity of $J_s(t)$ will be shown next. This can be seen by noting that the solution of (2.10), subject to the initial value for $J_s(0)$ given in (2.19), is

$$J_s(t) = \int_{t-\tau}^{t} b(A_s(\xi)) \exp\left(-\int_{\xi}^{t} [d_{sj} + \beta_j m_i(u)]\, du\right) d\xi, \tag{2.20}$$

which is nonnegative because A_s is nonnegative.

We still have to prove the nonnegativity of $A_i(t)$, $J_i(t)$, and $m_i(t)$. It will be helpful to note that the solution of (2.14), subject to the initial value for $J_i(0)$ given in (2.19), is

$$J_i(t) = \int_{t-\tau}^{t} b(A_s(\xi)) \left\{ \int_{\xi}^{t} \beta_j m_i(\eta) e^{-d_{ij}(t-\eta)} e^{-\int_{\xi}^{\eta} [d_{sj} + \beta_j m_i(v)]\, dv} d\eta \right\} d\xi, \tag{2.21}$$

which is nonnegative if $m_i(t)$ is nonnegative. Therefore, it suffices to prove nonnegativity of $A_i(t)$ and $m_i(t)$. These two functions can be viewed as the solution $(A_i(t), m_i(t))$ of the system of differential equations consisting of (2.13) and

$$\begin{aligned} \frac{dm_i(t)}{dt} &= -d_m m_i(t) + \beta_m (m_T^* - m_i(t)) \\ &\times \left(\int_{t-\tau}^{t} b(A_s(\xi)) \left\{ \int_{\xi}^{t} \beta_j m_i(\eta) e^{-d_{ij}(t-\eta)} e^{-\int_{\xi}^{\eta} [d_{sj} + \beta_j m_i(v)]\, dv} d\eta \right\} d\xi + \alpha A_i(t) \right) \end{aligned} \tag{2.22}$$

for $t > 0$, with initial data taken from (2.19), but with $A_s(t)$ thought of simply as some prescribed nonnegative function. Recalling that $m_i(t) \le m_T^*$, we now note that, even though this system does not satisfy a quasi monotonicity condition, Theorem 2.1 of Smith [23, p. 81] is applicable and gives us the nonnegativity of $A_i(t)$ and $m_i(t)$ immediately. The proof of the nonnegativity of each component of the solution is then complete.

The boundedness of $A_s(t)$ is simple since, by (2.9),

$$\frac{d}{dt} A_s(t) \le b_{\sup} - d_{sa} A_s(t) - \beta_a m_i(t) A_s(t),$$

where $b_{\sup} = \sup_{A \ge 0} b(A) < \infty$. The boundedness of $A_i(t)$ follows from (2.13) and the boundedness of $m_i(t)$. The boundedness of $J_s(t)$ and $J_i(t)$ follows from (2.20) and (2.21) directly. This completes the proof.

2.2. Global convergence to disease-free state. In this section we shall prove a theorem giving sufficient conditions for the system to evolve to the disease-free state (i.e., conditions that ensure A_i, J_i, and m_i go to zero as $t \to \infty$). Since the differential equations (2.10) and (2.14) can be solved to give (2.20) and (2.21), respectively, it is sufficient to study the system consisting of (2.9), (2.13), and (2.22), with initial data taken from (2.19). These equations form a closed system for $A_s(t)$, $A_i(t)$, and $m_i(t)$. Our aim will be to establish, using these three equations, a differential inequality for the variable $m_i(t)$ only, and to use this to find conditions which ensure that $m_i(t) \to 0$ as $t \to \infty$. Note that if $m_i(t) \to 0$, then from (2.21) it follows immediately that $J_i(t) \to 0$ and, furthermore, (2.13) then becomes an asymptotically autonomous ODE, from which it is easily seen that $A_i(t)$ tends to zero.

We will make certain assumptions concerning the birth rate function $b(\cdot)$ for hosts. These assumptions are

$$\left.\begin{array}{l} b(0)=0,\ b(A)>0 \text{ when } A>0,\ b_{\sup}:=\sup_{A\geq 0} b(A)<\infty, \text{ there exists} \\ A_s^*>0 \text{ such that } e^{-d_{sj}\tau}b(A)>d_{sa}A \text{ when } A<A_s^* \text{ and } e^{-d_{sj}\tau}b(A)< \\ d_{sa}A \text{ when } A>A_s^*. \end{array}\right\} \tag{2.23}$$

These assumptions are not the same as those for the birth rate function $B(\cdot)$ for mosquitoes (assumptions (2.17)); note in particular that we do not require $b(\cdot)$ to be monotone.

The reader will realize that the quantity A_s^* in (2.23) is, in fact, a nonzero equilibrium value for $A_s(t)$ in the case when the disease is absent. Assumptions (2.23) are geared towards ensuring that the population $A_s(t)$ of adult susceptible hosts does not go to zero even without the disease; otherwise the model is not interesting. This is important because if $e^{-d_{sj}\tau}b(A)<d_{sa}A$ for all $A>0$ (which means that, in the absence of the disease, adult recruitment of susceptible hosts is insufficient to offset natural death of adult susceptible hosts), then it is natural to expect that $A_s(t)\to 0$ even without the disease, and this can be mathematically shown to be the case, using (2.9).

We will prove the following theorem. Assumption (2.17) is needed to ensure the existence of m_T^*. We shall need the functions a_1 and a_0 defined by

$$\begin{aligned} a_1(\epsilon) &= d_m d_{ia}+d_m d_{ij}+d_{ia}d_{ij} \\ &\quad -\frac{\beta_m m_T^* b_{\sup}\beta_j}{d_{sj}}-\beta_m m_T^*\alpha\beta_a\left(\frac{b_{\sup}e^{-d_{sj}\tau}}{d_{sa}}+\epsilon\right) \\ &\quad -e^{-d_{sj}\tau}\left(\frac{1-e^{-\tau(d_{ij}-d_m-d_{sj})}}{d_{ij}-d_m-d_{sj}}\right)\beta_m m_T^*\alpha\beta_j b_{\sup} \end{aligned} \tag{2.24}$$

and

$$\begin{aligned} a_0(\epsilon) &= d_m d_{ia}d_{ij}-\frac{d_{ia}\beta_m m_T^* b_{\sup}\beta_j}{d_{sj}}-d_{ij}\beta_m m_T^*\alpha\beta_a\left(\frac{b_{\sup}e^{-d_{sj}\tau}}{d_{sa}}+\epsilon\right) \\ &\quad -d_{ij}e^{-d_{sj}\tau}\left(\frac{1-e^{-\tau(d_{ij}-d_m-d_{sj})}}{d_{ij}-d_m-d_{sj}}\right)\beta_m m_T^*\alpha\beta_j b_{\sup}. \end{aligned} \tag{2.25}$$

THEOREM 2.2. *Let* (2.17) *and* (2.23) *hold, and let* $A_s(t)$, $A_i(t)$, *and* $m_i(t)$ *satisfy* (2.9), (2.13), *and* (2.22), *with initial data taken from* (2.19). *Assume further that*

$$a_1(0)>0, \quad a_0(0)>0, \quad \textit{and} \quad (d_m+d_{ia}+d_{ij})a_1(0)>a_0(0), \tag{2.26}$$

where the functions a_1, a_0 *are defined by* (2.24) *and* (2.25). *Then* $(A_i(t),m_i(t))\to(0,0)$ *as* $t\to\infty$.

Remark. It is not hard to check that (2.26) can be satisfied for some parameter values. It is satisfied, for example, when the contact rates β_a, β_j, and β_m are sufficiently small, or when the mosquito capacity m_T^* is sufficiently small. These are situations in which we intuitively expect the theorem to hold. As such, an obvious control measure for achieving disease eradication is to reduce the mosquito capacity. Reducing β_m is an alternative approach.

Proof of Theorem 2.2. For the reasons explained above, we may concentrate on showing that $m_i(t)\to 0$ as $t\to\infty$. From positivity of solutions, we find from (2.9)

that

$$\begin{aligned}\frac{dA_s}{dt} &\le b(A_s(t-\tau))e^{-d_{sj}\tau} - d_{sa}A_s(t)\\ &\le b_{\sup}e^{-d_{sj}\tau} - d_{sa}A_s(t).\end{aligned}$$

Hence

$$\limsup_{t\to\infty} A_s(t) \le \frac{b_{\sup}e^{-d_{sj}\tau}}{d_{sa}}.$$

By hypothesis (2.26) and by a continuity argument we may choose $\epsilon > 0$ sufficiently small that

$$a_1(\epsilon) > 0, \quad a_0(\epsilon) > 0, \quad \text{and} \quad (d_m + d_{ia} + d_{ij})a_1(\epsilon) > a_0(\epsilon). \tag{2.27}$$

There exists $T_1 > 0$ such that, for $t \ge T_1$,

$$A_s(t) \le \frac{b_{\sup}e^{-d_{sj}\tau}}{d_{sa}} + \epsilon.$$

Using this estimate in (2.13), we find that, for $t \ge T_1$,

$$\begin{aligned}\frac{dA_i(t)}{dt} &\le -d_{ia}A_i(t) + \beta_a m_i(t)\left(\frac{b_{\sup}e^{-d_{sj}\tau}}{d_{sa}} + \epsilon\right)\\ &+ \beta_j b_{\sup}\int_{t-\tau}^{t} m_i(\xi)e^{-d_{ij}(t-\xi)}\exp\left(-\int_{t-\tau}^{\xi}(d_{sj} + \beta_j m_i(v))\,dv\right)d\xi.\end{aligned} \tag{2.28}$$

Solving this differential inequality and ignoring a transient term involving $A_i(0)$, we find that

$$\begin{aligned}A_i(t) &\le \beta_a\left(\frac{b_{\sup}e^{-d_{sj}\tau}}{d_{sa}} + \epsilon\right)\int_0^t e^{-d_{ia}(t-\psi)}m_i(\psi)\,d\psi\\ &+ \beta_j b_{\sup}\int_0^t e^{-d_{ia}(t-\psi)}\int_{\psi-\tau}^{\psi} m_i(\xi)e^{-d_{ij}(\psi-\xi)}\exp\left(-\int_{\psi-\tau}^{\xi}(d_{sj} + \beta_j m_i(v))\,dv\right)d\xi\,d\psi.\end{aligned} \tag{2.29}$$

We shall use this estimate for $A_i(t)$ to obtain a differential inequality for $m_i(t)$ as follows. From (2.22), and using positivity of $m_i(t)$ and the bound on $b(\cdot)$,

$$\begin{aligned}\frac{dm_i(t)}{dt} &\le -d_m m_i(t) + \beta_m m_T^*\\ &\times\left(b_{\sup}\int_{t-\tau}^{t}\int_{\xi}^{t}\beta_j m_i(\eta)e^{-d_{ij}(t-\eta)}e^{-\int_{\xi}^{\eta}[d_{sj}+\beta_j m_i(v)]\,dv}d\eta\,d\xi + \alpha A_i(t)\right),\end{aligned}$$

so that, from (2.29),

$$\begin{aligned}
\frac{dm_i(t)}{dt} &\le -d_m m_i(t) \\
&+ \beta_m m_T^* b_{\sup} \int_{t-\tau}^{t} \int_{\xi}^{t} \beta_j m_i(\eta) e^{-d_{ij}(t-\eta)} e^{-\int_{\xi}^{\eta} [d_{sj}+\beta_j m_i(v)]\, dv} d\eta\, d\xi \\
&+ \beta_m m_T^* \alpha \beta_a \left(\frac{b_{\sup} e^{-d_{sj}\tau}}{d_{sa}} + \epsilon \right) \int_0^t e^{-d_{ia}(t-\psi)} m_i(\psi)\, d\psi \\
&+ \beta_m m_T^* \alpha \beta_j b_{\sup} \int_0^t e^{-d_{ia}(t-\psi)} \int_{\psi-\tau}^{\psi} m_i(\xi) e^{-d_{ij}(\psi-\xi)} \\
&\qquad \times \exp\left(- \int_{\psi-\tau}^{\xi} (d_{sj} + \beta_j m_i(v))\, dv \right) d\xi\, d\psi.
\end{aligned}$$

From this it is easy to see, using the positivity of $m_i(t)$, that $m_i(t)$ also obeys the following simpler linear differential inequality:

$$\begin{aligned}
\frac{dm_i(t)}{dt} &\le -d_m m_i(t) \\
&+ \beta_m m_T^* b_{\sup} \int_{t-\tau}^{t} \int_{\xi}^{t} \beta_j m_i(\eta) e^{-d_{ij}(t-\eta)} e^{-d_{sj}(\eta-\xi)} d\eta\, d\xi \\
&+ \beta_m m_T^* \alpha \beta_a \left(\frac{b_{\sup} e^{-d_{sj}\tau}}{d_{sa}} + \epsilon \right) \int_0^t e^{-d_{ia}(t-\psi)} m_i(\psi)\, d\psi \\
&+ \beta_m m_T^* \alpha \beta_j b_{\sup} \int_0^t e^{-d_{ia}(t-\psi)} \int_{\psi-\tau}^{\psi} m_i(\xi) e^{-d_{ij}(\psi-\xi)} e^{-d_{sj}(\xi-\psi+\tau)} d\xi\, d\psi.
\end{aligned} \tag{2.30}$$

To make progress we need to estimate some of these integrals. If we change the order of integration in the first double integral of (2.30), we reach the following estimate:

$$\begin{aligned}
&\int_{t-\tau}^{t} \int_{\xi}^{t} \beta_j m_i(\eta) e^{-d_{ij}(t-\eta)} e^{-d_{sj}(\eta-\xi)} d\eta\, d\xi \\
&= \int_{t-\tau}^{t} \int_{t-\tau}^{\eta} \beta_j m_i(\eta) e^{-d_{ij}(t-\eta)} e^{-d_{sj}(\eta-\xi)} d\xi\, d\eta \\
&\le \frac{\beta_j}{d_{sj}} \int_{t-\tau}^{t} m_i(\eta) e^{-d_{ij}(t-\eta)} d\eta \\
&\le \frac{\beta_j}{d_{sj}} \int_{0}^{t} m_i(\eta) e^{-d_{ij}(t-\eta)} d\eta,
\end{aligned} \tag{2.31}$$

assuming $t > \tau$.

From (2.18) and Proposition 2.1 we have

$$\frac{dm_i(t)}{dt} \ge -d_m m_i(t).$$

Integrating from ξ to ψ gives

$$m_i(\xi) \le m_i(\psi) e^{d_m(\psi-\xi)}, \qquad \xi \le \psi.$$

Using this and (2.31), we obtain

$$\begin{aligned}\frac{dm_i(t)}{dt} &\le -d_m m_i(t) + \frac{\beta_m m_T^* b_{\sup}\beta_j}{d_{sj}} \int_0^t m_i(\eta) e^{-d_{ij}(t-\eta)}\, d\eta \\ &+ \beta_m m_T^* \alpha\beta_a \left(\frac{b_{\sup} e^{-d_{sj}\tau}}{d_{sa}} + \epsilon\right) \int_0^t e^{-d_{ia}(t-\psi)} m_i(\psi)\, d\psi \\ &+ \beta_m m_T^* \alpha\beta_j b_{\sup} e^{-d_{sj}\tau} \left(\frac{1 - e^{-\tau(d_{ij}-d_m-d_{sj})}}{d_{ij}-d_m-d_{sj}}\right) \int_0^t e^{-d_{ia}(t-\psi)} m_i(\psi)\, d\psi.\end{aligned} \tag{2.32}$$

By the theory of monotone systems [23], $m_i(t) \le M_i(t)$, where $M_i(t)$ is the solution of the differential equation obtained from (2.32) by replacing $\le$ by $=$, subject to the same initial data as that for m_i. Applying to this differential equation the Laplace transform, letting p be the transform variable and $\bar{M}_i(p)$ denote the Laplace transform of $M_i(t)$, we find after some algebra that

$$\bar{M}_i(p)\,\Lambda(p) = m_i(0)(p + d_{ia})(p + d_{ij}), \tag{2.33}$$

where

$$\Lambda(p) = p^3 + (d_m + d_{ia} + d_{ij})p^2 + a_1(\epsilon)p + a_0(\epsilon) \tag{2.34}$$

with $a_1(\epsilon)$ and $a_0(\epsilon)$ given by (2.24) and (2.25). Recall that the small number $\epsilon > 0$ has been chosen such that (2.27) holds. This fact, together with the Routh Hurwitz criteria, implies that all the roots of the cubic equation $\Lambda(p) = 0$ satisfy $\operatorname{Re} p < 0$, and so the same is true of all singularities of $\bar{M}_i(p)$. By the inversion formula for Laplace transforms, $M_i(t) \to 0$ as $t \to \infty$. Since $0 \le m_i(t) \le M_i(t)$, $m_i(t) \to 0$ as $t \to \infty$. By (2.13), $A_i(t) \to 0$ as $t \to \infty$. The proof of Theorem 2.2 is complete.

2.3. Local stability of disease-free equilibrium. If (2.23) holds, then the model (2.9), (2.10), (2.13), (2.14), and (2.18) has a disease-free equilibrium (DFE), obtained by substituting $J_i = 0$, $A_i = 0$, and $m_i = 0$ into the right-hand sides of those equations and setting them to zero, given by

$$E_0 = (A_s^*, J_s^*, 0, 0, 0), \tag{2.35}$$

where $A_s^* > 0$ and $J_s^* > 0$ are given by

$$\begin{cases} b(A_s^*)e^{-d_{sj}\tau} - d_{sa}A_s^* = 0, \\ J_s^* = \dfrac{b(A_s^*)}{d_{sj}}(1 - e^{-d_{sj}\tau}). \end{cases} \tag{2.36}$$

The previous section of this paper presented sufficient conditions for disease eradication (Theorem 2.2). In this section we investigate the linear stability of the DFE E_0 to gain further insight, and we shall present a condition (namely, condition (2.38) below) which is both necessary and sufficient for E_0 to be linearly stable. Though we do not establish disease eradication *globally* under this particular condition, it is clearly the weakest possible condition for disease eradication.

We first require the following simple preliminary result, which provides a condition for the linear stability of the DFE E_0 to perturbations in which the disease remains absent.

LEMMA 2.3. *Let* (2.23) *hold. Then* (A_s^*, J_s^*)*, given by* (2.36)*, is a locally asymptotically stable equilibrium of the subsystem*

$$\begin{cases} \dfrac{d\bar{J}_s(t)}{dt} = b(\bar{A}_s(t)) - b(\bar{A}_s(t-\tau))e^{-d_{sj}\tau} - d_{sj}\bar{J}_s(t), \\ \dfrac{d\bar{A}_s(t)}{dt} = b(\bar{A}_s(t-\tau))e^{-d_{sj}\tau} - d_{sa}\bar{A}_s(t) \end{cases} \tag{2.37}$$

if $d_{sa} > |b'(A_s^*)|e^{-d_{sj}\tau}$.

Proof. Obviously, (A_s^*, J_s^*) is an equilibrium of system (2.37). The linearization of (2.37) at this equilibrium has solutions of the form $\exp(\lambda t)$ whenever λ satisfies

$$\begin{vmatrix} -\lambda - d_{sj} & b'(A_s^*)(1 - e^{-(\lambda+d_{sj})\tau}) \\ 0 & -\lambda - d_{sa} + b'(A_s^*)e^{-(\lambda+d_{sj})\tau} \end{vmatrix} = 0.$$

Therefore, (A_s^*, J_s^*) is a locally stable solution of (2.37) if and only if all the roots λ of $-\lambda - d_{sa} + b'(A_s^*)e^{-(\lambda+d_{sj})\tau} = 0$ have negative real part. It is straightforward to show that this is the case if $d_{sa} > |b'(A_s^*)|e^{-d_{sj}\tau}$. The proof is complete.

Our main result of this section is the following theorem, which gives a necessary and sufficient condition for the linear stability of the disease-free state.

THEOREM 2.4. *Let* (2.17) *and the hypotheses of Lemma* 2.3 *hold, and assume additionally that*

$$\begin{aligned} d_m > \beta_m m_T^* \Bigg\{ & \frac{b(A_s^*)\beta_j}{d_{ij} - d_{sj}} \left[\frac{1 - e^{-d_{sj}\tau}}{d_{sj}} - \frac{(1 - e^{-d_{ij}\tau})}{d_{ij}} \right] \\ & + \frac{\alpha}{d_{ia}} \left[\beta_a A_s^* + \beta_j b(A_s^*) e^{-d_{sj}\tau} \frac{(1 - e^{-(d_{ij}-d_{sj})\tau})}{d_{ij} - d_{sj}} \right] \Bigg\}. \end{aligned} \tag{2.38}$$

Then the disease-free equilibrium E_0 *given by* (2.35) *is linearly asymptotically stable as a solution of the full model* (2.9), (2.10), (2.13), (2.14), (2.18).

Remark. The hypotheses of Theorem 2.4 are the weakest possible hypotheses that can guarantee the stated result. Recall from earlier remarks that if (2.17) or (2.23) is violated, then the mosquito or host population is doomed, irrespective of the disease. If the two sides of (2.38) are equal, then zero is an eigenvalue of the characteristic equation of the linearization about E_0 ((2.40) below), signaling the bifurcation of an endemic equilibrium. As will be shown numerically at the end of this section, a Hopf bifurcation of periodic solutions may further bifurcate from this endemic equilibrium. It remains a challenging problem to determine whether the hypotheses of Theorem 2.4 are sufficient to guarantee the global stability of E_0.

Proof. We aim for a linear equation in m_i only. Making use of the expression (2.21) for $J_i(t)$, solving for $A_i(t)$ the differential equation (2.13) on the interval $(-\infty, t)$, and then linearizing about $m_i = 0$, we obtain

$$\begin{aligned} \frac{dm_i(t)}{dt} &= -d_m m_i(t) \\ &+ \beta_m m_T^* b(A_s^*) \int_{t-\tau}^{t} \int_{\xi}^{t} \beta_j m_i(\eta) e^{-d_{ij}(t-\eta)} e^{-d_{sj}(\eta-\xi)} d\eta \, d\xi \\ &+ \beta_m m_T^* \alpha \beta_a A_s^* \int_{-\infty}^{t} e^{-d_{ia}(t-\psi)} m_i(\psi) \, d\psi \\ &+ \beta_m m_T^* \alpha \beta_j b(A_s^*) \int_{-\infty}^{t} e^{-d_{ia}(t-\psi)} \int_{\psi-\tau}^{\psi} m_i(\xi) e^{-d_{ij}(\psi-\xi)} e^{-d_{sj}(\xi-\psi+\tau)} d\xi \, d\psi. \end{aligned} \tag{2.39}$$

Solutions of the form $m_i(t) = e^{\lambda t}$ exist whenever λ satisfies

$$\lambda + d_m = \beta_m m_T^* \left\{ \frac{b(A_s^*)\beta_j}{\lambda + d_{ij} - d_{sj}} \left[\frac{1 - e^{-d_{sj}\tau}}{d_{sj}} - \frac{(1 - e^{-(\lambda + d_{ij})\tau})}{\lambda + d_{ij}} \right] + \frac{\alpha}{\lambda + d_{ia}} \left[\beta_a A_s^* + \beta_j b(A_s^*) e^{-d_{sj}\tau} \frac{(1 - e^{-(\lambda + d_{ij} - d_{sj})\tau})}{\lambda + d_{ij} - d_{sj}} \right] \right\}. \tag{2.40}$$

The structure of the linear equation (2.39) is such that the linear stability of its zero solution can be determined by considering only the real roots of the characteristic equation (2.40). This follows from Theorem 5.1 of Smith [23, p. 92] and Theorem 3.2 of Wu [28]. Our aim is therefore to show that, under condition (2.38), equation (2.40) does not have any nonnegative real roots. From simple graphical arguments, we see that it is sufficient to show that the right-hand side of (2.40) is monotonically decreasing as a function of $\lambda \in \mathbf{R}$ for $\lambda \geq 0$.

Let $F(\lambda)$ denote the right-hand side of (2.40), excluding the $\beta_m m_T^*$ factor. It is sufficient to show that $F'(\lambda) < 0$ for all $\lambda \geq 0$. Now

$$\begin{aligned} F(\lambda) &= \frac{\tau b(A_s^*)\beta_j}{\lambda + d_{ij} - d_{sj}} \left[f(d_{sj}\tau) - f((\lambda + d_{ij})\tau) \right] \\ &\quad + \frac{\alpha}{\lambda + d_{ia}} \left[\beta_a A_s^* + \tau \beta_j b(A_s^*) e^{-d_{sj}\tau} f((\lambda + d_{ij} - d_{sj})\tau) \right] \\ &=: F_1(\lambda) + \alpha F_2(\lambda), \end{aligned} \tag{2.41}$$

in which the function f is defined by

$$f(x) = \frac{1 - e^{-x}}{x}.$$

It is reasonably straightforward to see that f satisfies

$$f(x) > 0, \qquad f'(x) < 0, \qquad f''(x) > 0 \qquad \text{for all } x \in \mathbf{R}. \tag{2.42}$$

Indeed, (2.42) follows from the following inequalities:

$$(x + 1)e^{-x} \leq 1, \qquad x \in \mathbf{R},$$

and

$$\begin{aligned} e^{-x}(x^2 + 2x + 2) &< 2, \qquad x > 0, \\ e^{-x}(x^2 + 2x + 2) &> 2, \qquad x < 0. \end{aligned}$$

It is sufficient to show that $F_1'(\lambda) < 0$ and $F_2'(\lambda) < 0$ for all $\lambda \geq 0$, with the $F_i(\lambda)$ defined by (2.41). It is very easily seen, using (2.42), that $F_2'(\lambda) < 0$ for all $\lambda \geq 0$ (in fact for all $\lambda > -d_{ia}$). To show that $F_1'(\lambda) < 0$, introduce $\xi = \lambda + d_{ij} - d_{sj}$ and the function $g(\xi)$ defined by

$$g(\xi) = \frac{1}{\xi}(f(d_{sj}\tau) - f((\xi + d_{sj})\tau));$$

then it is more than sufficient to show that $g'(\xi) < 0$ for all $\xi \in \mathbf{R}$. However,

$$\begin{aligned} g'(\xi) &= \frac{1}{\xi^2}[f((\xi + d_{sj})\tau) - f(d_{sj}\tau)] - \frac{\tau}{\xi} f'((\xi + d_{sj})\tau) \\ &= \frac{\tau}{\xi} \left[f'((\theta\xi + d_{sj})\tau) - f'((\xi + d_{sj})\tau) \right] \\ &= (\theta - 1)\tau^2 f''(c) \end{aligned}$$

for some numbers $\theta \in (0,1)$ and $c \in \mathbf{R}$ which arise from applications of the mean value theorem. Since $f''(c) > 0$ by (2.42) it follows that $g'(\xi) < 0$, as desired. Thus, (2.40) does not have any nonnegative real roots.

With $m_i(t) \to 0$ it follows from (2.21) and (2.13) that $J_i(t) \to 0$ and $A_i(t) \to 0$. Then the hypotheses of Lemma 2.3, which are embodied within those of Theorem 2.4, imply that $A_s(t) \to A_s^*$ and $J_s(t) \to J_s^*$ in the linearized equations. The proof of Theorem 2.4 is complete.

2.4. Numerical simulations. Let us introduce the new variable W_1 defined by

$$W_1(t) = \int_{t-\tau}^{t} m_i(\xi) e^{-d_{ij}(t-\xi)} \exp\left(- \int_{t-\tau}^{\xi} (d_{sj} + \beta_j m_i(v))\, dv \right) d\xi,$$

so that we can rewrite the model (2.9), (2.10), (2.13), (2.14), and (2.18) in the form

(2.43)

$$\begin{aligned}
\frac{dJ_s(t)}{dt} &= b(A_s(t)) - b(A_s(t-\tau)) e^{-d_{sj}\tau} e^{-\int_{t-\tau}^{t} \beta_j m_i(v) dv} - d_{sj} J_s(t) - \beta_j m_i(t) J_s(t), \\
\frac{dA_s(t)}{dt} &= b(A_s(t-\tau)) e^{-d_{sj}\tau} e^{-\int_{t-\tau}^{t} \beta_j m_i(v) dv} - d_{sa} A_s(t) - \beta_a m_i(t) A_s(t), \\
\frac{dJ_i(t)}{dt} &= -d_{ij} J_i(t) + \beta_j m_i(t) J_s(t) - \beta_j b(A_s(t-\tau)) W_1(t), \\
\frac{dA_i(t)}{dt} &= -d_{ia} A_i(t) + \beta_a m_i(t) A_s(t) + \beta_j b(A_s(t-\tau)) W_1(t), \\
\frac{dm_i(t)}{dt} &= -d_m m_i(t) + (m_T(t) - m_i(t)) \beta_m (J_i(t) + \alpha A_i(t)), \\
\frac{dW_1(t)}{dt} &= W_1(t)(d_{sj} - d_{ij} + \beta_j m_i(t-\tau)) + m_i(t) e^{-d_{sj}\tau} e^{-\int_{t-\tau}^{t} \beta_j m_i(v) dv} \\
&\qquad - e^{-d_{ij}\tau} m_i(t-\tau).
\end{aligned}$$

The DFE of model (2.43) is the equilibrium in which

$$(J_s, A_s, J_i, A_i, m_i, W_1) \equiv (J_s^*, A_s^*, 0, 0, 0, 0).$$

In the simulations reported below, we take the birth function of mosquitoes and that of birds as

$$B(m_T) = b_m m_T e^{-a_m m_T}, \qquad b(A_s) = b_b A_s e^{-a_b A_s}, \tag{2.44}$$

respectively. These forms for the birth function have been used, for example, in the well-studied Nicholson blowflies equation [9].

Various parameter values are given in Table 1, taken from [18, 19, 3, 26] with reference to West Nile virus. We took the initial conditions to be

$$A_s(t) = 500, \qquad M_I(t) = 0$$

for $t \in [-\tau, 0]$ and $A_i(0) = 2$. This, together with the matching condition (2.19), gives $J_s(0) = 16700$ and $J_i(0) = 0$.

In Figure 1 the condition (2.38) is satisfied, and the infected populations go to zero. However, as we increase the contact rates, the condition (2.38) fails, and the disease sustains in the bird and mosquito population, as shown in Figure 2. If we continue to increase the contact rates, we eventually find oscillatory behaviors, as shown in Figure 3, suggesting the possibility of a Hopf bifurcation to periodic solutions.

TABLE 1
Meaning of parameters.

Parameter	Meaning of the parameter	Value
b_b	Maximum per capita daily bird production rate	0.5
$1/a_b$	Size of bird population at which the number of newborn birds is maximized	1000
b_m	Maximum per capita daily mosquito egg production rate	5
$1/a_m$	Size of mosquito population at which egg laying is maximized	10000
d_{sj}	Mortality rate of uninfected juveniles (per day)	0.005
d_{ij}	Mortality rate of infected juveniles (per day)	0.05
d_{sa}	Mortality rate of uninfected adults (per day)	0.0025
d_{ia}	Mortality rate of infected adults (per day)	0.015
d_m	Mortality rate of mosquito (per day)	0.05
β_j	Contact rate between uninfected juvenile and infected mosquito	Variable
β_a	Contact rate between uninfected adult and infected mosquito	Variable
β_m	Contact rate between uninfected mosquito and infected juvenile	Variable
$\alpha\beta_m$	Contact rate between uninfected mosquito and infected juvenile	Variable
τ	Duration of more vulnerable period of bird (day)	160
σ	Maturation time of mosquito (day)	10
d_l	Mortality rate of larva mosquito (per day)	0.1

3. Spatial speed of spread. In this section we will derive a reaction-diffusion analogue of the system we have studied thus far, and we will use this system to formally estimate the speed at which the disease epidemic would spread through space. For simplicity, diffusion will be modeled using Fick's law. Equations (2.1) and (2.2) become

$$\frac{\partial s}{\partial t} + \frac{\partial s}{\partial a} = D_s(a)\frac{\partial^2 s}{\partial x^2} - d_s(a)s(t,a,x) - \beta(a)s(t,a,x)m_i(t,x) \tag{3.1}$$

and

$$\frac{\partial i}{\partial t} + \frac{\partial i}{\partial a} = D_i(a)\frac{\partial^2 i}{\partial x^2} - d_i(a)i(t,a,x) + \beta(a)s(t,a,x)m_i(t,x) \tag{3.2}$$

on a one-dimensional spatial domain $x \in (-\infty, \infty)$, where $m_i(t,x)$ is the number of infected adult mosquitoes at (t,x) satisfying a reaction-diffusion equation mentioned later. We shall assume that the age-dependent diffusivities $D_s(a)$, $D_i(a)$ have the special form

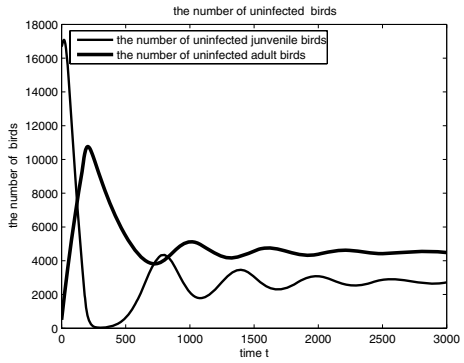
$$D_s(a) = \begin{cases} D_{sj}, & a < \tau, \\ D_{sa}, & a > \tau, \end{cases} \qquad D_i(a) = \begin{cases} D_{ij}, & a < \tau, \\ D_{ia}, & a > \tau. \end{cases} \tag{3.3}$$

With this choice for the diffusivities, our concern for the moment is with deriving a system of four reaction-diffusion equations for the quantities

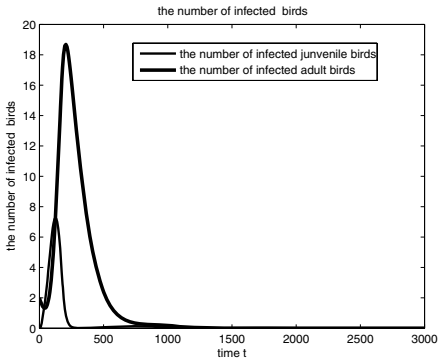
$$\begin{aligned} A_s(t,x) &= \int_\tau^\infty s(t,a,x)\,da, \qquad A_i(t,x) = \int_\tau^\infty i(t,a,x)\,da, \\ J_s(t,x) &= \int_0^\tau s(t,a,x)\,da, \qquad J_i(t,x) = \int_0^\tau i(t,a,x)\,da, \end{aligned} \tag{3.4}$$

which are analogous to the total numbers in (2.5). Differentiating the expression for $A_s(t,x)$ and using (3.1) and (3.3) gives

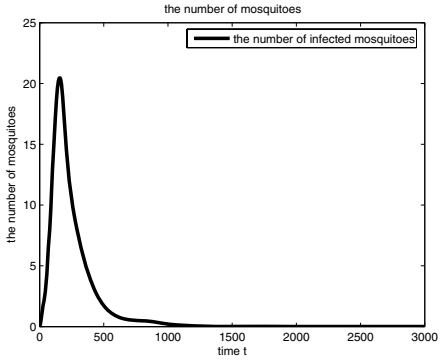
$$\frac{\partial A_s}{\partial t} = s(t,\tau,x) + D_{sa}\frac{\partial^2 A_s}{\partial x^2} - d_{sa}A_s - \beta_a m_i(t,x)A_s,$$



(a) Uninfected birds



(b) Infected birds



(c) Infected mosquitoes

FIG. 1. *Parameter values are* $\beta_j = 3.5 \times 10^{-6}$, $\beta_a = 1.5 \times 10^{-6}$, $\beta_m = 3.25 \times 10^{-6}$, $\alpha\beta_m = 7.5 \times 10^{-7}$, *and other parameters have the values shown in Table* 1. *In this case* d_m *is larger than the right-hand side of* (2.38), *which equals* 0.0343. *The DFE is stable.*

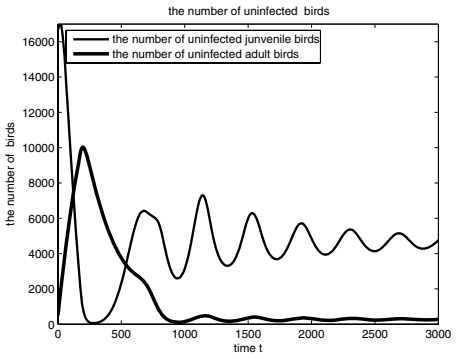
and we need to find $s(t, \tau, x)$. Set

$$s_\xi(a, x) = s(\xi + a, a, x).$$

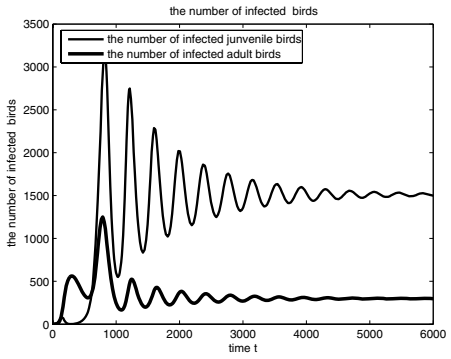
Differentiating with respect to a and using (3.1) gives

$$\frac{\partial s_\xi}{\partial a} = D_s(a)\frac{\partial^2 s_\xi}{\partial x^2} - d_s(a)s_\xi(a, x) - \beta(a)s_\xi(a, x)m_i(\xi + a, x). \tag{3.5}$$

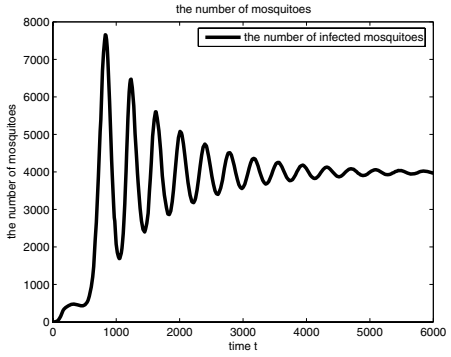
We would like to solve (3.5) exactly for $s_\xi(a, x)$, but this is impossible because the equation is nonautonomous. (The variable m_i satisfies a separate nonlinear partial differential equation, which appears below.) Our aim, however, will be to study the spatial spread of the disease by looking for traveling wave solutions which move leftwards through the spatial domain $x \in (-\infty, \infty)$, and which constitute a connection between the disease-free state and an endemic state. The PDEs we derive for the



(a) Uninfected birds



(b) Infected birds



(c) Infected mosquitoes

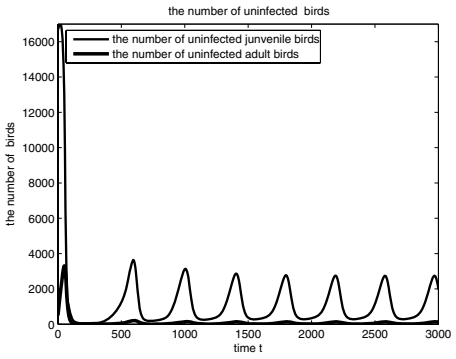
FIG. 2. *Parameter values are* $\beta_j = 4.0883 \times 10^{-6}$, $\beta_a = 2.3705 \times 10^{-6}$, $\beta_m = 3.7962 \times 10^{-6}$, $\alpha\beta_m = 1.1853 \times 10^{-6}$, *and other parameters have the values shown in Table* 1. *In this case* d_m *is less than the right-hand side of* (2.38), *which equals* 0.0613. *The DFE is unstable, and the solution evolves to an endemic equilibrium.*

variables (3.4), and for $m_i(t,x)$, will be studied only in the region far ahead of the advancing epidemic, i.e., as $x \to -\infty$, because we shall be assuming that the linearized equations in this region determine the speed of the epidemic wave. In the disease-free region $x \approx -\infty$, the variables $A_i(t,x)$, $J_i(t,x)$, and $m_i(t,x)$ are all close to zero. Thus, we solve (3.5) in the case when m_i is zero to find that in this case the solution subject to the first condition appearing below,

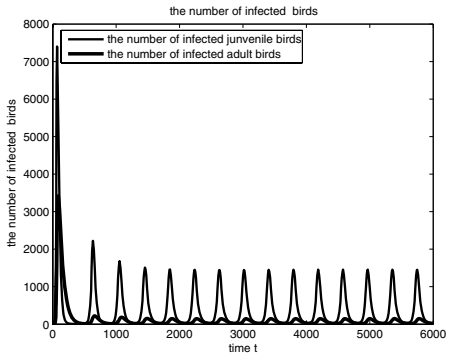
$$s(t,0,x) = b(A_s(t,x)), \qquad i(t,0,x) = 0 \tag{3.6}$$

(the analogue of (2.6)) is, for $a \le \tau$ and $\xi \ge 0$,

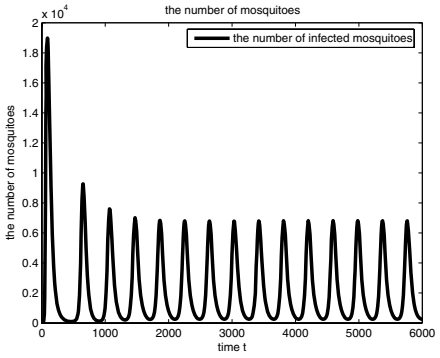
$$s_\xi(a,x) = s(\xi + a, a, x) = \int_{-\infty}^{\infty} \Gamma(D_{sj}a, x-y) b(A_s(\xi,y)) e^{-d_{sj}\tau}\, dy, \tag{3.7}$$



(a) Uninfected birds



(b) Infected birds



(c) Infected mosquitoes

FIG. 3. *Parameter values are* $\beta_j = 9.4030 \times 10^{-6}$, $\beta_a = 2.2388 \times 10^{-6}$, $\beta_m = 8.7313 \times 10^{-6}$, $\alpha\beta_m = 4.0299 \times 10^{-6}$, *and other parameters have the values shown in Table* 1. *In this case* d_m *is less than the right-hand side of* (2.38), *which equals* 0.2475. *The DFE is unstable, and the solution is oscillating.*

where

$$\Gamma(t,x) = \frac{1}{\sqrt{4\pi t}} e^{-x^2/4t}. \tag{3.8}$$

From (3.7) we find an expression for $s(t,\tau,x)$, and we deduce that for $t \geq \tau$ the PDE for $A_s(t,x)$ is

$$\begin{aligned} \frac{\partial A_s}{\partial t} &= \int_{-\infty}^{\infty} \Gamma(D_{sj}\tau, x-y) b(A_s(t-\tau,y)) e^{-d_{sj}\tau}\, dy \\ &\quad + D_{sa}\frac{\partial^2 A_s}{\partial x^2} - d_{sa}A_s(t,x) - \beta_a m_i(t,x) A_s(t,x), \end{aligned} \tag{3.9}$$

valid in the far left of the spatial domain $x \in (-\infty, \infty)$. Similarly, we obtain the following approximate equation for $J_s(t,x)$, also valid only in the far field $x \to -\infty$:

$$\begin{aligned}\frac{\partial J_s}{\partial t} &= b(A_s(t,x)) - \int_{-\infty}^{\infty} \Gamma(D_{sj}\tau, x-y) b(A_s(t-\tau,y)) e^{-d_{sj}\tau}\, dy \\ &\quad + D_{sj}\frac{\partial^2 J_s}{\partial x^2} - d_{sj} J_s(t,x) - \beta_j m_i(t,x) J_s(t,x).\end{aligned} \tag{3.10}$$

Next we shall derive the PDE for $A_i(t,x)$. Differentiating the expression for A_i in (3.4) and using (3.2) and (3.3) gives

$$\frac{\partial A_i}{\partial t} = i(t,\tau,x) + D_{ia}\frac{\partial^2 A_i}{\partial x^2} - d_{ia}A_i + \beta_a m_i(t,x) A_s,$$

and we need to find $i(t,\tau,x)$. Set

$$i_\xi(a,x) = i(\xi + a, a, x).$$

Since the calculation of $i(t,\tau,x)$ involves immature ages $a \in [0,\tau]$ only, from (3.2) we obtain

$$\frac{\partial i_\xi}{\partial a} = D_{ij}\frac{\partial^2 i_\xi}{\partial x^2} - d_{ij} i_\xi(a,x) + \beta_j m_i(\xi + a, x) s(\xi + a, a, x).$$

The solution of this equation satisfying the second condition in (3.6) is

$$i_\xi(a,x) = \beta_j \int_0^a e^{-d_{ij}(a-\zeta)} \int_{-\infty}^{\infty} \Gamma(D_{ij}(a-\zeta), x-y) m_i(\xi+\zeta, y) s(\xi+\zeta, \zeta, y)\, dy\, d\zeta,$$

where Γ is again given by (3.8). For $s(\xi+\zeta,\zeta,y)$ we use expression (3.7). Then, setting $a = \tau$ and $\xi = t - \tau$ in the above expression gives us $i(t,\tau,x)$, and thus we conclude that the evolution PDE for the variable $A_i(t,x)$ representing the number of adult infected hosts is, for $t \geq \tau$,

$$\begin{aligned}\frac{\partial A_i}{\partial t} &= D_{ia}\frac{\partial^2 A_i}{\partial x^2} - d_{ia}A_i(t,x) + \beta_a m_i(t,x) A_s(t,x) \\ &\quad + \beta_j \int_0^\tau e^{-d_{ij}(\tau-\zeta)} \int_{-\infty}^{\infty} \Gamma(D_{ij}(\tau-\zeta), x-y) m_i(t-\tau+\zeta, y) \\ &\quad \times \int_{-\infty}^{\infty} \Gamma(D_{sj}\zeta, y-\eta) b(A_s(t-\tau,\eta)) e^{-d_{sj}\zeta}\, d\eta\, dy\, d\zeta.\end{aligned} \tag{3.11}$$

This is again valid only in the far field $x \to -\infty$, since we have used expression (3.7). The last term in the right-hand side of (3.11) is the rate at which infected immatures become infected adults and has a similar interpretation to a term in the right-hand side of (2.13). This time the term involves additional integrals because of diffusion, but the reader may notice that in certain other respects the term in (3.11) is a little simpler than we might expect based on comparison with (2.13); this is due to the approximations we have made to derive (3.11) because of the restriction to the $x \approx -\infty$ zone. The interpretation of the term we are discussing is as follows. Each individual that reaches adulthood at point x at time t as an infected individual was born as a susceptible at time $t - \tau$ at some other point η. For an amount of time ζ that individual drifted around as a susceptible individual with diffusivity D_{sj} until

reaching a point y, where it became infected at time $t-\tau+\zeta$. For an amount of time $\tau-\zeta$, constituting the remainder of its childhood, it drifted around as an infected individual with diffusivity D_{ij} to reach point x at time t, where it becomes an adult. The two exponential factors represent the probability of surviving the susceptible and infected portions of childhood.

The PDE for $J_i(t,x)$ is derived similarly and turns out to be

$$\begin{aligned}\frac{\partial J_i}{\partial t} &= D_{ij}\frac{\partial^2 J_i}{\partial x^2} - d_{ij}J_i(t,x) + \beta_j m_i(t,x)J_s(t,x)\\ &\quad - \beta_j \int_0^\tau e^{-d_{ij}(\tau-\zeta)} \int_{-\infty}^{\infty} \Gamma(D_{ij}(\tau-\zeta), x-y) m_i(t-\tau+\zeta, y)\\ &\quad \times \int_{-\infty}^{\infty} \Gamma(D_{sj}\zeta, y-\eta) b(A_s(t-\tau,\eta)) e^{-d_{sj}\zeta}\, d\eta\, dy\, d\zeta.\end{aligned} \tag{3.12}$$

Finally we need a reaction-diffusion equation for the infected adult mosquitoes $m_i(t,x)$. We shall take

$$\frac{\partial m_i}{\partial t} = D_m\frac{\partial^2 m_i}{\partial x^2} - d_m m_i(t,x) + \beta_m(m_T^* - m_i(t,x))(J_i(t,x)+\alpha A_i(t,x)). \tag{3.13}$$

The system of PDEs to be solved thus consists of (3.9), (3.10), (3.11), (3.12), and (3.13). As explained previously, we shall look for solutions which constitute a leftward moving traveling wave-front and which invade what was formerly a disease-free zone; in other words, as $x \to -\infty$ we shall assume that the variables tend to the disease-free values in which A_i, J_i, and m_i are zero while $A_s^* > 0$ and $J_s^* > 0$ are given by (2.36), assuming that (2.23) holds. (If (2.23) does not hold, then the host population is eradicated even in the absence of the disease.)

We shall, in fact, look for a wave-front that constitutes a transition from the disease-free state to an endemic steady state, and so we need to be assured of the existence of an endemic state. The endemic state cannot be found explicitly, but fortunately we know the condition for its existence. This condition is the opposite of (2.38). Therefore, we assume in this section that

$$\begin{aligned}d_m < \beta_m m_T^* \Bigg\{ &\frac{b(A_s^*)\beta_j}{d_{ij}-d_{sj}}\left[\frac{1-e^{-d_{sj}\tau}}{d_{sj}} - \frac{(1-e^{-d_{ij}\tau})}{d_{ij}}\right]\\ &+ \frac{\alpha}{d_{ia}}\left[\beta_a A_s^* + \beta_j b(A_s^*)e^{-d_{sj}\tau}\frac{(1-e^{-(d_{ij}-d_{sj})\tau})}{d_{ij}-d_{sj}}\right]\Bigg\}.\end{aligned} \tag{3.14}$$

We linearize the equations for A_i, J_i, and m_i ((3.11), (3.12), and (3.13)) in the region $x \to -\infty$, where $A_s \to A_s^*$, $J_s \to J_s^*$, and the other variables approach zero. The linearized equations are then converted to traveling wave form by looking for solutions that are functions only of the variable $z = x + ct$ with $c \geq 0$. Then we look for nontrivial solutions of the linearized traveling wave equations of the form $(A_i, J_i, m_i) = (q_1, q_2, q_3)\exp(\lambda z)$. After a fair amount of algebra we find that the characteristic equation determining λ is

$$G_1(\lambda; c) = G_2(\lambda; c)G_3(\lambda; c), \tag{3.15}$$

where

$$\begin{aligned}G_1(\lambda;c) &= (D_{ia}\lambda^2 - d_{ia} - c\lambda)(D_{ij}\lambda^2 - d_{ij} - c\lambda)(D_m\lambda^2 - d_m - c\lambda)\\ &\quad - \beta_m m_T^*[\beta_j J_s^*(D_{ia}\lambda^2 - d_{ia} - c\lambda) + \alpha\beta_a A_s^*(D_{ij}\lambda^2 - d_{ij} - c\lambda)],\end{aligned} \tag{3.16}$$

$$G_2(\lambda;c) = \alpha(D_{ij}\lambda^2 - d_{ij} - c\lambda) - (D_{ia}\lambda^2 - d_{ia} - c\lambda), \tag{3.17}$$

and

$$G_3(\lambda;c) = \beta_m m_T^* \left(\frac{\beta_j b(A_s^*)(e^{-d_{sj}\tau} - e^{-d_{ij}\tau - \lambda c\tau + \lambda^2 D_{ij}\tau})}{d_{ij} - d_{sj} + \lambda c - \lambda^2 D_{ij}} \right). \tag{3.18}$$

Recall that A_s^* and J_s^* are given by (2.36) and that m_T^* is given by (2.17).

An epidemiologically feasible wave-front is one in which all the variables remain nonnegative as $x \to -\infty$ (as $z \to -\infty$ in the traveling wave variable formulation). The decay of A_i, J_i, and m_i to zero as $z \to -\infty$ must not be oscillatory. It is therefore necessary that there should exist at least one strictly positive real root λ of the characteristic equation (3.15) with the property that the corresponding eigenvector (q_1, q_2, q_3) points into the positive octant in $\mathbf{R}^3$. This actually happens only for c above some minimum value $c_{\min} > 0$. Define

$$c_{\min} = \inf\{c : \quad \exists \lambda \in (0, \tfrac{1}{2D_{ia}}(c + \sqrt{c^2 + 4d_{ia}D_{ia}})] \text{ satisfying (3.15)}\}. \tag{3.19}$$

The reason why the search for positive real roots λ of (3.15) is confined to the finite interval in (3.19) is that the eigenvector (q_1, q_2, q_3) corresponding to an eigenvalue λ exceeding $\frac{1}{2D_{ia}}(c + \sqrt{c^2 + 4d_{ia}D_{ia}})$ has q_1 and q_3 of opposite sign (implying that one of A_i or m_i is negative) so that such an eigenvalue corresponds to an infeasible solution. Note that the interval of λ in (3.19) is c dependent.

A calculation shows that, because of (3.14),

$$G_1(0;c) - G_2(0;c)G_3(0;c) > 0.$$

If for a fixed c one plots the graph of $G_1(\lambda;c) - G_2(\lambda;c)G_3(\lambda;c)$ against λ on the feasible domain $\lambda \in [0, \frac{1}{2D_{ia}}(c + \sqrt{c^2 + 4d_{ia}D_{ia}})]$, one finds that for a very small value of c the graph is always above the horizontal axis. The effect of increasing c is that a minimum begins to form within the feasible domain, and this minimum moves down and touches the horizontal axis at a critical c, the value $c_{\min}$ defined in (3.19) above. Figure 4 shows the critical situation for a particular set of parameter values shown in the caption, and for the two birth functions $b(\cdot)$ and $B(\cdot)$ chosen as in section 2.4. The value $c_{\min}$ can be found by numerically solving the simultaneous equations

$$G_1(\lambda;c) - G_2(\lambda;c)G_3(\lambda;c) = 0,$$
$$\frac{d}{d\lambda}[G_1(\lambda;c) - G_2(\lambda;c)G_3(\lambda;c)] = 0,$$

for c and λ with $c > 0$ and $\lambda \in (0, \frac{1}{2D_{ia}}(c + \sqrt{c^2 + 4d_{ia}D_{ia}})]$.

4. Discussion. The minimum speed of spread computed in the previous section according to the predictions of the linearized analysis was about 2.62 km/day, i.e., about 956 km/year. This is certainly roughly consistent with the speed at which West Nile virus has spread across the USA. The disease first emerged in New York in 1999 and had reached the West coast five years later. We should point out, however, that there is some uncertainty regarding the choice of parameter values, especially the diffusivities. We have availed ourselves of what data there is concerning the diffusivity of adult crows, but our choice of a value for the fledgling crows, which do not fly so well and may well spend some time on the ground (where they are, of course, vulnerable to predators such as cats) is purely our estimate. While the speed of spread does show a dependence on the diffusivities, we noted a lack of sensitivity to the values of

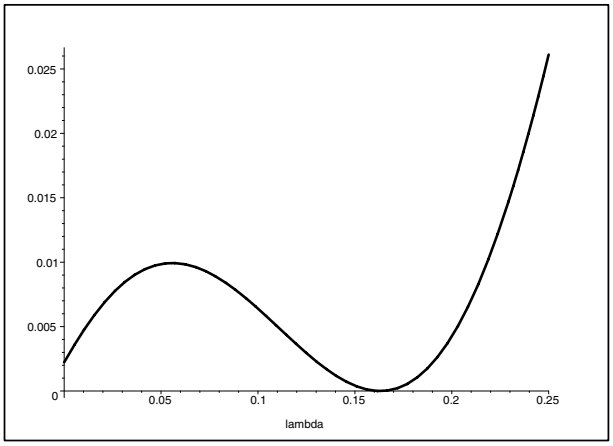


FIG. 4. *Parameter values are* $\beta_j = 3.15 \times 10^{-5}$, $\beta_a = 1.5 \times 10^{-5}$, $\beta_m = 2.925 \times 10^{-5}$, $\alpha\beta_m = 0.75 \times 10^{-5}$, $D_{ia} = 13$ km^2/day [16, 21] *(the diffusion rate of infected adult),* $D_{ij} = 6$ km^2/day *(the diffusion rate of infected juvenile),* $D_m = 0.1$ km^2/day *(the diffusion rate of mosquito), and other parameters have the values shown in Table* 1. *For these values, the minimum speed* $c_{\min}$*, computed as described in the text, equals* 2.623164094 *km/day.*

some of them (e.g., the diffusivity of mosquitoes) and a sensitivity to the values of other parameters, particularly the contact rates.

Ideally it would be desirable to have some information on whether the minimum speed $c_{\min}$ computed as described in section 3 is really the speed that solutions would evolve to, from ecologically realistic initial data such as a localized introduction of infectives. One must remember that in deriving the reaction-diffusion model, we were restricted to the vicinity of the DFE because the model derivation requires an explicit solution to a certain linear parabolic PDE that is nonautonomous except near that equilibrium. The inability to formulate a model that is valid everywhere in the spatial domain has made it impossible to numerically simulate the spatially extended model (such a simulation might have confirmed that the spread rate is indeed the minimal wave speed $c_{\min}$). The mathematical theory of the speed of spread in reaction-diffusion equations with functional terms is still far from complete, especially for coupled systems such as those in this paper. Relating the spread rate of the disease to the traveling wave with the minimal wave speed relies on the so-called linear conjecture (see [25, 15]). The fact that the minimal speed coincides with the spread rate has been theoretically verified only for dynamical systems enjoying certain order-preserving properties (see the two recent articles [24, 17]), and counterexamples when these properties do not hold have been reported [11]. Establishing this fact for our system (3.9)–(3.13) is even more difficult due to the interaction of time delay and spatial diffusion, in addition to the nonlocality of the nonlinear terms. Therefore, it has to be emphasized that our calculation of $c_{\min}$ is nothing more than a formal calculation of the minimum ecologically feasible speed according to the linearized equations ahead of the front.

Throughout this paper simple mass action terms have been used. In some virus infections, possibly including mosquito borne disease, one might argue for the inclusion of a term which represents the fact that a female mosquito takes a fixed number of blood meals per unit time (Anderson and May [1]). Such a modification involves dividing by bird density and has recently been utilized by Lewis, Renclawowicz, and

van den Driessche [16] and by Bowman et al. [3] in some simpler models for West Nile virus. In the present paper such a modification can be implemented only in the model without diffusion, which we have studied in section 2, and unfortunately not for the reaction-diffusion model of section 3, which becomes intractable. The type of modification we are discussing involves replacing (2.1) by

$$\frac{\partial s}{\partial t} + \frac{\partial s}{\partial a} = -d_s(a)s(t,a) - \frac{\beta(a)s(t,a)m_i(t)}{N(t)}, \tag{4.1}$$

with another similar modification to (2.2). The variable $N(t)$ stands for the total bird population,

$$N(t) = A_s(t) + A_i(t) + J_s(t) + J_i(t),$$

in which the variables are defined by (2.5). Equation (2.18) would be replaced by

$$\frac{dm_i(t)}{dt} = -d_m m_i(t) + \frac{\beta_m(m_T^* - m_i(t))}{N(t)}(J_i(t) + \alpha A_i(t)). \tag{4.2}$$

For this modified model it is possible to develop a parallel theory including equations for the total number variables analogous to (2.9), (2.10), (2.13), (2.14) and to prove theorems concerning positivity, boundedness, and global convergence. We shall confine ourselves in this paragraph only to a discussion of the linear stability of the DFE in the modified model involving division by bird density. The DFE itself is still given precisely by (2.35). Lemma 2.3, which concerns stability to perturbations in which the disease remains absent, still holds. For the modified model a necessary and sufficient condition for the DFE to be linearly asymptotically stable to arbitrary small perturbations is

$$\begin{aligned} d_m > \frac{\beta_m m_T^*}{N^*}\Bigg\{ & \frac{b(A_s^*)\beta_j}{N^*(d_{ij}-d_{sj})}\left[\frac{1-e^{-d_{sj}\tau}}{d_{sj}} - \frac{(1-e^{-d_{ij}\tau})}{d_{ij}}\right] \\ & + \frac{\alpha}{d_{ia}N^*}\left[\beta_a A_s^* + \beta_j b(A_s^*)e^{-d_{sj}\tau}\frac{(1-e^{-(d_{ij}-d_{sj})\tau})}{d_{ij}-d_{sj}}\right]\Bigg\}, \end{aligned} \tag{4.3}$$

which is similar to condition (2.38). Here, $N^* = A_s^* + J_s^*$, where A_s^* and J_s^* are given by (2.36).

There are a number of ways in which one could interpret conditions (2.38) and (4.3) for the simple mass action model and the modified model, respectively. First let us note that as far as the stability of the DFE is concerned the two models are similar: to get from one to the other we simply divide the contact rates by the total bird population at the equilibrium. Not surprisingly, in reality in the control of West Nile virus a great deal of emphasis goes into mosquito control. This may mean larval control, i.e., reducing the number of places mosquito larvae may inhabit such as old tires, blocked gutters, bird baths, flower pots, swimming pool covers, etc. Adult mosquito control using adulticides, which are sprayed into the air from a sprayer truck as very tiny droplets, is also practiced, especially when larval control measures are clearly inadequate or disease is imminent. The per capita mortality rate for adult mosquitoes manifests itself in our model as the parameter d_m. The per capita mortality rate for mosquito larvae is d_l, which does not feature directly in (2.38) or (4.3) but features indirectly through the quantity m_T^*. (In fact, m_T^* depends on both d_l and d_m.) If the birth function $B(\cdot)$ for mosquitoes is chosen as

in (2.44), then $m_T^* = \frac{1}{a_m}\ln(b_m/d_m) - d_l\sigma/a_m$, and so the right-hand side of (2.38) or (4.3) decreases linearly with d_l so that sufficiently effective larval control eradicates the disease. On the other hand, as d_m increases, the left-hand side increases linearly while the right-hand side decreases, suggesting that in percentage terms an increase in d_m might be more effective than an increase in larval mortality d_l. However, adult mosquito control is more expensive and more difficult to organize.

There are a number of other factors we have not considered in this paper at all. It seems that in reality seasonal effects probably play an important role and should be modeled. It is really only in the breeding season that crows, once paired, seek to establish individual territories to raise their broods. In the nonbreeding season crow activities tend to be centered around large communal roosts to which they return in the evenings after searching for food during the day (some roost locations may have been gathering points for crows for many decades). Crows also have a strong flocking instinct, something which Fickian diffusion does not model at all. Northern birds tend to fly south during the winter. All these considerations indicate possible areas for further investigation.

Acknowledgment. We are grateful to the referees for their valuable comments and suggestions which have led to an improved paper.

REFERENCES

[1] R. M. Anderson and R. M. May, *Infectious Diseases of Humans*, Oxford University Press, London, 1991.

[2] J. S. Blackmore and R. P. Dow, *Differential feeding of Culex tarsalis on nestling and adult birds*, Mosq. News, 18 (1958), pp. 15–17.

[3] C. Bowman, A. B. Gumel, P. van den Driessche, J. Wu, and H. Zhu, *A mathematical model for assessing control strategies against West Nile virus*, Bull. Math. Biol., 67 (2005), pp. 1107–1133.

[4] F. Brauer and C. Castillo-Chavez, *Mathematical Models in Population Biology and Epidemiology*, Springer, New York, 2001.

[5] O. Diekmann and J. A. P. Heesterbeek, *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*, Wiley Ser. Math. Comput. Biol., John Wiley and Sons, Chichester, UK, 2000.

[6] L. Edelstein-Keshet, *Mathematical Models in Biology*, Birkhäuser Math. Ser., McGraw-Hill, Toronto, 1988.

[7] S. A. Gourley, J. W.-H. So, and J. Wu, *Nonlocal reaction-diffusion equations with delay: Biological models and nonlinear dynamics*, J. Math. Sci., 124 (2004), pp. 5119–5153.

[8] S. A. Gourley and J. Wu, *Delayed non-local diffusive systems in biological invasion and disease spread*, in Nonlinear Dynamics and Evolution Equations, Fields Inst. Commun. 48, Amer. Math. Soc., Providence, RI, 2006, pp. 137–200.

[9] W. S. C. Gurney, S. P. Blythe, and R. M. Nisbet, *Nicholson's blowflies revisited*, Nature, 287 (1980), pp. 17–21.

[10] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Rev., 42 (2000), pp. 599–653.

[11] Y. Hosono, *The minimal speed of traveling fronts for a diffusive Lotka Volterra competition model*, Bull. Math. Biol., 60 (1998), pp. 435–448.

[12] D. S. Jones and B. D. Sleeman, *Differential Equations and Mathematical Biology*, Chapman & Hall/CRC Math. Biol. Med. Ser., Chapman & Hall/CRC, Boca Raton, FL, 2003.

[13] M. Kot, *Elements of Mathematical Ecology*, Cambridge University Press, Cambridge, UK, 2001.

[14] Y. Kuang, *Delay Differential Equations with Applications in Population Dynamics*, Math. Sci. Engrg. 191, Academic Press, Boston, MA, 1993.

[15] M. A. Lewis, B. Li, and H. F. Weinberger, *Spreading speed and linear determinancy for two-species competition models*, J. Math. Biol., 45 (2002), pp. 219–233.

[16] M. Lewis, J. Renclawowicz, and P. van den Driessche, *Traveling waves and spread rates for a West Nile virus model*, Bull. Math. Biol., 68 (2006), pp. 3–23.

[17] X. Liang and X.-Q. Zhao, *Asymptotic speeds of spread and traveling waves for monotone semiflows with applications*, Comm. Pure Appl. Math., 60 (2007), pp. 1–40.

[18] C. C. Lord and J. F. Day, *Simulation studies of St. Louis encephalitis virus in south Florida*, Vector Borne and Zoonotic Diseases, 1 (2001), pp. 299–315.
[19] C. C. Lord and J. F. Day, *Simulation studies of St. Louis encephalitis and West Nile viruses: The impact of bird mortality*, Vector Borne and Zoonotic Diseases, 1 (2001), pp. 317–329.
[20] J. D. Murray, *Mathematical Biology*, Springer-Verlag, New York, 1993.
[21] A. Okubo, *Diffusion-type models for avian range expansion*, in Proceedings of the 19th International Ornithological Congress (Ottawa), Vol. 1, H. Queslet, ed., 1988, pp. 1038–1049.
[22] T. W. Scott, L. H. Lorenz, and J. D. Edman, *Effects of house sparrow age and arbovirus infection on attraction of mosquitoes*, J. Med. Entomol., 27 (1990), pp. 856–863.
[23] H. L. Smith, *Monotone Dynamical Systems. An Introduction to the Theory of Competitive and Cooperative Systems*, Amer. Math. Soc., Providence, RI, 1995.
[24] H. R. Thieme and X.-Q. Zhao, *Asymptotic speeds of spread and traveling waves for integral equations and delayed reaction-diffusion models*, J. Differential Equations, 195 (2003), pp. 430–470.
[25] H. F. Weinberger, M. A. Lewis, and B. Li, *Analysis of linear determinacy for spread in cooperative models*, J. Math. Biol., 45 (2002), pp. 183–218.
[26] M. J. Wonham, T. de-Camino-Beck, and M. A. Lewis, *An epidemiological model for West Nile virus: Invasion analysis and control applications*, Proc. R. Soc. London Ser. B, 271 (2004), pp. 501–507.
[27] M. J. Wonham, M. A. Lewis, J. Renclawowicz, and P. van den Driessche, *Transmission assumptions generate conflicting predictions in host-vector disease models: A case study in West Nile virus*, Ecol. Lett., 9 (2006), pp. 706–725.
[28] J. Wu, *Global dynamics of strongly monotone retarded equations with infinite delay*, J. Integral Equations Appl., 4 (1992), pp. 273–307.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 434–446

ANALYSIS OF THE DYNAMICS AND TOUCHDOWN IN A MODEL OF ELECTROSTATIC MEMS*

G. FLORES†, G. MERCADO‡, J. A. PELESKO§, AND N. SMYTH¶

Abstract. We study a reaction-diffusion equation in a bounded domain in the plane, which is a mathematical model of an idealized electrostatically actuated microelectromechanical system (MEMS). A relevant feature in these systems is the "pull-in" or "jump-to contact" instability, which arises when applied voltages are increased beyond a critical value. In this situation, there is no longer a steady state configuration of the device where mechanical members of the device remain separate. It may present a limitation on the stable operation regime, as with a micropump, or it may be used to create contact, as with a microvalve. The applied voltage appears in the equation as a parameter. We prove that this parameter controls the dynamics in the sense that before a critical value the solution evolves to a steady state configuration, while for larger values of the parameter, the "pull-in" instability or "touchdown" appears. We estimate the touchdown time. In one dimension, we prove that the touchdown is self-similar and determine the asymptotic rate of touchdown. The same type of results are obtained in a disk. We also present numerical simulations in some two-dimensional domains which allow an estimate of the critical voltage and of the touchdown time. This information is relevant in the design of the devices.

Key words. microelectromechanical system, touchdown, quenching

AMS subject classifications. 34A34, 34C11, 35B30, 35K60

DOI. 10.1137/060648866

1. Introduction.

Lots of micro- and nanoelectromechanical systems rely upon electrostatic forces to make things move. Devices such as micropumps, microswitches, etc., can be modeled as electrostatically deflected elastic membranes.

Typically, the device consists of an elastic membrane suspended above a rigid ground plate, placed in series with a fixed voltage source and a fixed capacitor. In the limit of small aspect ratio, that is, small gap size relative to device length, the model can be reduced to a single scalar equation for the deflection of the membrane. Denoting this deflection by u, we have, in dimensionless variables,

$$\epsilon^2 u_{tt} + u_t - \triangle u = -\frac{\lambda f(x,y,t)}{(1+u)^2(1+\chi\int_\Omega \frac{1}{1+u})^2}$$

in Ω, with $u = 0$ on $\partial\Omega$. Here

$$\epsilon^2 = \frac{\rho E}{\nu^2} = \frac{\text{inertial terms}}{\text{damping terms}} \quad \text{and} \quad \lambda = \frac{V^2 L^2 \epsilon_0}{2Tl^2}.$$

*Received by the editors January 2, 2006; accepted for publication (in revised form) September 19, 2006; published electronically February 2, 2007. A preliminary version of this paper appeared in *Proceedings of ASME DETC'*03, Chicago, IL, 2003, pp. 1–8.
http://www.siam.org/journals/siap/67-2/64886.html

†Departamento de Matemáticas y Mecánica, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Apdo. Postal 20-726, 01000 México City, México (gfg@mym.iimas.unam.mx).

‡Facultad de Matemáticas, Universidad Autónoma de Zacatecas, 98068 Zacatecas, México (gema @mate.reduaz.mx).

§Mathematical Sciences, University of Delaware, Newark, DE 19716-2553 (pelesko@math.udel. edu).

¶School of Mathematics, University of Edinburgh, Edinburgh EH9 3JZ, Scotland, UK (N.Smyth@ ed.ac.uk).

For a derivation of the equation see [15].

The parameter V is the applied voltage, T the tension in the membrane, L a characteristic length of the domain, l a characteristic width of the gap between the membrane and the fixed electrode, and ϵ_0 the permittivity of free space. It is useful to think of λ as a control parameter proportional to the applied voltage. The function $f(x, y, t)$ may model varying dielectric properties of the membrane and an applied alternating current. Physically f is constrained to be positive. The integral in the equation arises when the device is placed in a circuit with a capacitor of fixed capacitance. The parameter χ is the ratio of this fixed capacitance to a reference capacitance of the device. The limit where $\chi = 0$ corresponds to removing the fixed capacitor from the circuit.

We shall study the viscosity dominated limit of the model above, that is, the limit where $\epsilon \to 0$. Further, we can remove the nonlocal integral term by introducing a new parameter and establishing a mapping between solutions to the nonlocal and a local problem. See [15] for details. The local problem corresponds to setting $\chi = 0$. For simplicity, we take $f \equiv 1$. The case of devices with dielectric properties which vary in space is treated in [9]. We also assume that the device starts from rest. So, we shall study

$$u_t - \triangle u = -\frac{\lambda}{(1+u)^2} \quad \text{in } \Omega, \quad \text{with } u = 0 \text{ on } \partial\Omega, \quad u(x, y, 0) = 0. \tag{1}$$

Electrostatically deflected elastic systems exhibit an instability known as "pull-in." For moderate voltages, the system is in a stable operation regime: the membrane approaches a steady state configuration and remains separate from the ground plate. When applied voltages are increased beyond a critical value, there is no longer a steady state configuration of the device, causing the membrane to collapse onto the ground plate. This phenomenon is what we call "touchdown."

The "pull-in" instability was observed in 1967 by Nathanson and coworkers in their investigation of electrostatic actuation as a method for designing a resonant gate transistor [12]. At around the same time, this instability was also observed by Taylor in his work on the coalescence of liquid drops held at different electric potentials [16].

Mathematically, the "pull-in" instability corresponds to $u = -1$ being achieved in Ω in finite time, which creates a singularity in (1).

What we call "touchdown" is also known as "pinching" or "quenching" in the mathematical literature, where it appears in the equation for the motion of a surface with a normal velocity proportional to its mean curvature. See [2], [8]. The analysis of parabolic equations with negative power nonlinearities began with the work of Kawarada in 1975 [10]. Quenching has been studied for semilinear and quasi-linear equations; see [4] and [11].

Our formulation, based on the fact that gravity acts downwards, leads to (1). For this reason, our solution takes negative values. To relate this to the study of quenching for positive solutions, it is enough to substitute u by $-u$ to obtain the parabolic equation in the standard form as it appears, for instance, in [11].

In this work, we describe the behavior of solutions in terms of the parameter λ: There exists a critical value λ^* such that solutions converge to a steady state for $\lambda \leq \lambda^*$, while for $\lambda > \lambda^*$ touchdown occurs.

We also characterize the asymptotic self-similar nature of touchdown by analyzing the one-dimensional equation in self-similar variables. We find a continuum of steady states, which correspond to self-similar solutions. We show that the nonconstant

steady states grow at infinity like e^{y^2}, thereby obtaining the constant solution as the only possible limiting configuration, as well as the asymptotic rate of touchdown. This asymptotic rate was also obtained by Guo, Pan, and Ward, using different techniques. See [9].

Finally, we present some numerical results for two-dimensional domains: ellipses and annular regions. We obtain bounds for the critical value of λ as well as for the touchdown time. One interesting feature of annular regions is that the touchdown set is a circle. Also, a thin annulus sustains a very large stable operation regime.

Some of the results presented here were either announced or proved in [6]. For the sake of completeness, we include some of the proofs from our previous work.

2. Steady state solutions. Let Ω be a convex, bounded domain in the plane with smooth boundary. We first prove the existence of stationary solutions for small values of λ.

LEMMA 2.1. *There exists $\lambda_0 > 0$ such that the Dirichlet problem for*

$$\Delta u = \frac{\lambda}{(1+u)^2} \tag{2}$$

has a solution for $\lambda \leq \lambda_0$.

Proof. Since $u \equiv 0$ is an upper solution, it is enough to construct nonpositive lower solutions. To this end, let u^* be the solution of $\Delta u^* = 1$ in Ω, $u = 0$ on $\partial\Omega$. By the maximum principle, $u^* < 0$ in Ω.

Let $m^* := \inf\{u^*(x) \,|x \in \Omega\}$ and $\alpha := \frac{-1}{2m^*}$. Now choose $\lambda \leq \alpha/4$; then

$$\Delta(\alpha u^*) - \frac{\lambda}{(1+\alpha u^*)^2} \geq \alpha - 4\lambda \geq 0. \tag{3}$$

Hence, u^* is a nonpositive lower solution for $\lambda \leq \alpha/4$. A standard monotone iteration scheme yields the existence of a stable stationary solution for $\lambda \leq \lambda_0$. This finishes the proof. □

In the case of a disk we get explicit bounds for λ_0.

COROLLARY 2.1. *If Ω is a disk of radius R, there exist stationary solutions for $\lambda \leq R^{-2}$.*

Proof. It follows from the fact that $u^*(x) = \frac{1}{4}(\sum_{i=1}^{2} x_i^2 - R^2)$ and that we can choose $\alpha = R^{-2}$.

Our next result gives a description of the structure of stationary solutions in terms of the parameter λ. □

THEOREM 2.1. *There exists $\lambda_* > 0$ such that there is at least one stationary solution for $\lambda < \lambda_*$ and none for $\lambda > \lambda_*$.*

Proof. This is a consequence of the fact that if u_1 is a stationary solution for $\lambda = \lambda_1$, then, for $\lambda < \lambda_1$,

$$\Delta u_1 = \frac{\lambda_1}{(1+u_1)^2} > \frac{\lambda}{(1+u_1)^2} \tag{4}$$

so that u_1 is a lower solution for all $\lambda < \lambda_1$.

Now let $E = \{\lambda > 0$ for which there is a stationary solution$\}$. By the paragraph above, E is a nonempty interval. Let $\lambda_* := \sup E$.

Our next task is to prove that λ_* is finite.

To this end, let μ_0 be the first eigenvalue of the Dirichlet Laplacian. We claim that for $\lambda > -\frac{4}{27}\mu_0$, the steady state problem has no solution.

This result is slightly more precise than Theorem 4.4 of [15]. The proof is similar: let u_0 be a positive eigenfunction corresponding to μ_0; then

$$\int_\Omega u_0 \left[\Delta u - \frac{\lambda}{(1+u)^2}\right] dx = \mu_0 \int_\Omega \frac{u_0}{(1+u)^2} \left[u(1+u)^2 - \frac{\lambda}{\mu_0}\right] dx. \tag{5}$$

If $\lambda > -\frac{4}{27}\mu_0$, then the right-hand side in the equation above is negative. This shows that there are no stationary solutions for such values of λ. The theorem is proved. □

In the one-dimensional case, with $\Omega = (-1/2, 1/2)$, we give a complete description of the structure of solutions of

$$u_{xx} = \frac{\lambda}{(1+u)^2} \quad \text{in} \quad \Omega \tag{6}$$

with $u(\pm 1/2) = 0$. Indeed, with the scaling $\xi = \sqrt{\lambda}x$, the previous equation becomes

$$u'' = \frac{1}{(1+u)^2} \quad \text{in} \quad (-\sqrt{\lambda}/2, \sqrt{\lambda}/2) \tag{7}$$

and $u(\pm\sqrt{\lambda}/2) = 0$.

The following result is a special case of Theorem 2.1 in [11]. However, our method of proof allows the explicit determination of the critical value of the parameter λ.

THEOREM 2.2. *There exists a constant C^* such that the previous equation has zero, one, or two solutions according as $\sqrt{\lambda}/2 > C^*$, $\sqrt{\lambda}/2 = C^*$, or $\sqrt{\lambda}/2 < C^*$. Moreover, the exact value is $C^* = .591611$, and this determines the critical value $\lambda^* = 1.400016469$.*

Proof. In the phase plane of u, $v = u'$, the integral curves satisfy

$$\frac{v^2}{2} + \frac{1}{1+u} = E_0,$$

where E_0 is the initial energy.

An integral curve starting at $(\overline{u}, 0)$ with $-1 < \overline{u} < 0$ reaches the v axis in finite time at $(0, \overline{v})$. It follows that

$$E_0 = \frac{\overline{v}^2}{2} + 1.$$

The equations for the integral curves take the form

$$v = \pm\sqrt{\overline{v}^2 + \frac{2u}{1+u}}, \tag{8}$$

which is defined for $u \geq -\frac{\overline{v}^2}{2+\overline{v}^2} = \overline{u}$. The travel time from $(\overline{u}, 0)$ to $(0, \overline{v})$ is given by the map T defined by

$$T(\overline{v}^2) = \int_{\overline{u}}^{0} \frac{du}{\sqrt{\overline{v}^2 + \frac{2u}{1+u}}}. \tag{9}$$

Note that $T \to 0$ as $\overline{v} \to 0$ and $+\infty$. We shall show that T has a unique maximum, from which the structure of stationary solutions is obtained.

To analyze the travel time map we let $\alpha = \overline{v}^2$ and $\beta = \frac{\alpha}{2+\alpha}$; then

$$T(\alpha) = \frac{1}{\sqrt{2+\alpha}} \int_{-\beta}^{0} \frac{\sqrt{1+u}}{\sqrt{u+\beta}}\, du = \frac{\sqrt{\alpha}}{2+\alpha} \int_{-\pi/2}^{0} \sqrt{1+\beta\sin(\theta)}\sqrt{1-\sin(\theta)}\, d\theta$$
$$= \int_{-\pi/2}^{0} F(\alpha,\theta) d\theta.$$

The critical value C^* is the maximum value of the travel time. The point where the maximum is achieved is computed explicitly and used in the integral representation of the travel time to determine C^*. This critical value is used to determine the critical parameter λ^*.

One can prove that the function $T(\alpha)$ has a unique maximum, as its second derivative is negative at critical points.

The proof is finished. □

The shape of the stationary solutions can be obtained from the trajectories in the phase plane. For instance, for small λ, there is one solution of small amplitude and a second solution with a peak at the center which gets close to -1. Both are even functions.

For a disk shaped domain we also have a complete picture of solutions. Details appear in [14]. Note that by Gidas, Ni, and Nirenberg, all stationary solutions are radially symmetric.

3. Dynamics: The stable operation and touchdown regimes. In this section we characterize the stable operation and touchdown regimes in terms of the parameter λ. We begin by establishing some general properties of solutions of the evolution equation (1).

THEOREM 3.1. *Assume that $u(x,y,t;\lambda) > -1$ for $\lambda > 0$, and for all $(x,y) \in \Omega$ and $t \in [0,T]$. Then we have the following:*

(i) *$u(x,y,t;\lambda)$ is decreasing in t at each $(x,y) \in \Omega$.*

(ii) *$u(x,y,t;\lambda)$ is a decreasing function of λ.*

(iii) *When $\Omega = (-1/2, 1/2)$, $u(x,t;\lambda)$ is an even function of x, it achieves its minimum at $x = 0$, and it is increasing for $x \in [0, 1/2]$.*

Proof.

(i) u_t is a solution of a linear parabolic equation with zero boundary values and initial value equal to $-\lambda$. By the maximum principle, $u_t \leq 0$, and the strict inequality holds in the interior of the domain.

(ii) It is an easy consequence of the maximum principle.

(iii) The symmetry of the solution is a consequence of the symmetry of the domain, of the initial, condition, and of the heat operator. It follows that $u_x(x,t)$ is an odd function of x; therefore $u_x(0,t) = 0$. Moreover, u_x is a solution of a parabolic equation for $x \in [0,1/2]$, and $t \in [0,T]$, with $u_x(0,t) = 0$ and, by the maximum principle, $u_x(1/2,t) > 0$. It follows that u_x cannot achieve a negative minimum, and hence $u_x \geq 0$. By the strong maximum principle, u_x cannot achieve its minimum in the region $x \in (0,1/2)$, and $t \in (0,T]$. Therefore, $u_x(x,t) > 0$ for $x \in (0,1/2)$, and $t \in (0,T]$. It follows that for a fixed $t \in (0,T]$, $u(x,t)$ achieves its minimum at $x = 0$ only.

The proof of the theorem is finished. □

Amplification. The plots of the profile u as a function of x suggest that it is a convex function. We have not been able to prove this, yet there is some sort of convexity in the approximations: if we solve (1) by iterations, starting with $u_0 \equiv 0$,

then the next approximation u_1 is a solution of the corresponding linear heat equation with forcing term $\equiv -\lambda$, and hence

$$u_1(x,t) = -\lambda \int_0^t \int_{-1/2}^{1/2} K(x,y,\tau)dy\,d\tau,$$

where $K(x,y,t)$ is the Dirichlet Green's function. It follows that

$$\frac{\partial^2 u_1}{\partial x^2} = \frac{\partial u_1}{\partial t} + \lambda = \lambda \left[1 - \int_{-1/2}^{1/2} K(x,y,\tau)\right] dy$$

and the last expression is positive by the maximum principle.

3.1. The stable operation regime. For $\lambda < \lambda_*$, the solution stabilizes to a steady state.

THEOREM 3.2. *For $\lambda < \lambda_*$, the solution $u(x,y,t;\lambda)$ converges to a stationary solution as $t \to \infty$.*

Proof. The solution $u(x,y,t;\lambda)$ is bounded below by any stationary solution. Hence, $u(x,y,t;\lambda)$ is defined for all $t > 0$ and, by the previous theorem, is decreasing in t at each $(x,y) \in \Omega$. Therefore, $u(x,y,t;\lambda)$ converges as $t \to \infty$.

The energy

$$\int_\Omega \left[\frac{1}{2}|\nabla u|^2 - \frac{\lambda}{1+u}\right] dx\,dy \tag{10}$$

decreases and is bounded below along such solutions. Therefore, the only points in the ω-limit set of such trajectories are steady states. The proof is finished. ∎

3.2. The touchdown regime. Our first result is an estimate on the values of λ for which touchdown occurs.

THEOREM 3.3. *For $\lambda > -\frac{4}{27}\mu_0$, $u(x,y,t;\lambda) = -1$ in finite time.*

Proof. Let μ_0 be the smallest eigenvalue of the Dirichlet Laplace operator on Ω and let u_0 be the corresponding eigenfunction. We note that u_0 may be chosen strictly positive in Ω and normalized so that

$$\int_\Omega u_0 = 1. \tag{11}$$

Now we derive an energy for our system. Multiply (1) by u_0 and integrate over the domain Ω. This yields

$$\frac{d}{dt}\int_\Omega u u_0 - \mu_0 \int_\Omega u u_0 = -\lambda \int_\Omega \frac{u_0}{(1+u)^2}, \tag{12}$$

where we have used Green's theorem to integrate by parts. We define

$$E(t) = \int_\Omega u u_0 \tag{13}$$

and rewrite as

$$\frac{dE}{dt} - \mu_0 E = -\lambda \int_\Omega \frac{u_0}{(1+u)^2}. \tag{14}$$

Then, applying Jensen's inequality and the initial condition on our problem, we arrive at

$$\frac{dE}{dt} - \mu_0 E \leq \frac{-\lambda}{(1+E)^2}, \tag{15}$$

$$E(0) = 0. \tag{16}$$

Now notice

$$E(t) = \int_\Omega u u_0 \geq \inf u \int_\Omega u_0 = \inf u. \tag{17}$$

Next, define ϕ to be the solution of

$$\frac{d\phi}{dt} = \mu_0 \phi - \frac{\lambda}{(1+\phi)^2}, \tag{18}$$

$$\phi(0) = 0. \tag{19}$$

By standard comparison principles we have

$$E(t) \leq \phi(t) \tag{20}$$

for all time. Hence

$$\inf u \leq \phi(t) \tag{21}$$

and the "worst" behavior of u is captured by $\phi(t)$. In the ODE for ϕ we can separate variables and integrate to time T:

$$T = -\int_0^{\phi(T)} \frac{d\phi}{-\mu_0 \phi + \frac{\lambda}{(1+\phi)^2}}. \tag{22}$$

If the integral

$$\int_{-1}^{0} \frac{d\phi}{-\mu_0 \phi + \frac{\lambda}{(1+\phi)^2}} \tag{23}$$

is finite, we have existence for ϕ only for a finite interval and touchdown must occur in finite time. The integral remains finite if the denominator is never zero. But, we can guarantee that the integral remains finite if $\lambda > -\frac{4}{27}\mu_0$. The proof is finished. □

As in the case of stationary solutions, the set of parameter values for which touchdown occurs is an interval.

THEOREM 3.4. *There exists $\lambda^* > 0$ such that touchdown occurs if $\lambda > \lambda^*$. Moreover, touchdown does not occur for $\lambda < \lambda^*$.*

Proof. It is a consequence of the maximum principle: solutions of the parabolic equation (1) are strictly decreasing functions of λ in Ω. Therefore $B := \{\lambda | \text{touchdown occurs}\}$ is an interval, it is nonempty by the previous theorem, and $\lambda^* := \text{infimum}\, B$ has the required property. The proof is finished. □

Note that $\lambda_* \le \lambda^*$. Fila and Kawohl have proved in [5] that quenching in infinite time is not possible for convex domains in the plane. It follows that there is no gap for the type of domains we are considering, that is, $\lambda_* = \lambda^*$. Another consequence of the above mentioned result is the fact that $u(x, y, t, \lambda^*)$ is uniformly bounded away from -1. The argument in the proof of Theorem 3.2 guarantees that $u(x, y, t, \lambda^*)$ converges to a stationary solution as $t \to \infty$. Thus, we have proved the following.

THEOREM 3.5. *There exists a critical value λ^* such that the solution $u(x, y, t, \lambda)$ of the parabolic equation* (1) *converges as $t \to \infty$ to a stationary solution if $\lambda \le \lambda^*$. For $\lambda > \lambda^*$, touchdown in finite time occurs.*

If we define T^* as the touchdown time, we can bound T^* from above. Assume λ is such that touchdown occurs; then

$$T^* \le \int_{-1}^{0} \frac{ds}{-\mu_0 s + \frac{\lambda}{(1+s)^2}}. \tag{24}$$

We can also bound T^* from below. Suppose $v(t)$ solves

$$v_t = -\frac{\lambda}{(1+v)^2} \tag{25}$$

with

$$v(0) = 0; \tag{26}$$

then v is a lower solution for our problem. But, v touches down at time $t = \frac{1}{3\lambda}$, and hence

$$\frac{1}{3\lambda} \le T^* \le \int_{-1}^{0} \frac{ds}{-\mu_0 s + \frac{\lambda}{(1+s)^2}}. \tag{27}$$

3.3. Self-similarity and asymptotics of touchdown. For convenience, we restrict ourselves to a one-dimensional domain and consider the new dependent variable $w = 1 + u$, which is a solution of

$$w_t = w_{xx} - \frac{\lambda}{w^2}. \tag{28}$$

Following Giga and Kohn [7], we analyze the structure of solutions near touchdown by means of self-similar variables $y = \frac{x}{\sqrt{T-t}}$, $\tau = \ln\left(\frac{1}{T-t}\right)$, and

$$v(y, \tau) = (T-t)^{-1/3} w(x, t) = e^{\tau/3} w(e^{-\tau/2} y; T - e^{-\tau}); \tag{29}$$

then $v(y, \tau)$ is a solution of

$$v_\tau = v_{yy} - \frac{y}{2} v_y + \frac{1}{3} v - \frac{\lambda}{v^2} \tag{30}$$

in the region $|y| \le \frac{1}{2} e^{\tau/2}, \tau \ge \tau_0 := \ln\left(\frac{1}{T}\right)$, with boundary conditions $v\left(\pm\frac{1}{2} e^{\tau/2}, \tau\right) = e^{\tau/3}$. The initial condition for v is $v(y, \tau_0) = T^{-1/3}$.

It follows that $v(y, \tau)$ is an even function of y for each $\tau \ge \tau_0$.

Generically, $v(y, \tau)$ will converge to a stationary solution. However, there is a continuum of steady states.

The behavior as $y \to \infty$ of stationary solutions as well as the self-similar nature of touchdown for (28) has been described by Fila and Hulshof [3], under the assumption of the existence of nonconstant global solutions. We prove the existence of a continuum of even stationary global solutions.

We study the problem of existence of even stationary solutions, as our solution of (30) is an even function of y.

Thus, we consider solutions of

$$v'' - \frac{y}{2}v' + \frac{1}{3}v - \frac{\lambda}{v^2} = 0 \tag{31}$$

in $y \geq 0$ with $v(0) = \alpha > 0, v'(0) = 0$.

THEOREM 3.6. *For each $\alpha > 0$, the solution of the initial value problem above is defined for all $y \in [0, \infty)$.*

Proof. Assume that $v(y)$ is a solution defined for $y \in [0, y_0]$ and satisfying $v(y) > 0$ in that interval.

Consider the energy

$$E(y) := \frac{v'^2(y)}{2} + \frac{v^2(y)}{6} + \frac{\lambda}{v(y)}; \tag{32}$$

then $\frac{dE}{dy} = \frac{y}{2}v'^2(y) \leq yE(y)$. Therefore $E(y) \leq E(0)e^{y^2/2}$ with $E(0) = \frac{\alpha^2}{6} + \frac{\lambda}{\alpha}$. It follows that for $y \in [0, y_0]$,

$$\frac{\lambda}{E(0)}e^{-y^2/2} \leq v(y) \leq \sqrt{6E(0)}e^{y^2/4}. \tag{33}$$

This estimate proves the theorem. □

The next step is to get bounds on our solution $v(y, \tau)$.

LEMMA 3.1. *There exist positive constants c_j, for $j = 1, 2, 3$, such that*

$$c_3 \leq v(y, \tau) \leq c_1 + c_2|y| \tag{34}$$

for all y and τ in the domain of (30).

Proof. Choose $c_3 < T^{-1/3}$; then c_3 is a lower solution for the parabolic initial-boundary problem defined by (30) and our solution.

The upper bound for v is obtained as follows: choose c_1 in such a way that $c_1^3 < 3\lambda$; then

$$f(y) := \frac{c_2}{6}y - \frac{c_1}{3} + \frac{\lambda}{(c_1 + c_2 y)^2}$$

satisfies $f(0) > 0$, $f'(0) < 0$, it is convex, and it achieves a positive minimum, provided $c_2 > 4\lambda$. Now it is easy to verify that $c_1 + c_2 y$ satisfies the appropriate differential inequality for (30) in the part of the domain corresponding to $y > 0$. The inequality for the boundary values is achieved as long as $c_1 > 0$ and $c_2 > e^{-\tau_0/6}$. For the initial values it is enough to require that $c_1 \geq T^{-1/3}$, which is consistent with the estimate (27). It now follows that the upper bound holds in the part of the domain corresponding to $y > 0$. The upper bound in the entire domain is obtained from the symmetry of the functions involved in the comparison.

The results of Fila and Hulshof guarantee that stationary solutions are increasing and convex for $\alpha < (3\lambda)^{1/3}$. If the inequality is reversed, stationary solutions have

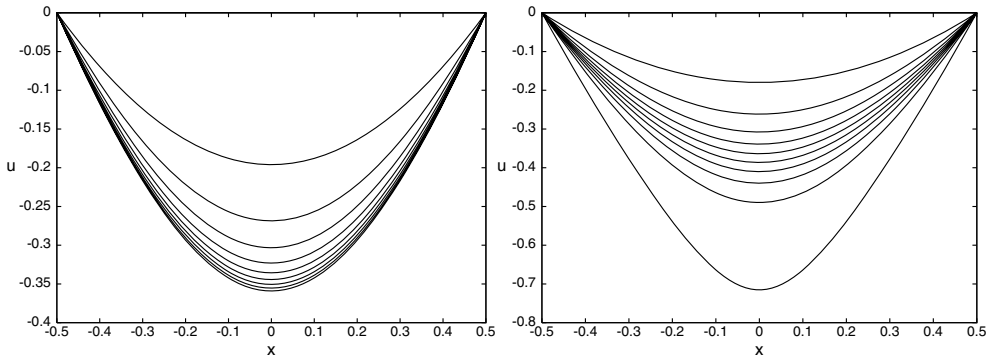


FIG. 1. *One dimension,* $\lambda = 1.4$*,* $t = 2$*, and* $\lambda = 1.45$*,* $t = 1.6$*.*

a single positive minimum in $(0, \infty)$. After that positive minimum, the solutions are increasing and convex.

By using variations of constants in (31), it is easy to see that a nonconstant stationary solution has a derivative that grows like $e^{y^2/4}$ as $y \to \infty$. This shows that our solution $v(y, \tau)$ converges to the constant stationary solution $(3\lambda)^{1/3}$, the convergence being uniform in bounded intervals of y. □

In summary, we have the following.

THEOREM 3.7. *Let* $u(x, t)$ *be the solution of* (5), (6), *and* (7)*; then*

$$u(x,t) = -1 + [3\lambda(T-t)]^{1/3}(1 + o(1)) \quad \text{as } t \to T. \tag{35}$$

The asymptotics is valid in the parabolic regions defined by $|y| \le C$*.*

Amplification. The asymptotic rate of touchdown obtained in Theorem 3.7 is valid if Ω is the unit disk. The same method of proof works since the only difference with respect to (31) is the extra term coming from the radial Laplacian. Thus, Theorem 3.6, Lemma 3.1, and Theorem 3.7 remain valid in a disk.

4. Numerics. We extensively use computing simulations to numerically solve (1) in the following domains: a strip shaped domain, an ellipse, and annular domains. In the one-dimensional case, we also show the validity of our analytical results by comparing numerical and analytical results in several regimes for the value of the parameter λ. The numerical simulation for the solutions of the initial boundary value problem is obtained by means of the Crank–Nicolson scheme, which is implicit and second order in space and time.

To present our numerical results in an accessible manner, we fixed the values of the parameters $\gamma = 1$, $q = 2$ and $\nu = 0.01$ and $\rho = 10$. Then we analyzed different regimes governed by the relative size of the dimensionless parameters λ. In particular we studied the regimes determined by $\lambda < \lambda^*$ and $\lambda > \lambda^*$, where the value of $\lambda^* = 1.400016469$ was obtained for a strip shaped domain in section 2.

4.1. One dimension, $\lambda < \lambda^*$ regime. Our numerical simulations confirm what our analytical results have shown above. As a solution of (1), we observe the generation of a smooth function on time and space with positive concavity which approximates the steady state solution as time goes to infinity. This is illustrated in Figure 1, where the zero initial condition evolves in the way just described. The value of λ for this particular example is 1.4. Note that we are very close to the critical value. The calculation ran up to time $T = 2$.

4.2. One dimension, $\lambda > \lambda^*$ regime. Here we used $\lambda = 1.45$, and the simulation went up to time $T = 1.6$. In Figure 1 we see clearly that at this time the solution is approaching the value -1 at $x = 0$. The solution $u(x,t)$ decreases in time up to what we have called touchdown time T_* when the solution ceases to be smooth. Due to the positive concavity of u, the local minimum also occurs at $x = 0$ where $\inf(u)$ is found. When the function reaches the value of -1 first at $x = 0$ the solution ceases to be continuously smooth and becomes a continuous function with a peak developed at this point. This occurs at a finite time that we called the "the touchdown time T_*." For $t \geq T_*$ we observe a smooth spatial profile, suggesting the existence of a weak solution after touchdown. See Figure 2, where the solution is plotted at time $t = 1.611$.

4.3. Two-dimensional geometries. We have also performed numerical calculation of the solutions for annular regions and ellipses. In the case of an ellipse, the solutions touch down at the center of the ellipse. In this case, the ellipse was transformed into a circle; then the most basic scheme for the heat equation was used, with first-order forward differences for the time derivative and second-order, centered differences for the space derivatives.

The graphs in Figure 3 correspond to the slice $y = 0$ and show the behavior of solutions for the ellipse with $a = 1$, $b = 1/2$. The values of λ are 1.3 and 1.8.

In the case of an annular region with outer radius 1, the solutions touch down at

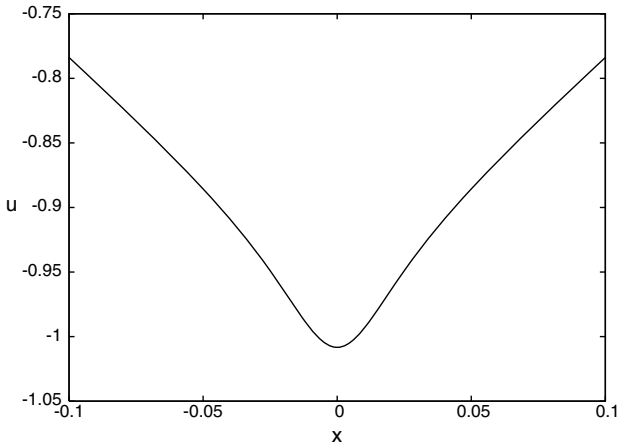


FIG. 2. *One dimension,* $\lambda = 1.45$, $t = 1.611$.

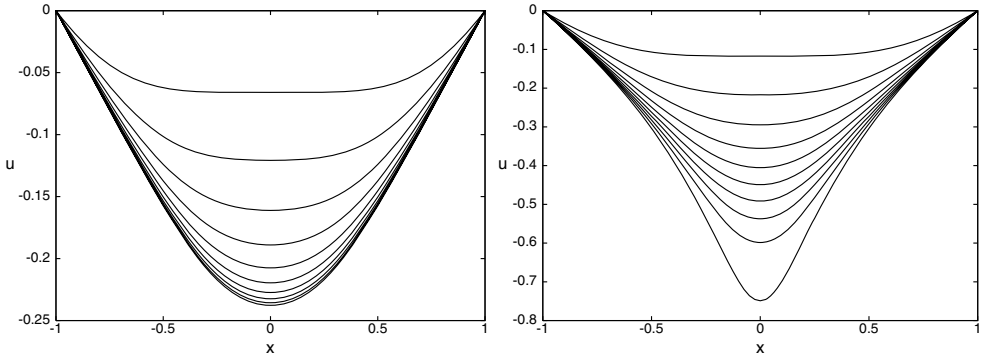


FIG. 3. *Ellipse,* $a = 1$, $b = .5$, $\lambda = 1.3$, $t = .5$, *and* $\lambda = 1.8$, $t = .63$.

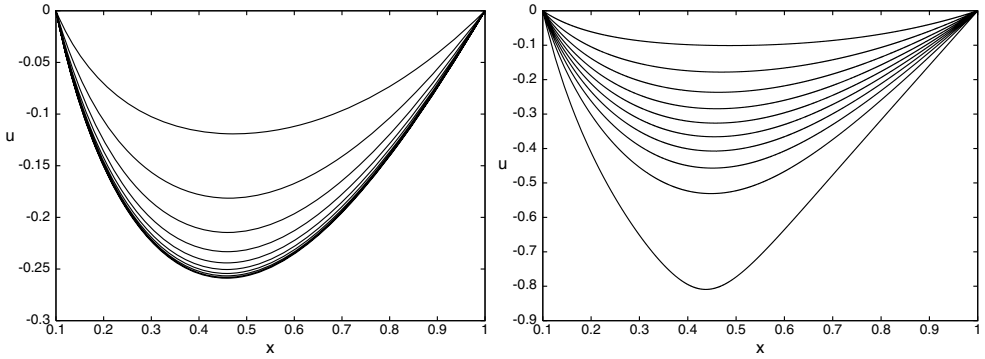


FIG. 4. *Annulus,* $r = .1$*,* $R = 1$*,* $\lambda = 1.4$*,* $t = 1$*, and* $\lambda = 1.8$*,* $t = .57$*.*

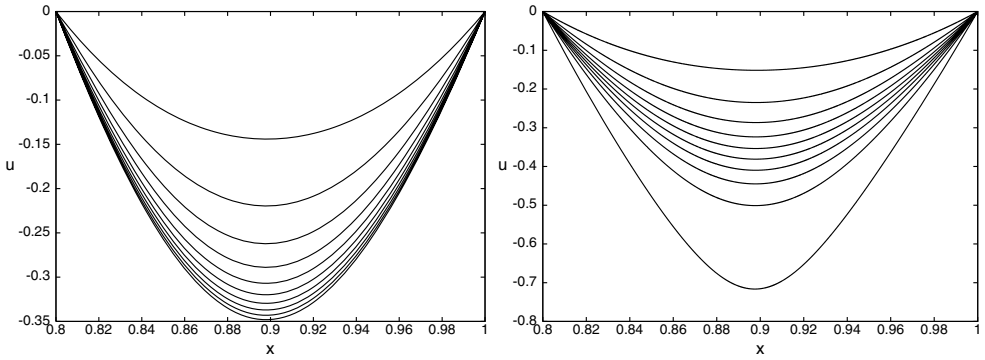


FIG. 5. *Annulus,* $r = .8$*,* $R = 1$*,* $\lambda = 35$*,* $t = .05$*, and* $\lambda = 37$*,* $t = .048$*.*

a circle, whose radius is an increasing function of the inner radius. In this case, we use the Crank-Nicolson method, as the problem is really one-dimensional.

It is interesting to observe the variation of the critical value λ^*: if the inner radius is .1, then $1.4 < \lambda^* < 1.8$. This is illustrated in Figure 4.

If the inner radius is .8, then $35 < \lambda^* < 37$, which resembles Corollary 2.1: thin annular regions have a very large critical value. See Figure 5.

Acknowledgments. We are grateful to A. Minzoni for pointing out the asymptotic behavior of stationary self-similar solutions and to A. Olvera for the numerical verification of this fact. We are also grateful to an anonymous referee for several helpful comments.

REFERENCES

[1] D. BERNSTEIN, P. GUIDOTTI, AND J. A. PELESKO, *Analytical and numerical analysis of electrostatically actuated MEMS devices*, in Proceedings of the International Conference on Modeling and Simulation of Microsystems (1), San Diego, CA, 2000,pp.489–492 (MSM 2000).

[2] K. A. BRAKKE, *The Motion of a Surface by Its Mean Curvature*, Princeton University Press, Princeton, NJ, 1978.

[3] M. FILA AND J. HULSHOF, *A note on the quenching rate*, Proc. Amer. Math. Soc., 112 (1991), pp. 473–477.

[4] M. FILA, B. KAWOHL, AND H. LEVINE, *Quenching for quasilinear equations*, Comm. Partial Differential Equations, 17 (1992), pp. 593–614.

[5] M. Fila and B. Kawohl, *Is quenching in infinite time possible?*, Quart. Appl. Math., 48 (1990), pp. 531–534.
[6] G. Flores, G. Mercado, and J. Pelesko, *Dynamics and touchdown in electrostatic MEMS*, Proceedings of ASME DETC'03 Chicago, IL, 2003, pp. 1–8.
[7] Y. Giga and R. Kohn, *Asymptotically self-similar blow-up for semilinear heat equations*, Comm. Pure Appl. Math., 38 (1985), pp. 297–319.
[8] M. Grayson, *A short note on the evolution of surfaces via mean curvature*, Duke Math. J., 58 (1989), pp. 555–558.
[9] Y. Guo, Z. Pan, and M. J. Ward, *Touchdown and pull-in voltage behavior of a MEMS device with varying dielectric properties*, SIAM J. Appl. Math, 66 (2005), pp. 309–338.
[10] H. Kawarada, *On solutions of the initial value problem for* $u_t = u_{xx} + \frac{1}{1-u}$, Publ. Res. Inst. Math. Sci., 10 (1975), pp. 729–736.
[11] H. Levine, *Quenching, nonquenching and beyond quenching*, Ann. Mat. Pura Appl. (4), 155 (1989), pp. 243–260.
[12] H. C. Nathanson, W. E. Newell, R. A. Wickstrom, and J. R. Davis, *The resonant gate transistor*, IEEE Trans. Electron. Devices, 14 (1967), pp. 117–133.
[13] J. A. Pelesko, *Mathematical modeling of electrostatic MEMS with tailored dielectric properties*, SIAM J. Appl. Math., 62 (2002), pp. 888–908.
[14] J. A. Pelesko and X. Y. Chen, *Electrostatic deflections of circular elastic membranes*, J. Elec., 57 (2003), pp. 1–12.
[15] J. A. Pelesko and A. A. Triolo, *Nonlocal problems in MEMS device control*, J. Engrg. Math., 41 (2001), pp. 345–366.
[16] G. I. Taylor, *The coalescence of closely spaced drops when they are at different electric potentials*, Proc. Roy. Soc. Ser. A, 306 (1968), pp. 423–434.

SIAM J. Appl. Math.
Vol. 67, No. 2, pp. 447–463

BOUNDARY-ROUGHNESS EFFECTS IN NEMATIC LIQUID CRYSTALS*

PAOLO BISCARI† AND STEFANO TURZI†

Abstract. We study the equilibrium configuration of a nematic liquid crystal bounded by a rough surface. The wrinkling of the surface induces a partial melting in the degree of orientation. This softened region penetrates the bulk up to a length scale which turns out to coincide with the characteristic wavelength of the corrugation. Within the boundary layer where the nematic degree of orientation decreases, the tilt angle steepens and gives rise to a nontrivial structure, which may be interpreted in terms of an effective weak anchoring potential. We determine how the effective surface extrapolation length is related to the microscopic anchoring parameters. We also analyze the crucial role played by the boundary conditions assumed on the degree of orientation. Quite different features emerge depending on whether they are Neumann- or Dirichlet-like. These features may be useful to ascertain experimentally how the degree of orientation interacts with an external boundary.

Key words. nematic liquid crystals, surface roughness, surface melting, weak anchoring

AMS subject classifications. 76A15, 74A50, 82D30

DOI. 10.1137/060656711

Introduction. Nematic liquid crystals are fluid aggregates of elongated molecules. When the nematic rods interact with an external surface, the elastic strain energy induces them to align parallel to the unit normal, even if the surface is not perfectly flat [1]. Recent experimental observations confirm that the surface alignment of the nematic director is completely determined by the roughness-induced surface anisotropy [2]. Further analytical calculations, performed within the classical Frank model with unequal elastic constants, have detected the bulk effects induced by a periodically molded external boundary [3, 4].

A crucial effect, still related to surface roughness, escapes the framework of Frank theory, where the only order parameter is the director. Indeed, it is physically intuitive that nematic molecules will disorder if we force them to follow a rapidly varying boundary condition. This *surface melting* was first experimentally detected in [5, 6]. Recent experimental observations have also measured a boundary-layer structure in the degree of orientation [7]. The surface melting has been confirmed by approximated analytical solutions [8], numerical calculations [9, 10], and molecular Monte Carlo simulations [11].

The combined effect of a rapidly varying director anchoring and surface melting gives rise to an effective weak-anchoring effect that was first observed in [12]. The problem of relating the effective anchoring extrapolation length to the microscopic roughness parameters has been studied in several theoretical papers, all framed within the Frank theory [13, 14, 15, 16]. This observation is of basic significance, since weak anchoring potentials have proven to influence deeply all nematic phenomena, beginning with Freedericksz transitions [17, 18, 19]. Indeed, several theoretical studies have already determined the influence on anchoring energies of the presence of permanent surface dipoles [20] or diluted surface potentials [21, 22].

*Received by the editors April 7, 2006; accepted for publication (in revised form) October 19, 2006; published electronically February 2, 2007.
http://www.siam.org/journals/siap/67-2/65671.html
†Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy (paolo.biscari@polimi.it, stefano.turzi@mate.polimi.it).

In this paper we analyze in analytical detail the boundary-layer structure induced by a rough surface which bounds a nematic liquid crystal. We frame within the Landau–de Gennes order-tensor theory, to be able to detect the effects on both the director and the degree of orientation. Our results confirm the surface melting already obtained in [8] but allow us to detect new phenomena. First, the nematic director steepens close to the boundary, so giving rise to an effective weak anchoring potential, which turns out to be deeply related to the surface-melting effect, and thus can be correctly handled only within the order-tensor theory. Furthermore, the boundary layers display a strong dependence on the type of boundary conditions imposed on the degree of orientation. Indeed, the orders of magnitude of all the expected effects depend on whether the boundary conditions are Neumann- or Dirichlet-like. We discuss how these effects may help in ascertaining in experiments how the mesoscopic parameter, which measures the degree of order, interacts with an external surface.

The paper is organized as follows. In section 1 we present the model, we set the geometry of a roughly bounded sample, and derive the Euler–Lagrange partial differential equations that determine the equilibrium configurations. In section 2 we perform the perturbative two-scale analysis that provides all the analytical details of the boundary-layer structure. In section 3 we solve an effective problem, in which the rough surface is replaced by a weak-anchoring potential. As a result, we show that a weak-anchoring potential may be given a microscopic interpretation, and we relate the surface extrapolation length to the microscopic roughness parameters. In section 4 we draw our conclusions and discuss the validity of our geometric approximations. Two appendices collect the technical details of some lengthy calculations.

1. Equilibrium configurations. The degree of order decrease has been recognized by many authors as a crucial effect of surface roughness [8, 10]. We thus describe nematic configurations in the framework of the Landau–de Gennes $\mathbf{Q}$-tensor theory [23]. The order tensor is defined as the trace-free part of the second moment of the probability distribution of molecular orientations:

$$\mathbf{Q}(\mathbf{r}) := \int_{\mathbb{S}^2} (\mathbf{m} \otimes \mathbf{m})\, f_r(\mathbf{m})\, da - \frac{1}{3}\,\mathbf{I}\,, \tag{1.1}$$

where $\mathbf{I}$ denotes the identity tensor. $\mathbf{Q}$ is a second-order traceless symmetric tensor, with $\operatorname{sp}\mathbf{Q} \subset \left[-\frac{1}{3}, \frac{2}{3}\right]$ [18].

In order to keep computations simple, we adopt the one-constant approximation for the elastic part of the free-energy functional

$$f_{\mathrm{el}}[\mathbf{Q}] = \frac{1}{2}\, K\, |\nabla \mathbf{Q}|^2\,, \tag{1.2}$$

where K is an average elastic constant. We stress, however, that it is straightforward to generalize all what follows to take into account unequal material elastic constants.

The free-energy functional includes the Landau–de Gennes thermodynamic potential as well:

$$f_{\mathrm{LdG}}(\mathbf{Q}) = A \operatorname{tr}\mathbf{Q}^2 - B \operatorname{tr}\mathbf{Q}^3 + C \operatorname{tr}\mathbf{Q}^4\,. \tag{1.3}$$

The material parameter A depends on the temperature, and in particular it becomes negative deep in the nematic phase. By contrast, B, C can be assumed to be positive and temperature-independent. The potential (1.3) strongly favors uniaxial phases, in which at least two of the three eigenvalues of $\mathbf{Q}$ coincide. In fact, $\mathbf{Q}$ is expected

to abandon uniaxiality mainly close to director singularities [24, 25, 26]. We will not deal with any defect structure. Thus, though the uniaxiality constraint is not essential for our purposes, we follow the attitude of avoiding unnecessary complications and restrict our attention to uniaxial states

$$\mathbf{Q}(\mathbf{r}) = s(\mathbf{r}) \left(\mathbf{n}(\mathbf{r}) \otimes \mathbf{n}(\mathbf{r}) - \frac{1}{3}\mathbf{I} \right) . \tag{1.4}$$

We stress that we are not claiming that biaxiality effects are absent close to an external surface, since indeed the converse holds [27, 28, 29]. However, our results show that, even in the absence of biaxiality, a surface melting exists and an effective weak anchoring arises. A detailed treatment of the complete order-tensor theory would yield more precise quantitative estimates of the effects we will determine anyhow.

The scalar $s \in \left[-\frac{1}{2}, 1\right]$ and the unit vector $\mathbf{n}$ in (1.4) are, respectively, the *degree of orientation* and the *director*. With the aid of (1.4), the potentials (1.2), (1.3) can be written as

$$f_{\rm el}[s,\mathbf{n}] = K\left(s^2|\nabla\mathbf{n}|^2 + \tfrac{1}{3}|\nabla s|^2\right) \quad \text{and} \quad f_{\rm LdG}(s) = \tfrac{2}{3}A\,s^2 - \tfrac{2}{9}B\,s^3 + \tfrac{2}{9}C\,s^4 . \tag{1.5}$$

When $A \le B^2/(12C)$, the absolute minimum of the function $f_{\rm LdG}(s)$ occurs at the *preferred degree of orientation*

$$s_{\rm pr} := \frac{3B + \sqrt{9B^2 - 96AC}}{8C} > 0 . \tag{1.6}$$

In order to gain physical interpretation of the results, we also introduce the *nematic coherence length* ξ and the dimensionless (positive) parameter ω, defined as

$$\xi^2 := \frac{9K}{C} \qquad \text{and} \qquad \omega^2 := \frac{6}{C}\left(s_{\rm pr}B - 4A\right) . \tag{1.7}$$

The nematic coherence length compares the strength of the elastic and thermodynamic contributions to the free-energy functional. It characterizes the size of the domains where the degree of orientation may abandon its preferred value $s_{\rm pr}$. The number ω depends on the depth of the potential well associated with $s_{\rm pr}$. Indeed, it is defined in such a way that $f''_{\rm LdG}(s_{\rm pr}) = K\omega^2/\xi^2$.

By using (1.6), (1.7) we write the bulk free-energy density $f_{\rm b} := f_{\rm el} + f_{\rm LdG}$ as

$$\frac{f_{\rm b}[s,\mathbf{n}]}{K} = s^2|\nabla\mathbf{n}|^2 + \frac{1}{3}|\nabla s|^2 + \frac{1}{\xi^2}\left(s^4 - \frac{4}{3}s^3\left(2s_{\rm pr} - \frac{\omega^2}{s_{\rm pr}}\right) + 2s^2(s_{\rm pr}^2 - \omega^2)\right) . \tag{1.8}$$

1.1. Modeling a rough surface. We aim at analyzing the effects that a rough boundary induces in a nematic liquid crystal. Once again, we try to keep our analysis as simple as possible, while still catching the essential features. We thus follow, e.g., [13] in modeling roughness by imposing a sinusoidally perturbed homeotropic anchoring condition on a flat surface (see Figure 1.1). The amplitude and the wavelength characterizing the perturbation will be the crucial parameters in our results.

There are two nontrivial simplifications in our geometric setting. First, we are assuming that the boundary is perfectly sinusoidal, while a physical surface will clearly exhibit a whole roughness spectrum. Second, we are replacing an undulating boundary by an undulating boundary condition on a flat surface. We postpone until the final section a more detailed discussion on the validity of these simplifying assumptions. We anticipate, however, that none of them introduces qualitative errors. More precisely,

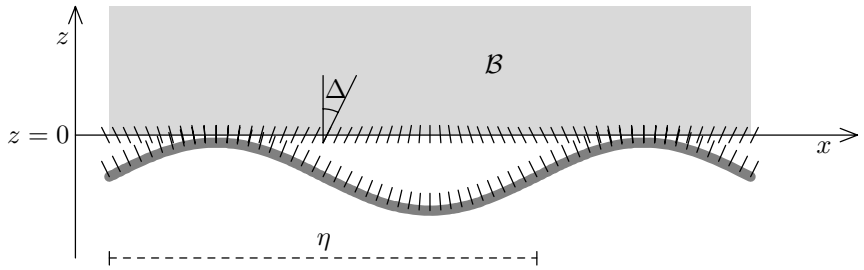


FIG. 1.1. *Geometric modeling of a rough surface. The physical surface oscillates with a characteristic wavelength η. The homeotropic anchoring at the oscillating boundary induces an oscillation of amplitude Δ in the boundary director. The bulk volume $\mathcal{B}$ is the grey region. Besides the microscopic roughness wavelength η, the two-scale analysis performed below is governed as well by the microscopic nematic coherence length ξ, introduced in* (1.7).

the latter amounts to performing an expansion in the roughness amplitude and keeping the leading order in the expansion. As for the treatment of the whole roughness spectrum, it turns out that at the same order of approximation one is allowed to treat a single wavelength at a time and eventually add up all the contributions.

We focus attention on a thin boundary layer, attached to the external surface. Consequently, we disregard the detailed structure of the bulk equilibrium configuration, which will enter our results only as asymptotic *outer* solution for the surface boundary layer. We introduce a Cartesian frame of reference $\{\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\}$ and assume that the nematic spreads in the whole half-space $\mathcal{B} = \{z \geq 0\}$. We further simplify the geometry by assuming that $\mathbf{n}(\mathbf{r}) = \sin\theta(\mathbf{r})\,\mathbf{e}_x + \cos\theta(\mathbf{r})\,\mathbf{e}_z$ and that the asymptotic bulk configuration depends only on z:

$$\theta(\mathbf{r}) \approx \theta_{\mathrm{b}}(z) \qquad \text{as} \quad z \to +\infty\,. \tag{1.9}$$

A crucial role in the developments below is played by $\mathrm{m} := \theta_{\mathrm{b}}'(0)$, the derivative of the asymptotic solution at $z = 0$, which has the physical dimensions of an inverse length. It represents the effect of any tilted bulk field that competes with the surface anchoring and may well be an electric or magnetic coherence length. Throughout our calculations we will assume that m^{-1} is much greater than both the nematic coherence length and the roughness wavelength. In the presence of an external field so strong that m^{-1} becomes of the order of, or even smaller than, the microscopic lengths above, the following asymptotic expansions fail. In particular, in this extreme regime the surface roughness effects may invade the whole bulk and cannot be replaced by an effective weak-anchoring potential.

In the presence of strong homeotropic anchoring on a flat surface, the boundary condition to be imposed on the director would be $\theta^{(\mathrm{flat})}(x, y, 0) = 0$. On the contrary, we will require

$$\theta(x, y, 0) = \Delta \cos\frac{x}{\eta}\,. \tag{1.10}$$

The boundary condition (1.10) mimics the rugosity of the external surface by introducing two new parameters: the (dimensionless) roughness amplitude Δ and the roughness wavelength η (see Figure 1.1). We remark that the oscillation rate increases as $\eta \to 0^+$, while all roughness effects are expected to vanish in the limit $\Delta \to 0^+$. The requirements (1.9), (1.10) imply that the free-energy minimizer will not exhibit

any dependence on the transverse y-coordinate, so that we will henceforth restrict our attention to the dependence on the coordinates (x, z).

It is more complex to ascertain the correct type of boundary conditions which are to be imposed on the degree of orientation s. From the mathematical point of view, it would be natural to imitate the (Dirichlet) strong anchoring imposed on the director and thus set $s(x, y, 0)$ to be equal to some fixed boundary value $\tilde{s}$. Nevertheless, while it is well established that we can induce an easy axis for the director on an external boundary, it is questionable whether we may fix the value of a mesoscopic parameter, which measures the degree of order in a distribution. From the physical point of view, stress-free (Neumann) boundary conditions on the degree of orientation deserve attention as well. In this latter case, we simply leave to the thermodynamic potential (1.3) the assignment of inducing the preferred value $s_{\rm pr}$ in the bulk ($z \to \infty$), while we perform no boundary action on the degree of orientation. To be safe, both possibilities (Dirichlet and Neumann) will be analyzed in section 2.

1.2. Euler–Lagrange equations. Once we consider that $|\nabla \mathbf{n}|^2 = |\nabla \theta|^2$, it is easy to derive the Euler–Lagrange partial differential equations associated with the functional (1.8). They read

$$s^2 \Delta\theta + 2\, s\, \nabla s \cdot \nabla\theta = 0 \qquad \text{and} \qquad \Delta s - 3\, s\, |\nabla\theta|^2 - 3\, \frac{\sigma(s)}{\xi^2} = 0\,, \tag{1.11}$$

where

$$\sigma(s) := s(s - s_{\rm pr}) \left(s - s_{\rm pr} + \frac{\omega^2}{s_{\rm pr}} \right) . \tag{1.12}$$

Since the boundary conditions (1.10) are x-periodic, with a period of $2\pi\eta$, we look for solutions of (1.11) in $C^2_{2\pi\eta}$ (the space of C^2-functions, $2\pi\eta$-periodic in the x-direction). To complete the differential system (1.10), in section 2.1 we will require

$$\begin{cases} \theta(x,0) = \Delta \cos \dfrac{x}{\eta} \\ \dfrac{\partial s}{\partial z}(x,0) = 0 \end{cases} \qquad \text{and} \qquad \begin{cases} \theta(x,z) \approx \theta_{\rm b}(z) \\ s(x,z) \approx s_{\rm pr} \end{cases} \qquad \text{as } z \to \infty\,, \tag{1.13}$$

while in section 2.2 we will choose

$$\begin{cases} \theta(x,0) = \Delta \cos \dfrac{x}{\eta} \\ s(x,0) = \tilde{s} \end{cases} \qquad \text{and} \qquad \begin{cases} \theta(x,z) \approx \theta_{\rm b}(z) \\ s(x,z) \approx s_{\rm pr} \end{cases} \qquad \text{as } z \to \infty\,. \tag{1.14}$$

2. Two-scale analysis. Before proceeding with the perturbation analysis of the differential equations, we state them in dimensionless form. It will turn out that the correct scaling is obtained by measuring lengths in η-units, so that we introduce the new dimensionless coordinates $\bar{x} = x/\eta$, $\bar{z} = z/\eta$ and define the dimensionless parameter $\varepsilon = \xi/\eta$. Equations (1.11) thus become

$$s^2 \Delta\theta + 2s\, \nabla s \cdot \nabla\theta = 0 \qquad \text{and} \qquad \varepsilon^2 \Delta s - 3\varepsilon^2 s\, |\nabla\theta|^2 - 3\sigma(s) = 0\,, \tag{2.1}$$

where both the gradient and the Laplacian are now to be intended with respect to the scaled variables. The nematic coherence length is usually much smaller than all other characteristic lengths. Consequently, we will look for uniformly asymptotic

solutions to (2.1), by treating ε as a small parameter. In this limit, $(2.1)_2$ is singular, so that a regular asymptotic expansion would not provide a uniform approximation of the solution. Indeed, the small parameter ε multiplies the highest derivative, so that we may expect the solution to have a steep behavior in a layer of thickness δ (to be determined), close to the boundary $z = 0$. We refer the reader to the books [30, 31, 32, 33] for the details of the singular perturbation theory we will apply henceforth and, in particular, for the technique of the two-scale method which directly yields a uniform approximation of the solution.

A standard dominant balance argument (that requires us to introduce a stretched variable $Z = \bar{z}/\delta$) allows us to recognize that the boundary layer thickness is $\delta = \varepsilon$. We then introduce the *fast* variable $Z = \bar{z}/\varepsilon$. The two-scale chain rule requires us to replace $\partial_{\bar{z}}$ by $\left(\partial_{\bar{z}} + \varepsilon^{-1}\partial_Z\right)$, and equations (2.1) take the form (when $s \neq 0$)

$$s\left(\varepsilon^2\theta_{,\bar{x}\bar{x}} + \varepsilon^2\theta_{,\bar{z}\bar{z}} + 2\varepsilon\theta_{,\bar{z}Z} + \theta_{,ZZ}\right) + 2\varepsilon^2 s_{,\bar{x}}\theta_{,\bar{x}} + 2(\varepsilon s_{,\bar{z}} + s_{,Z})(\varepsilon\theta_{,\bar{z}} + \theta_{,Z}) = 0\,, \tag{2.2}$$

$$\varepsilon^2 s_{,\bar{x}\bar{x}} + \varepsilon^2 s_{,\bar{z}\bar{z}} + 2\varepsilon s_{,\bar{z}Z} + s_{,ZZ} - 3s\left[\varepsilon^2(\theta_{,\bar{x}})^2 + (\varepsilon\theta_{,\bar{z}} + \theta_{,Z})^2\right] - 3\sigma(s) = 0\,, \tag{2.3}$$

where a comma denotes differentiation with respect to the indicated variable. In agreement with the two-scale method, θ and s are now to be intended as $\theta(\bar{x}, \bar{z}, Z)$ and $s(\bar{x}, \bar{z}, Z)$. In other words, θ and s are functions of $\bar{x}$, $\bar{z}$, and Z, which are to be regarded as *independent* variables. It will be only at the very end of our calculations that we will recast the connection between $\bar{z}$ and Z: $Z = \bar{z}/\varepsilon$. We seek solutions which may be given the formal expansions

$$\theta(\bar{x}, \bar{z}, Z) = \theta_0(\bar{x}, \bar{z}, Z) + \varepsilon\theta_1(\bar{x}, \bar{z}, Z) + \varepsilon^2\theta_2(\bar{x}, \bar{z}, Z) + O(\varepsilon^3)\,, \tag{2.4}$$

$$s(\bar{x}, \bar{z}, Z) = s_0(\bar{x}, \bar{z}, Z) + \varepsilon s_1(\bar{x}, \bar{z}, Z) + \varepsilon^2 s_2(\bar{x}, \bar{z}, Z) + O(\varepsilon^3)\,. \tag{2.5}$$

If we insert (2.4)–(2.5) into (2.2)–(2.3), we obtain the following differential equations to $\mathcal{O}(1)$, $\mathcal{O}(\varepsilon)$, and $\mathcal{O}\left(\varepsilon^2\right)$:

$$\begin{cases} \frac{1}{s_0}\left(s_0^2\theta_{0,Z}\right)_{,Z} = 0\,, \\ s_{0,ZZ} - 3s_0(\theta_{0,Z})^2 - 3\sigma(s_0) = 0\,, \end{cases} \tag{2.6}$$

$$\begin{cases} \frac{1}{s_0}\left(s_0^2\theta_{1,Z}\right)_{,Z} + \frac{1}{s_1}\left(s_1^2\theta_{0,Z}\right)_{,Z} = -2\left(s_0\theta_{0,Z}\right)_{,\bar{z}} - 2s_{0,Z}\theta_{0,\bar{z}}\,, \\ s_{1,ZZ} - 6s_0\theta_{0,Z}\theta_{1,Z} - 3s_1\left(\sigma'(s_0) + (\theta_{0,Z})^2\right) = 6s_0\theta_{0,Z}\theta_{0,\bar{z}} - 2s_{0,\bar{z}Z}\,, \end{cases} \tag{2.7}$$

$$\begin{cases} \frac{1}{s_0}(s_0^2\theta_{2,Z})_{,Z} + \frac{1}{s_2}(s_2^2\theta_{0,Z})_{,Z} = -\frac{1}{s_1}(s_1^2\theta_{1,Z})_{,Z} - \frac{1}{s_0}(s_0^2\theta_{0,\bar{z}})_{,\bar{z}} - \frac{1}{s_0}(s_0^2\theta_{0,\bar{x}})_{,\bar{x}} \\ \qquad\qquad - 2\left(s_0\theta_{1,Z}\right)_{,\bar{z}} - 2\left(s_1\theta_{0,Z}\right)_{,\bar{z}} - 2s_{1,Z}\theta_{0,\bar{z}} - 2s_{0,Z}\theta_{1,\bar{z}}\,, \\ s_{2,ZZ} - 3s_2\left[\sigma'(s_0) + (\theta_{0,Z})^2\right] - 6s_0\theta_{0,Z}\theta_{2,Z} = \frac{3}{2}s_1^2\sigma''(s_0) \\ \qquad + 3s_0\left[(\theta_{0,\bar{z}} + \theta_{1,Z})^2 + (\theta_{0,x})^2\right] \\ \qquad + 6\theta_{0,Z}\left(s_1\theta_{1,Z} + s_1\theta_{0,\bar{z}} + s_0\theta_{1,\bar{z}}\right) - 2s_{1,\bar{z}Z} - s_{0,\bar{z}\bar{z}} - s_{0,xx}\,. \end{cases} \tag{2.8}$$

Analogous equations can be easily derived at any desired order. For any $n \geq 1$, the differential system obtained at $\mathcal{O}(\varepsilon^n)$ is linear in the unknowns θ_n, s_n and may be solved analytically. By virtue of the multiscale method, we find the correct dependence on $\bar{z}, Z$ by requiring that all s_n, θ_n are uniformly bounded as $\varepsilon \to 0^+$ for expanding intervals of the type $0 \leq Z \leq Z^*/\varepsilon$, for a suitable positive constant Z^*. In most practical cases this requirement is equivalent to asking for the removal of secular terms (i.e., terms that diverge as $Z \to +\infty$).

2.1. Free surface degree of orientation. In terms of the scaled variables, the boundary conditions (1.13) require

$$\begin{cases} \theta(\bar{x},0) = \Delta\cos\bar{x} \\ s_{,\bar{z}}(x,0) = 0 \end{cases} \qquad \text{and} \qquad \begin{cases} \theta(\bar{x},\bar{z}) \approx \theta_{\mathrm{b}}(\eta\bar{z}) \\ s(\bar{x},\bar{z}) \approx s_{\mathrm{pr}} \end{cases} \qquad \text{when } \bar{z} \gg \eta\,. \tag{2.9}$$

The leading solutions in expansions (2.4), (2.5) are

$$s_0(x,z) = s_{\mathrm{pr}} \qquad \text{and} \qquad \theta_0(x,z) = \mathrm{m}\,z + \Delta \mathrm{e}^{-z/\eta}\cos\frac{x}{\eta}\,, \tag{2.10}$$

where $\mathrm{m} := \theta_{\mathrm{b}}'(0)$. Higher-order asymptotic solutions are gathered by means of laborious but straightforward calculations. After recasting the solutions in terms of the dimensional variables $x = \eta\,\bar{x}$ and $z = \eta\,\bar{z}$, we find

$$\begin{aligned} s(x,z) = s_{\mathrm{pr}} &- \frac{s_{\mathrm{pr}}\xi^2}{\omega^2}\left(\mathrm{m}^2 - \frac{2\,\mathrm{m}\,\Delta}{\eta}\,\mathrm{e}^{-z/\eta}\,\cos\frac{x}{\eta} + \frac{\Delta^2}{\eta^2}\,\mathrm{e}^{-2z/\eta}\right) \\ &+ \frac{2s_{\mathrm{pr}}\,\xi^3}{\sqrt{3}\,\omega^3}\,\mathrm{e}^{-\sqrt{3}\omega z/\xi}\left(\frac{\Delta^2}{\eta^3} - \frac{\mathrm{m}\,\Delta}{\eta^2}\cos\frac{x}{\eta}\right) + \mathcal{O}\left(\varepsilon^4\right) \end{aligned} \tag{2.11}$$

and

$$\begin{aligned} \theta(x,z) = \mathrm{m}\,z &+ \Delta\,\mathrm{e}^{-z/\eta}\,\cos\frac{x}{\eta} + \frac{\xi^2}{\omega^2}\left(\frac{2\,\mathrm{m}\,\Delta^2}{\eta}\left(1 - \mathrm{e}^{-2z/\eta}\right)\right. \\ &\left. - \frac{\Delta^3}{2\,\eta^2}\left(\mathrm{e}^{-z/\eta} - \mathrm{e}^{-3z/\eta}\right)\cos\frac{x}{\eta} - \frac{2\,\mathrm{m}^2\,\Delta}{\eta}\,z\,\mathrm{e}^{-z/\eta}\,\cos\frac{x}{\eta}\right) + \mathcal{O}\left(\varepsilon^4\right)\,. \end{aligned} \tag{2.12}$$

The above expansions have been carried out up to the first nontrivial correction of the 0th-order approximation. Indeed, all calculations must be pushed to $\mathcal{O}\left(\varepsilon^3\right)$, since an internal ξ-layer is necessary to satisfy the boundary condition (1.13) in $z = 0$. This layer is of $\mathcal{O}\left(\varepsilon^3\right)$ because in the Neumann case the boundary condition (1.13) concerns the first derivative of s, instead of the degree of orientation itself. We remark that the solutions (2.11)–(2.12) are coherently ordered for every fixed value of $\eta \neq 0$. However, they are not uniformly ordered when $\eta \in (0,\bar{\eta}]$; namely, we do not have a uniform solution if η is allowed to become of order ξ or, still worse, tend to zero. In other words, the above solutions remain valid as $\eta \to 0^+$ if and only if $\xi = o\,(\eta)$. The main properties of the equilibrium configurations in the mathematically appealing but physically uncommon case in which η is of the order of, or even smaller than, ξ will be presented elsewhere [34].

2.1.1. Surface melting. We can highlight three different contributions in the degree of orientation (2.11). First, we notice a uniform decrease in the degree of order, equal to $-s_{\mathrm{pr}}\mathrm{m}^2\xi^2/\omega^2$. This disordering effect is triggered by the θ-derivative m and was certainly to be expected. In fact, a glance to the free-energy functional (1.8) suffices to show that a reduction in s decreases the free energy whenever the gradient of the director is not null. We then find two boundary layers. The former, of thickness η and $\mathcal{O}\left(\varepsilon^2\right)$, is a further reduction of the degree of orientation due to the boundary roughness, which induces a director variation in the x-direction. An internal boundary layer, of thickness ξ and order $\mathcal{O}\left(\varepsilon^3\right)$, is finally needed in order to cancel the normal derivative of s at the external surface. If we take into account all

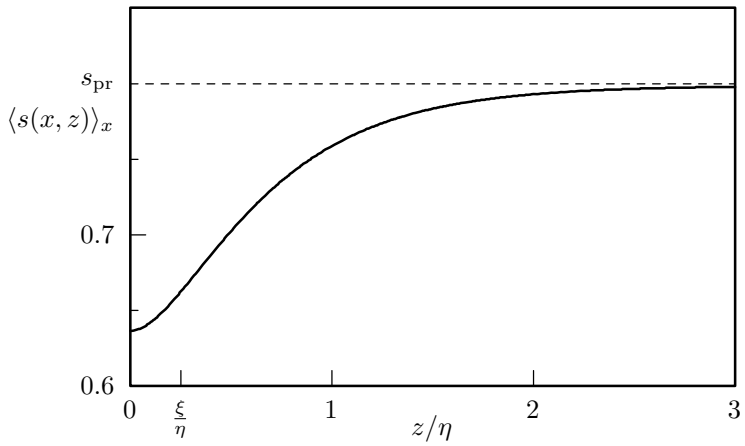


FIG. 2.1. *Boundary layers in the mean degree of orientation* $\langle s(x,z)\rangle_x$ *when* $\xi = 0.25\eta$, $s_{\rm pr} = 0.8$, $\omega = 0.6$, $\mathrm{m} = 0.1/\eta$, *and* $\Delta = 1.5$. *The plot exhibits the presence of* two *boundary layers, the internal one being required by the free boundary condition applied on* s.

the contributions, the mean surface degree of orientation, defined as the x-average of $s(x,0)$, turns out to be

$$\langle s(x,0)\rangle_x = s_{\rm pr}\left[1 - \frac{\mathrm{m}^2\xi^2}{\omega^2} - \frac{\Delta^2\xi^2}{\omega^2\eta^2} + \frac{2\Delta^2\xi^3}{\sqrt{3}\,\omega^3\eta^3}\right] + \mathcal{O}\left(\varepsilon^4\right) \ . \tag{2.13}$$

Figure 2.1 evidences the reported behavior of the mean degree of orientation as a function of the distance from the surface.

2.1.2. Effective surface angle. The tilt angle θ exhibits a boundary-layer structure as well. Equation (2.12) shows that such a layer is of $\mathcal{O}\left(\varepsilon^2\right)$ and thickness η. It gives rise to an interesting effective misalignment of the surface director. Indeed, if we allow $z \gg \eta$ in (2.12) we find that

$$\theta(x,z) \approx \theta_{\rm b}(z) = \frac{2\,\mathrm{m}\,\xi^2\Delta^2}{\eta\omega^2} + \mathrm{m}\,z \qquad \text{as} \quad z \gg \eta \ . \tag{2.14}$$

The asymptotic approximation (2.14) shows that an experimental observation, performed sufficiently far from the external plate (with respect to the microscopic scale η), would detect an *effective* tilt angle $\theta_{\rm b}$, whose value at the plate is different from zero, since

$$\theta_{\rm b}(0) = \frac{2\,\mathrm{m}\,\xi^2\Delta^2}{\eta\omega^2} \ . \tag{2.15}$$

Thus, a coarse observation of the nematic configuration measures a surface tilt angle slightly different from the homeotropic prescription $\theta_{\rm surf} = 0$. Figure 2.2 evidences this effect. In the next section we will analyze in more detail the result (2.15). Then we will show how it matches the predictions of an effective weak-anchoring potential. We remark that the tilt angle does not exhibit any further boundary layer at the smaller scale ξ.

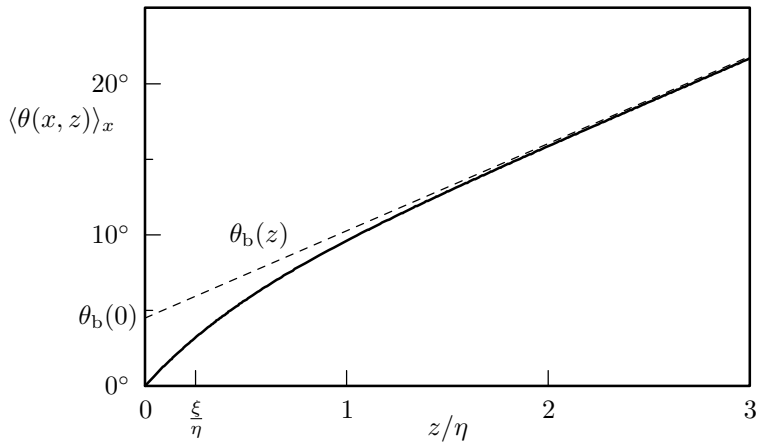


FIG. 2.2. *Boundary layer in the mean tilt angle* $\langle\theta(x,z)\rangle_x$ *when* $\xi = 0.25\eta$, $s_{\rm pr} = 0.8$, $\omega = 0.6$, $\mathrm{m} = 0.1/\eta$, *and* $\Delta = 1.5$. *The dashed line corresponds to the asymptotic, linear approximation* $\theta_{\rm b}(z)$.

2.2. Fixed surface degree of orientation. The perturbative analysis of the differential equations (1.11), with the Dirichlet boundary conditions (1.14), would be unnecessarily entangled because of the nonlinearity of the thermodynamic potential (1.12). In fact, in this case only implicit solutions for $s_0(x,z,Z)$ can be gathered. In order to pursue our analysis, and still catch the essential features of the solutions, we replace the function σ in (1.11) by its linear approximation $\sigma_1(s) = \omega^2(s - s_{\rm pr})$. This is tantamount to replacing the Landau–de Gennes potential in (1.5) by a tangent quadratic well, still centered in $s_{\rm pr}$. Such an approximation is certainly well justified deep in the nematic phase, when the isotropic state $s = 0$ becomes unstable, and the second well of the Landau–de Gennes potential can be neglected.

The asymptotic properties of the solutions in this case depend critically on the value $\tilde{s}$ forced on the surface. If $\tilde{s} \neq s_{\rm pr}$, the boundary layer induced by the Dirichlet condition dominates over the roughness effect. Indeed, the leading asymptotic solutions are given by

$$\begin{aligned} s(x,z) = {} & s_{\rm pr} - (s_{\rm pr} - \tilde{s})\, \mathrm{e}^{-\sqrt{3}\,\omega z/\xi} \\ & - \sqrt{3}\, (s_{\rm pr} - \tilde{s})\, \frac{\xi}{\omega}\, \mathrm{e}^{-\sqrt{3}\,\omega z/\xi} \left[\frac{\Delta^2}{4\eta} \left(1 - \mathrm{e}^{-2z/\eta}\right) + \frac{3}{2}\, \mathrm{m}^2\, z \right. \\ & \left. - 3\mathrm{m}\Delta \left(1 - \mathrm{e}^{-z/\eta}\right) \cos\frac{x}{\eta} + \frac{\Delta^2}{2\eta} \left(1 - \mathrm{e}^{-2z/\eta}\right) \cos\frac{2x}{\eta} \right] + \mathcal{O}\left(\varepsilon^2\right), \end{aligned} \tag{2.16}$$

$$\begin{aligned} \theta(x,z) = {} & \mathrm{m}\, z + \Delta \mathrm{e}^{-z/\eta} \cos\frac{x}{\eta} \\ & + \frac{\xi}{\sqrt{3}\,\omega} \left[h\left(\frac{z}{\xi}\right) - h(0) \right] \left(\mathrm{m} - \frac{\Delta}{\eta} \mathrm{e}^{-z/\eta} \cos\frac{x}{\eta} \right) + \mathcal{O}\left(\varepsilon^2\right), \end{aligned} \tag{2.17}$$

where

$$h(\zeta) = \log\left[s_{\rm pr} - (s_{\rm pr} - \tilde{s})\, \mathrm{e}^{-\sqrt{3}\,\omega\,\zeta} \right] - \frac{(s_{\rm pr} - \tilde{s})\, \mathrm{e}^{-\sqrt{3}\,\omega\,\zeta}}{s_{\rm pr} - (s_{\rm pr} - \tilde{s})\, \mathrm{e}^{-\sqrt{3}\,\omega\,\zeta}} \tag{2.18}$$

determines the tilt angle variation within the boundary layer. The bulk-asymptotic tilt angle is then given by

$$\theta(x,z) \approx \theta_{\mathrm{b}}(z) = \frac{\mathrm{m}\,\xi}{\sqrt{3}\,\omega}\left(\log\frac{s_{\mathrm{pr}}}{\tilde{s}} + \frac{s_{\mathrm{pr}}-\tilde{s}}{\tilde{s}}\right) + \mathrm{m}\,z \qquad \text{as} \quad z \gg \eta\,. \tag{2.19}$$

We remark that, when $\tilde{s} \neq s_{\mathrm{pr}}$, the leading contribution to $\theta_{\mathrm{b}}(0)$ is independent of Δ and thus does not depend on the surface roughness. Furthermore, the effective surface tilt angle depends linearly on ξ, which makes it significantly larger than the prediction (2.15), derived with Neumann-like boundary conditions on s, which possesses an extra ξ/η (small) factor. Finally, we remark the fact that $\theta_{\mathrm{b}}(0)$ shares the sign of m if and only if $\tilde{s} < s_{\mathrm{pr}}$. We will return below to the physical origin and implications of this result.

When the induced degree of orientation $\tilde{s}$ does exactly coincide with s_{pr}, all calculations simplify, since $h(\zeta) \equiv \log s_{\mathrm{pr}}$, and all first-order correction in (2.17) vanish. We therefore push our perturbation analysis and obtain

$$\begin{aligned} s(x,z) = s_{\mathrm{pr}} - \frac{s_{\mathrm{pr}}\xi^2}{\omega^2}\Bigg[\mathrm{m}^2 + \frac{\Delta^2}{\eta^2}\mathrm{e}^{-2z/\eta} - \frac{2\mathrm{m}\,\Delta}{\eta}\mathrm{e}^{-z/\eta}\cos\frac{x}{\eta} \\ - \mathrm{e}^{-\sqrt{3}\,\omega z/\xi}\left(\mathrm{m}^2 + \frac{\Delta^2}{\eta^2} - \frac{2\mathrm{m}\,\Delta}{\eta}\cos\frac{x}{\eta}\right)\Bigg] + \mathcal{O}\left(\varepsilon^3\right) \end{aligned} \tag{2.20}$$

$$\begin{aligned} \theta(x,z) = \mathrm{m}\,z + \Delta\,\mathrm{e}^{-z/\eta}\,\cos\frac{x}{\eta} + \frac{\xi^2}{\omega^2}\left(\frac{2\,\mathrm{m}\,\Delta^2}{\eta}\left(1-\mathrm{e}^{-2z/\eta}\right)\right. \\ \left. -\frac{\Delta^3}{2\,\eta^2}\left(\mathrm{e}^{-z/\eta}-\mathrm{e}^{-3z/\eta}\right)\cos\frac{x}{\eta} - \frac{2\,\mathrm{m}^2\,\Delta}{\eta}\,z\,\mathrm{e}^{-z/\eta}\,\cos\frac{x}{\eta}\right) + \mathcal{O}\left(\varepsilon^3\right)\,. \end{aligned} \tag{2.21}$$

Equation (2.21) allows us to compute the asymptotic tilt angle θ_{b} when $\tilde{s} = s_{\mathrm{pr}}$. In fact, once we drop all exponentially decaying terms in (2.21), we arrive at the interesting result that $\theta_{\mathrm{b}}(z)$ does exactly coincide with (2.14), that is, with the expression we derived with a Neumann-like boundary condition on the degree of orientation. In fact, the complete expression (2.21) for the tilt angle $\theta(x,z)$ coincides with (2.12) up to $\mathcal{O}\left(\varepsilon^3\right)$. Thus, any observation on the tilt angle is not able to distinguish among a free and a fixed boundary condition on the degree of orientation, as long as the imposed value $\tilde{s}$ coincides with the preferred value s_{pr}. This similarity between the Neumann and Dirichlet cases can be pursued further. Indeed, we can determine the $\mathcal{O}\left(\varepsilon^2\right)$-contributions in (2.16)–(2.17) also when $\tilde{s} \neq s_{\mathrm{pr}}$. If we then use them to compute the $\mathcal{O}\left(\varepsilon^2\right)$-correction to the asymptotic tilt angle (2.19), we arrive at the following expression, valid at $\mathcal{O}\left(\varepsilon^2\right)$ for any value of $\tilde{s}$:

$$\theta(x,z) \approx \theta_{\mathrm{b}}(z) = \left[\frac{\mathrm{m}\,\xi}{\sqrt{3}\,\omega}\left(\log\frac{s_{\mathrm{pr}}}{\tilde{s}} + \frac{s_{\mathrm{pr}}-\tilde{s}}{\tilde{s}}\right) + \frac{2\,\mathrm{m}\,\xi^2\Delta^2}{\eta\omega^2}\right] + \mathrm{m}\,z \quad \text{as } z \gg \eta, \tag{2.22}$$

which yields

$$\theta_{\mathrm{b}}(0) = \frac{\mathrm{m}\,\xi}{\sqrt{3}\,\omega}\left(\log\frac{s_{\mathrm{pr}}}{\tilde{s}} + \frac{s_{\mathrm{pr}}-\tilde{s}}{\tilde{s}}\right) + \frac{2\,\mathrm{m}\,\xi^2\Delta^2}{\eta\omega^2}\,. \tag{2.23}$$

The $\mathcal{O}\left(\varepsilon^2\right)$-contribution to the effective surface angle $\theta_{\mathrm{b}}(0)$ is thus fully a roughness effect and does not depend at all on the type of boundary conditions imposed on s.

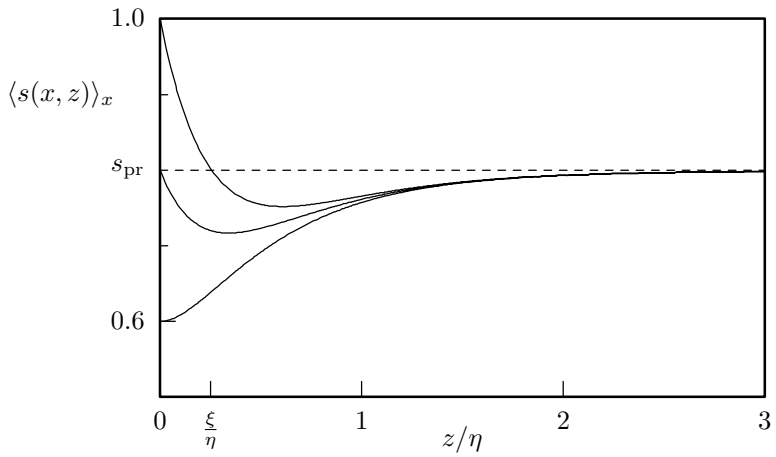


FIG. 2.3. *Boundary layers in the mean degree of orientation* $\langle s(x,z)\rangle_x$*, when* $\xi = 0.25\eta$*,* $s_{\rm pr} = 0.8$*,* $\omega = 0.6$*,* $\mathrm{m} = 0.1\eta$*, and* $\Delta = 1.5$*, when Dirichlet-like boundary conditions are applied on the degree of orientation. The boundary degree of orientation* $\tilde{s}$ *is, respectively, equal to* 1 *(top),* $s_{\rm pr}$ *(middle), and* 0.6 *(bottom).*

On the other hand, (2.23) confirms that the effective surface angle possesses also an $\mathcal{O}(\varepsilon)$-term when Dirichlet conditions are imposed on the degree of orientation, and $\tilde{s} \neq s_{\rm pr}$.

Figure 2.3 shows how the degree of orientation varies within the boundary layer as $\tilde{s}$ is fixed above, equal to, or below $s_{\rm pr}$. A double boundary-layer structure emerges. All plots exhibit a decrease of s in a region of characteristic size η: this effect comes from the $\mathcal{O}(\varepsilon^2)$-contribution. A similar surface melting was already presented and discussed in Figure 2.1. Close to the boundary, the $\mathcal{O}(1)$-term proportional to $(\tilde{s} - s_{\rm pr})\,\mathrm{e}^{-\sqrt{3}\,\omega z/\xi}$ settles the desired boundary value of s in a thin boundary layer of characteristic size ξ.

3. Effective weak anchoring. Once the boundary layer effects fade away, the main macroscopic effect of a rough surface on the director orientation is to allow for an effective surface tilt angle $\theta_{\rm b}(0)$, which apparently violates the homeotropic prescription $\theta(0) = 0$ (see (2.15) and (2.23)). It appears then natural to check whether the same macroscopic effect may be modeled through a weak anchoring potential, acting on a smooth surface. In this section we pursue this similarity, and we derive a relation connecting the microscopic roughness parameters with a macroscopic anchoring strength.

To solve the weak-anchoring problem, we consider a nematic liquid crystal which still spreads in the half-space $\mathcal{B} = \{z \geq 0\}$. To better compare our results with classical weak-anchoring models, we settle within Frank's director theory and thus look for the equilibrium distribution that minimizes the free-energy functional

$$\mathcal{F}[\mathbf{n}] := K \int_{\mathcal{B}} \left|\nabla \mathbf{n}\right|^2 dv + W \int_{\partial\mathcal{B}} f_{\rm w}[\mathbf{n}]\, da\,. \tag{3.1}$$

The bulk free-energy density in the functional (3.1) can be derived from its order-tensor theory counterpart by setting $s \equiv 1$ in (1.8). The anchoring potential $f_{\rm w}$ is required to attain its minimum at the homeotropic anchoring $\mathbf{n}\big|_{\partial\mathcal{B}} = \mathbf{e}_z$, while W is the *anchoring strength*.

We look again for equilibrium distributions of the type $\mathbf{n}(z) = \sin\theta(z)\,\mathbf{e}_x + \cos\theta(z)\,\mathbf{e}_z$. Thus, the free-energy functional (3.1) per unit transverse area can be written as

$$f[\theta] := K\int \theta'^2(z)\,dz + W\,f_{\mathrm{w}}\big(\theta(0)\big)\,, \tag{3.2}$$

where we assume $f'_{\mathrm{w}}(0) = 0$ and $f''_{\mathrm{w}}(0) > 0$, in order to guarantee the homeotropic preference. The minimizers of (3.2) satisfy the trivial Euler–Lagrange equation $\theta'' = 0$ and the boundary condition

$$K\theta'(0) - W f'_{\mathrm{w}}\big(\theta(0)\big) = 0\,. \tag{3.3}$$

When the anchoring strength W is large enough, the boundary condition (3.3) requires $\theta(0)$ to be small. When this is the case, a Taylor expansion in (3.3) supplies

$$\theta(0) \approx \frac{K\,\mathrm{m}}{W f''_{\mathrm{w}}(0)} = \zeta\,\mathrm{m}\,. \tag{3.4}$$

In (3.4) we have restored the notation $\mathrm{m} = \theta'(0)$, to better compare this estimate with our preceding results, and introduced the *surface extrapolation length*

$$\zeta := \frac{K}{W f''_{\mathrm{w}}(0)}\,, \tag{3.5}$$

a quantity that compares the relative strengths of the elastic and anchoring potentials.

The comparison between (3.4) and our results (2.15)–(2.23) relates the surface extrapolation length to the microscopic roughness parameters and/or the surface value of the degree of orientation. To further pursue this similarity we need to consider separately the different anchorings that may be applied to the degree of orientation.

- When s is free to choose its boundary value, (2.15) shows that the surface extrapolation length is given by

$$\frac{\zeta}{\xi} = \frac{2\Delta^2}{\omega^2}\,\frac{\xi}{\eta} + \mathcal{O}\left(\frac{\xi^2}{\eta^2}\right)\,. \tag{3.6}$$

  Thus, the anchoring strength increases when either the roughness amplitude Δ decreases (towards a smooth surface) or the roughness wavelength increases. An estimate of the order of magnitude of the effective roughness wavelength can be obtained by assuming typical values for the quantities involved in (3.6). Indeed, if we assume $\zeta \approx \xi$, $\Delta \approx 1$, and $\omega \approx \frac{1}{2}$, we arrive at $\eta \approx 10\xi$, which models a roughness wavelength in the hundredths of molecular lengths.
- When the boundary conditions fix the value of the degree of orientation at the surface, (2.23) yields

$$\frac{\zeta}{\xi} = \frac{1}{\sqrt{3}\,\omega}\left(\log\frac{s_{\mathrm{pr}}}{\tilde{s}} + \frac{s_{\mathrm{pr}} - \tilde{s}}{\tilde{s}}\right) + \frac{2\Delta^2}{\omega^2}\,\frac{\xi}{\eta} + \mathcal{O}\left(\frac{\xi^2}{\eta^2}\right)\,. \tag{3.7}$$

  Equation (3.7) shows that the surface extrapolation length includes two quite different contributions. The former depends on the difference between the boundary and the preferred values of the degree of orientation ($\tilde{s}$ and s_{pr}, respectively), while the latter depends on the surface roughness and indeed

coincides with (3.6). However, (3.7) may lose sense when $\tilde{s} > s_{\rm pr}$. Indeed, in this case ζ may become negative, so providing an *inverse* weak-anchoring effect. The physical origin of this odd result may be easily understood if we again resort to the $s^2|\nabla\theta|^2$-term in the free-energy density. By virtue of that term, the tilt angle prefers to limit its spatial variations in regions of higher s. If we force in the surface a higher degree of orientation than the bulk value, the tilt angle will flatten close to the surface, thus exhibiting the opposite behavior with respect to that shown in Figure 2.2. Equation (3.7) shows that this inverse effect may occur whenever

$$\frac{\tilde{s} - s_{\rm pr}}{s_{\rm pr}} \gtrsim \frac{\sqrt{3}\Delta^2}{\omega}\,\frac{\xi}{\eta} + \mathcal{O}\left(\frac{\xi^2}{\eta^2}\right)\,. \tag{3.8}$$

If we again replace the estimates above for Δ, ω, η, we arrive at the result that a fixed degree of orientation is able to completely hide the roughness-induced effective weak anchoring whenever $\tilde{s}$ exceeds $s_{\rm pr}$ by the 10% of the preferred value $s_{\rm pr}$ itself.

4. Discussion. We have examined both the boundary layer structure and the bulk effects of a rough surface bounding a nematic liquid crystal. Our main results may be summarized as follows.

- The roughness of the surface has been modeled by an oscillating anchoring condition, characterized by an oscillation amplitude Δ and a wavelength η. Figures 2.1 and 2.3 show that the rough boundary induces a partial melting in a neighborhood (of size η) of the external boundary. When Neumann-like boundary conditions are imposed on the degree of orientation, (2.13) quantifies the mean degree of order at the boundary. By contrast, were s to be forced to a prescribed value $\tilde{s}$ on the surface, (2.16) and (2.20) show that the boundary condition induces a thin boundary layer, determined by the nematic coherence length ξ.
- Once the degree of orientation decreases, the spatial variations of the tilt angle become cheaper, and thus θ is keen to steepen close to the external boundary. Figure 2.2 illustrates this effect. As a consequence, the effective boundary tilt angle $\theta_{\rm b}(0)$, extrapolated from the asymptotic outer solution $\theta_{\rm b}(z)$, becomes different from the null homeotropic prescription (see (2.15) and (2.23)). In section 3 we have shown that a similar effective anchoring breaking takes place when a weak-anchoring potential is assumed on a smooth surface (see (3.5) for the characteristic surface extrapolation length). The comparison between (3.4) and (2.15)–(2.23) allows one to relate the surface extrapolation length to the microscopic roughness parameters and/or the surface value of the degree of orientation (see (3.7) and (3.8)).

To conclude, we want to discuss the validity of two nontrivial simplifications we have introduced in our geometric setting. First, we have assumed that the boundary is perfectly sinusoidal, while a physical surface will exhibit a whole roughness spectrum. Second, we have replaced an undulating boundary by an undulating boundary condition on a flat surface. We collect in the appendices below the technical details of the calculations that may help in relaxing the above assumptions. Here we discuss the outcomes of such calculations.

As a first step towards dealing with a real, randomly wrinkled surface we have considered the case in which the boundary undulation may be described as a superposition of two sinusoidal undulations. We have thus solved the same equilibrium

problem by replacing the boundary condition (1.10) with the more general (A.1). Our main result, summarized in (A.2)–(A.3), is that the quantities that refer to the bulk *outer* solution simply add independently the contributions from each oscillation, without any interference effect. This result allows us to infer that, in the presence of a whole spectrum of independent roughness wavelengths, the global surface melting and the effective weak anchoring may be computed by simply adding each independent contribution in a Fourier integral.

We have then analyzed in detail how a homeotropic anchoring imposed on an undulating surface relaxes when entering the bulk. We have expanded the equilibrium configuration in power series of the dimensionless parameter δ, which represents the ratio between the height of the sinusoidal undulations and the roughness wavelength. A crucial result in the expansion is that the $O(\delta^n)$ contribution to the tilt angle possesses at most n nonzero Fourier components. Thus, taking into account only the first Fourier component when we evaluate the solution at a constant height (see (1.10)) amounts to neglecting $O(\delta^2)$ terms within the thin boundary layer. Equation (B.4) further supports our approximation. Indeed, it shows that the kth Fourier component induced by the undulating boundary decays in the bulk with a characteristic penetration length η/k. Thus, higher harmonics are penalized both by a higher-order δ^n coefficient and by a shorter penetration depth.

Appendix A. Roughness spectrum. Throughout the paper we have studied the bulk effects induced by the presence of a perfectly sinusoidal boundary. In real physical systems, however, the boundary roughness is mostly random, and a whole spectrum of roughness wavelengths is to be expected. In order to estimate whether the effects we have determined may be enforced or hidden by the interference between different wavelengths we briefly report here the results that may be obtained by replacing the boundary condition (1.10) by the more general

$$\theta(x,y,0) = \Delta_1 \cos\frac{x}{\eta_1} + \Delta_2 \cos\left(\frac{x}{\eta_2} + \phi_2\right) , \tag{A.1}$$

with $\eta_1/\eta_2 \notin \{\frac{1}{2}, 1, 2\}$, in order to avoid resonance effects.

We refer the reader to [34] for a complete overview and analysis of the results that follow. We here simply report how the main results are to be modified when stress-free (Neumann) boundary conditions are applied on the degree of orientation.

The first effect we have studied is the surface melting induced by the boundary roughness. Once we average along the x-direction and compute the solutions at the effective boundary $z = 0$, (2.13) is to be replaced by

$$\langle s(x,0)\rangle_x^{(2)} = s_{\mathrm{pr}}\left[1 - \frac{\mathrm{m}^2\xi^2}{\omega^2} - \frac{\xi^2}{\omega^2}\left(\frac{\Delta_1^2}{\eta_1^2} + \frac{\Delta_2^2}{\eta_2^2}\right) + \frac{2\xi^3}{\sqrt{3}\,\omega^3}\left(\frac{\Delta_1^2}{\eta_1^3} + \frac{\Delta_2^2}{\eta_2^3}\right)\right] + \mathcal{O}\left(\varepsilon^4\right) . \tag{A.2}$$

Expression (2.15) for the effective surface angle becomes

$$\theta_{\mathrm{b}}^{(2)}(0) = \frac{2\,\mathrm{m}\,\xi^2}{\omega^2}\left(\frac{\Delta_1^2}{\eta_1} + \frac{\Delta_2^2}{\eta_2}\right) . \tag{A.3}$$

Equations (A.2)–(A.3) show that the presence of more than one characteristic wavelength does not yield any dramatic result in the averaged quantities that interact with the bulk. In fact, they simply add their contributions, weighted by the roughness amplitudes. The situation is clearly more complex if we aim at computing the exact

solutions within the boundary layers [34]. In particular, the x-periodicity is lost as soon as the roughness wavelengths are not commensurable.

Appendix B. Modeling an undulating boundary. In section 1.1 we have modeled a homeotropic boundary condition imposed on an undulating surface through an oscillating boundary condition imposed on a flat surface. In this appendix we analyze the validity of such an approximation. In order to avoid unnecessarily lengthy calculations, we perform the present check within the Frank approximation, that is, by assuming that the nematic coherence length ξ is much smaller than all other lengths involved in the problem. When this is the case, the degree of orientation is constrained to the value $s_{\rm pr}$ that minimizes the Landau–de Gennes potential, the Euler–Lagrange equation $(1.11)_1$ becomes Laplace's equation, and thus the tilt angle θ is harmonic.

We consider the region $\mathcal{A} = \{(x,z) : z \geq \delta\eta \sin x\}$ and look for a x-periodic harmonic function $\theta : \mathcal{A} \to \mathbb{R}$ (with x-period η) that satisfies the boundary conditions

$$\theta(x, \delta\eta \sin \tfrac{x}{\eta}) = \arctan\bigl(\delta \cos \tfrac{x}{\eta}\bigr)\,, \qquad \theta(x,z) \approx \tilde{\theta}(z) \quad \text{as } z \to +\infty \tag{B.1}$$

for all values of x. The boundary condition $(B.1)_1$ guarantees that the unit vector $\mathbf{n} = \sin\theta\,\mathbf{e}_x + \cos\theta\,\mathbf{e}_z$ is homeotropically anchored to the physical boundary, while $(B.1)_2$ guarantees that the bulk configuration depends only on the z-coordinate. Let us expand the tilt angle in power series of the amplitude coefficient δ:

$$\theta(x,z) = \sum_{n=0}^{\infty} \theta_n(x,z)\,\delta^n\,. \tag{B.2}$$

We next Fourier-expand all functions θ_n along the periodic direction

$$\theta_n(x,z) = \sum_{k=0}^{\infty} a_{n,k}(z)\,\cos\frac{kx}{\eta} + \sum_{k=1}^{\infty} b_{n,k}(z)\,\sin\frac{kx}{\eta}\,. \tag{B.3}$$

The Laplace equation implies then

$$\theta_n(x,z) = \sum_{k=0}^{\infty} \alpha_{n,k}\,\mathrm{e}^{-kz/\eta}\cos\frac{kx}{\eta} + \sum_{k=1}^{\infty} \beta_{n,k}\mathrm{e}^{-kz/\eta}\,\sin\frac{kx}{\eta}\,, \tag{B.4}$$

where the coefficients $\{\alpha_{n,k}, \beta_{n,k}\}$ can be determined by requiring $(B.1)_1$ to hold.

Let us now compute the value the tilt angle attains at the (horizontal) height $z = \delta\eta$, which we aim to consider as effective flat boundary (see Figure 1.1). We obtain

$$\theta(x, \delta\eta) = \bigl(\delta - \delta^2\bigr)\cos\frac{x}{\eta} + \frac{\delta^2}{2}\sin\frac{2x}{\eta} + O(\delta^3)\,. \tag{B.5}$$

In general, it can be shown that the nth coefficient θ_n in expansion (B.2) contains only Fourier components up to $k \leq n$. Thus, the boundary condition (1.10) used in the text is exact up to $O(\delta^2)$. Furthermore, the roughness amplitude Δ simply coincides with δ, the (dimensionless) ratio between the height of the sinusoidal undulations and the roughness wavelength.

Acknowledgment. P. B. thanks Georges E. Durand for useful discussions on the present topics.

REFERENCES

[1] D. W. Berreman, *Solid surface shape and the alignment of an adjacent nematic liquid crystal*, Phys. Rev. Lett., 28 (1972), pp. 1683–1686.

[2] S. Kumar, J.-H. Kim, and Y. Shi, *What aligns liquid crystals on solid substrates? The role of surface roughness anisotropy*, Phys. Rev. Lett., 94 (2005), 077803.

[3] F. Batalioto, I. H. Bechtold, E. A. Oliveira, and L. R. Evangelista, *Effect of microtextured substrates on the molecular orientation of a nematic liquid-crystal sample*, Phys. Rev. E (3), 72 (2005), 031710.

[4] J. Elgeti and F. Schmid, *Nematic liquid crystals at rough and fluctuating interfaces*, Eur. Phys. J. E, 18 (2005), pp. 407–415.

[5] S. Faetti, M. Gatti, V. Palleschi, and T. J. Sluckin, *Almost critical behavior of the anchoring energy at the interface between a nematic liquid crystal and a SiO substrate*, Phys. Rev. Lett., 55 (1985), pp. 1681–1684.

[6] R. Barberi and G. E. Durand, *Order parameter of a nematic liquid-crystal on a rough-surface*, Phys. Rev. A, 41 (1990), pp. 2207–2210.

[7] L. Xuan, T. Tohyama, T. Miyashita, and T. Uchida, *Order parameters of the liquid crystal interface layer at a rubbed polymer surface*, J. Appl. Phys., 96 (2004), pp. 1953–1958.

[8] G. Barbero and G. E. Durand, *Curvature induced quasi-melting from rough surfaces in nematic liquid-crystals*, J. Physique II, 1 (1991), pp. 651–658.

[9] G. Skačej, A. L. Alexe-Ionescu, G. Barbero, and S. Žumer, *Surface-induced nematic order variation: Intrinsic anchoring and subsurface director deformations*, Phys. Rev. E (3), 57 (1998), pp. 1780–1788.

[10] V. Mocella, C. Ferrero, M. Iovane, and R. Barberi, *Numerical investigation of surface distortion and order parameter variation in nematics*, Liq. Cryst., 26 (1999), pp. 1345–1350.

[11] D. L. Cheung and F. Schmid, *Monte Carlo simulations of liquid crystals near rough walls*, J. Chem. Phys., 122 (2005), 074902.

[12] Y. Sato, K. Sato, and T. Uchida, *Relationship between rubbing strength and surface anchoring of nematic liquid crystal*, Jpn. J. Appl. Phys., 31 (1992), pp. L579–L581.

[13] L. R. Evangelista and G. Barbero, *Theoretical-analysis of actual surfaces: The effect on the nematic orientation*, Phys. Rev. E (3), 48 (1993), pp. 1163–1171.

[14] L. R. Evangelista and G. Barbero, *Walls of orientation induced in sematic-liquid-crystal samples by inhomogeneous surfaces*, Phys. Rev. E (3), 50 (1994), pp. 2120–2133.

[15] A. L. Alexe-Ionescu, R. Barberi, G. Barbero, and M. Giocondo, *Anchoring energy for nematic liquid-crystals: Contribution from the spatial variation of the elastic-constants*, Phys. Rev. E (3), 49 (1994), pp. 5378–5388.

[16] J.-B. Fournier and P. Galatola, *Effective anchoring and scaling in nematic liquid crystals*, Eur. Phys. J. E, 2 (2000), pp. 59–65.

[17] A. Strigazzi, *Surface elasticity and Freedericksz threshold in a nematic cell weakly anchored*, Nuovo Cimento D, 10 (1988), pp. 1335–1344.

[18] E. G. Virga, *Variational Theories for Liquid Crystals*, Chapman and Hall, London, 1994.

[19] G. Napoli, *Weak anchoring effects in electrically driven Freedericksz transitions*, J. Phys. A, 39 (2006), pp. 11–31.

[20] M. A. Osipov, T. J. Sluckin, and S. J. Cox, *Influence of permanent molecular dipoles on surface anchoring of nematic liquid crystals*, Phys. Rev. E (3), 55 (1997), pp. 464–476.

[21] A. M. Sonnet and E. G. Virga, *Dilution of nematic surface potentials: Statics*, Phys. Rev. E (3), 61 (2000), pp. 5401–5406.

[22] A. M. Sonnet, E. G. Virga, and G. E. Durand, *Dilution of nematic surface potentials: Relaxation dynamics*, Phys. Rev. E (3), 62 (2000), pp. 3694–3701.

[23] P.-G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, 2nd ed., Oxford University Press, Oxford, UK, 1995.

[24] N. Schopohl and T. J. Sluckin, *Defect core structure in nematic liquid crystals*, Phys. Rev. Lett., 59 (1987), pp. 2582–2584.

[25] P. Biscari, G. Guidone Peroli, and T. J. Sluckin, *The topological microstructure of defects in nematic liquid crystals*, Mol. Cryst. Liq. Cryst., 292 (1997), pp. 91–101.

[26] P. Biscari and T. J. Sluckin, *Expulsion of disclinations in nematic liquid crystals*, European J. Appl. Math., 14 (2003), pp. 39–59.

[27] P. Biscari, G. Capriz, and E. G. Virga, *On surface biaxiality*, Liq. Cryst., 16 (1994), pp. 479–489.

[28] P. Biscari and G. Guidone Peroli, *A hierarchy of defects in biaxial nematics*, Comm. Math. Phys., 186 (1997), pp. 381–392.

[29] P. Biscari, G. Napoli, and S. Turzi, *Bulk and surface biaxiality in nematic liquid crystals*, Phys. Rev. E (3), 74 (2006), 031708.
[30] M. Holmes, *Introduction to Perturbation Methods*, Springer-Verlag, New York, 1995.
[31] C. Bender and S. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, Springer-Verlag, New York, 1999.
[32] J. Murdock, *Perturbations. Theory and Methods*, Wiley-Interscience, New York, 1991.
[33] D. R. Smith, *Singular-Perturbation Theory. An Introduction with Applications*, Cambridge University Press, Cambridge, UK, 1985.
[34] S. Turzi, *Distortion-Induced Effects in Nematic Liquid Crystals*, Ph.D. thesis, Politecnico di Milano, Italy, 2007.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 464–486

HIGH LEWIS NUMBER COMBUSTION WAVEFRONTS: A PERTURBATIVE MELNIKOV ANALYSIS*

SANJEEVA BALASURIYA†, GEORG GOTTWALD‡, JOHN HORNIBROOK‡, AND STÉPHANE LAFORTUNE§

Abstract. The wavefronts associated with a one-dimensional combustion model with Arrhenius kinetics and no heat loss are analyzed within the high Lewis number perturbative limit. This situation, in which fuel diffusivity is small in comparison to that of heat, is appropriate for highly dense fluids. A formula for the wavespeed is established by a nonstandard application of Melnikov's method and slow manifold theory from dynamical systems, and compared to numerical results. A simple characterization of the wavespeed correction is obtained: it is proportional to the ratio between the exothermicity parameter and the Lewis number. The perturbation method developed herein is also applicable to more general coupled reaction-diffusion equations with strongly differing diffusivities. The stability of the wavefronts is also tested using a numerical Evans function method.

Key words. combustion waves, high Lewis number, Melnikov's method, slow manifold reduction, Evans function

AMS subject classifications. 80A25, 35K57, 35B35, 34E10, 34C37

DOI. 10.1137/050640849

1. Introduction.

In this article, we study the wavespeed of a combustion wavefront along a one-dimensional medium. This is a fundamental idealized problem towards understanding how flame fronts propagate and therefore has received a considerable amount of attention. There are several (nondimensional) parameters of importance: the Lewis number Le, the exothermicity parameter β, and the heat loss parameter ℓ. The first of these, the Lewis number, measures the relative importance of fuel diffusivity in comparison to that of heat. The exothermicity β is the ratio of the activation energy to the heat of reaction. The structure of the governing equations is such that an infinite Lewis number is considerably easier to deal with than allowing for fuel diffusivity. Many studies of this "solid" regime appear in the literature [5, 7, 28, 36, 37], and also the "gaseous" regime Le ≈ 1 has been frequently studied because of a symmetry in the equations [7, 20, 24, 37, 39]. Usually, the heat loss is neglected in these "adiabatic" studies. In several of these articles [28, 36, 37] the condition $\beta \gg 1$ is essential to the wavespeed and stability analysis. The case $\beta \ll 1$ has also been studied [8], in which a perturbative method is used to model the temperature. The bifurcation structure with respect to the heat loss parameter ℓ is addressed in [34], which obtains a stability diagram with respect to ℓ and the wavespeed.

*Received by the editors September 21, 2005; accepted for publication (in revised form) September 29, 2006; published electronically February 9, 2007.
http://www.siam.org/journals/siap/67-2/64084.html

†Corresponding author. Department of Mathematics, Connecticut College, 270 Mohegan Avenue, New London, CT 06320 (sanjeeva.balasuriya@conncoll.edu). Part of this work was done while this author was visiting the Department of Mathematics at the College of Charleston.

‡School of Mathematics & Statistics, University of Sydney, Sydney NSW 2006, Australia (gottwald@maths.usyd.edu.au, johnh@maths.usyd.edu.au). The second author was supported by the Australian Research Council, grant DP0452147.

§Department of Mathematics, College of Charleston, 66 George Street, Charleston, SC 29424 (lafortunes@cofc.edu). Part of this work was done while this author was visiting the School of Mathematics and Statistics at the University of Sydney. The research of this author was supported by the National Science Foundation under grant DMS-0509622.

We note that the limit of small fuel diffusivity (large, but not infinite, Lewis number) has not received much attention, perhaps because of the singularity of this limit in the governing equations. Yet this limit may be argued to be particularly appropriate for very high density fluids burning at high temperatures, such as would occur, for example, in the burning of toxic wastes at supercritical temperatures [25]. Even for solids, some mass diffusivity is to be expected at very high temperatures, particularly in the reaction zone in which liquification may occur. In [27], the mass diffusivity is modeled by an Arrhenius temperature dependence, which would result in a large effective Lewis number in certain situations (such as when the (scaled) adiabatic flame temperature is small in comparison to the activation energy for mass diffusion). It is this very large Lewis number limit which we study in this article, without restricting β. We do a detailed analysis of the wavespeed of combustion waves which can be supported. We also verify the linear stability of such wavefronts using an Evans function technique.

The model we use is for a premixed fuel in one dimension, with no heat loss and with an Arrhenius law for the reaction rate. These combustion dynamics can be represented in nondimensional form by [5, 8, 20, 24, 28, 34, 36, 37, 39]

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2} + y\, e^{-1/u}, \\ \dfrac{\partial y}{\partial t} = \dfrac{1}{\mathrm{Le}}\dfrac{\partial^2 y}{\partial x^2} - \beta\, y\, e^{-1/u}. \end{cases} \tag{1.1}$$

Here, $u(x,t)$ is the temperature, and $y(x,t)$ the fuel concentration, at a point x at time t. The parameters β and Le are as described earlier. We are neglecting heat loss (had we included it, an additional term $-\ell\,(u-u_a)$ for some ambient temperature u_a would be necessary on the right-hand side of the u equation in (1.1)). This one-dimensional model is also applicable to combustion in cylinders [30], with u and y being cross-sectionally averaged quantities in this case. See also [6, 7, 19, 26] for closely related governing equations. The nondimensionalization leading to (1.1) ensures that the cold boundary problem is circumvented (see [36] for a discussion). Since the Lewis number will be assumed large, set $\epsilon = 1/\mathrm{Le}$ with $0 \le \epsilon \ll 1$. This small ϵ limit clearly constitutes a singular perturbation in (1.1).

This article analyzes (1.1) as follows. In section 2, we determine the wavespeed as a function of β and ϵ. We initially consider the situation where $\mathrm{Le} = \infty$ (section 2.1), since this wavespeed is relevant to our subsequent perturbative analysis for $1 \ll \mathrm{Le} < \infty$ (sections 2.2, 2.3, and 2.4). While the infinite Lewis number situation is well studied, we are able to empirically determine a simple exponential formula for the wavespeed as a function of β. The case $1 \ll \mathrm{Le} < \infty$ is initially examined numerically in section 2.2, in which we obtain a method for computing the wavespeed. In the subsequent sections, we establish a theoretical estimate for the wavespeed with the help of two suitably modified tools from dynamical systems theory: a slow manifold reduction and Melnikov's method. In section 2.3, we reduce the dimensionality of the problem using a slow manifold reduction argument. This enables us in section 2.4 to utilize a nonstandard adaptation of Melnikov's method to find a theoretical estimate for the wavespeed. (This new technique is adaptable to other situations in which the wavespeed correction due to the presence of a small parameter is needed.) Our asymptotics enable the determination of a remarkably simple formula for the wavespeed, which is accurate for *all* β values (and not restricted to the "usual" large β limit). Essentially, we find that the relative wavespeed correction in going from an

infinite to a large Lewis number is proportional to (β/Le).

A brief stability analysis of the wavefronts is given in section 3. Having described the Evans function approach to stability in section 3, we compute the Evans function for high Lewis number combustion wavefronts using an exterior algebra [2, 16, 23, 40]. We note that in [20], an exterior algebra method has been successfully used to numerically investigate stability of wavefronts in combustion systems. A detailed stability analysis in the β-Le^{-1} plane is given therein for the system (1.1). As with infinite Lewis number fronts (see [5, 14, 28, 37]), [20] shows that stability occurs for small β but that, as β is increased, a Hopf bifurcation leads to an oscillatory instability. The β and Le values we test give results consistent with the stability boundary determined in [20]. Thus, stability properties remain essentially unaltered despite the singularity in the limit $\mathrm{Le} \to \infty$.

2. Wavespeed analysis. We seek wavefronts which travel in time, and hence set $u(x,t) = u(\xi)$ and $y(x,t) = y(\xi)$, where $\xi = x - c\,t$ and c is the traveling wave speed. Under this ansatz, (1.1) reduces to

$$\begin{cases} u'' + c\,u' + y\,e^{-1/u} = 0, \\ \epsilon\,y'' + c\,y' - \beta\,y\,e^{-1/u} = 0. \end{cases} \tag{2.1}$$

2.1. Wavefront for $\mathrm{Le} = \infty$. Set $\epsilon = 0$ in (2.1). Upon defining the new variable $v = u'$, the dynamics can be represented by a three-dimensional first-order system

$$\begin{cases} u' = v, \\ v' = -c\,v - y\,e^{-1/u}, \\ y' = \dfrac{\beta}{c}\,y\,e^{-1/u}. \end{cases} \tag{2.2}$$

The system (2.2) possesses a conserved quantity

$$H_c(u,v,y) = \beta\,v + \beta\,c\,u + c\,y, \tag{2.3}$$

since it is verifiable that $dH_c/d\xi = 0$ along trajectories of (2.2). Thus, motion is confined to planes defined by $H_c(u,v,y) = \text{constant}$. Now, for a wavefront, we require that $(u,v,y) \to (0,0,1)$ as $\xi \to \infty$; this corresponds to the region in which fuel is not yet burnt (and remains at its maximum nondimensional concentration of one) and the temperature (and its variation) is still zero. This point lies on $H_c(u,v,y) = c$, giving a well-known conservation relation [37]. At the other limit $\xi \to -\infty$, the fuel is completely burnt, and has reached a steady temperature, and so $(u,v,y) \to (u_B,0,0)$, where the temperature u_B is to be determined. Utilizing $H_c(u_B,0,0) = c$, we find that $u_B = 1/\beta$ is necessary; see also [8, 20, 37] for alternative ways to obtain this value.

Thus, we seek a heteroclinic solution of (2.2), which progresses between the fixed points $(1/\beta,0,0)$ and $(0,0,1)$, and is confined to the plane $\beta\,v + \beta\,c\,u + c\,y = c$; i.e., the fuel concentration obeys

$$y = 1 - \beta\,u - \frac{\beta}{c}\,v \tag{2.4}$$

at all values of ξ. Considering (2.2) under this restriction, we obtain

$$\begin{cases} u' = v, \\ v' = -\,c\,v - \left(1 - \beta\,u - \dfrac{\beta}{c}\,v\right)\,e^{-1/u}. \end{cases} \tag{2.5}$$

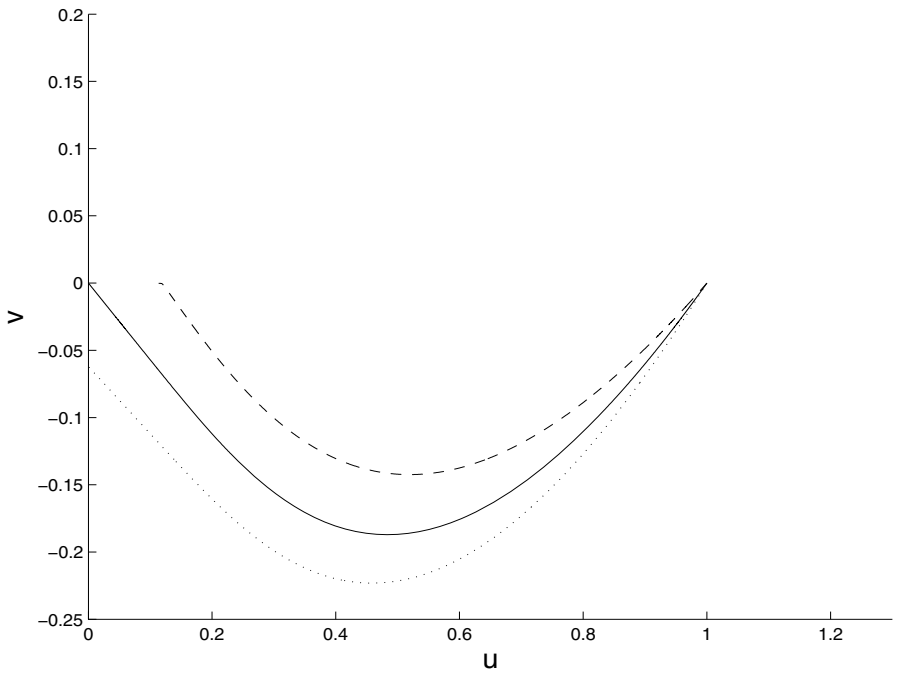


FIG. 2.1. *Projection onto the* (u, v)*-plane of trajectories of* (2.2) *lying on different planes* $H_c = c$*. Here,* $\beta = 1$*, and the three curves correspond to* $c = 0.5$ (*dotted*)*,* 0.5707 (*solid*)*, and* 0.7 (*dashed*)*.*

This is effectively a projection of the flow onto the particular invariant plane $H_c(u, v, y) = c$ onto the (u, v)-plane. Any value of c for which a heteroclinic connection exists between $(1/\beta, 0)$ and $(0, 0)$ is a permitted speed for the wavefront.

The unstable eigendirection of the point $(1/\beta, 0)$ is $(-c, -\beta e^{-\beta})$, and we determine c numerically by shooting along this direction and attempting to match up with a trajectory approaching the origin. In Figure 2.1 we show several numerically computed trajectories of this form, for different values of c, where we have chosen $\beta = 1$. Note that this is not a standard (u, v)-phase space for (2.5), since each curve corresponds to a different value of the parameter c. Rather, it is a projection onto the (u, v)-plane of specialized trajectories from the invariant planes $H_c(u, v, y) = c$ of the three-dimensional system (2.2). The one trajectory which makes the required connection lies in the invariant plane corresponding to $c = 0.5707$. The determination of this c value was obtained by making incremental adjustments of c until an appropriate connection is obtained.

We use this technique to numerically compute the wavespeeds for various values of the fuel parameter β, and we illustrate this dependence in Figure 2.2. The wavespeed decays with β. For fuels with larger β (poor fuels), the energy resulting from the reaction is insufficient to quickly activate combustion in nearby material, and combustion fronts propagate slowly. The data fits the exponential curve

$$c(\beta) = 0.927\, e^{-0.486\beta} \tag{2.6}$$

with correlation $\rho > 0.9999$. Equation (2.6) therefore provides an empirically determined formula of excellent accuracy for the speed of a wavefront in perfectly solid

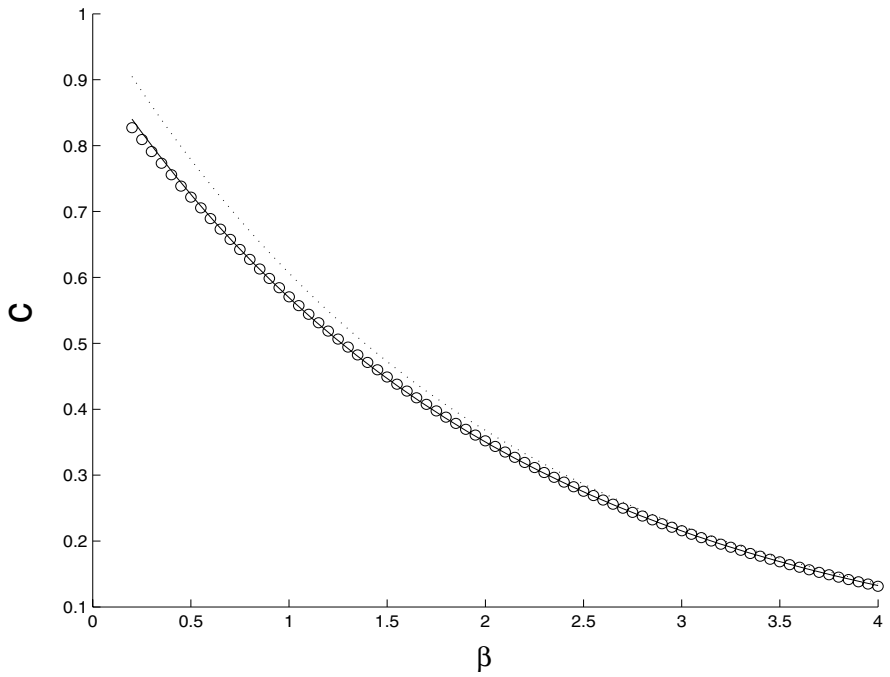


FIG. 2.2. *Variation of the wavespeed c with β: open circles (numerical results); unbroken curve (empirical curve (2.6)); dotted curve* $(\exp(-0.5\beta)$*, as obtained in* [7, 28, 36]).

adiabatic one-dimensional media. This result is close (and consistent with) a variety of sources: $\exp(-0.5\beta)$ is quoted in [36] for the small β limit; this same value is given as an upper bound in [7] and is also implied in eq. (10) in [28] using a large β limit within a discontinuous front approximation. See Figure 2.2 for a comparison with our results.

The structure of the temperature front is illustrated in Figure 2.3 for $\beta = 1$ (solid curve, left scale) and $\beta = 3$ (dashed curve, right scale), demonstrating that larger β fronts have a broader preheat layer preceding the front. Note that the preheat zone differs from the reaction zone [24]. The latter shrinks with increasing β [13]. Specifically, the reaction zone *as a fraction of the preheat zone* is $\mathcal{O}(1/\beta)$ [24, 28]. The reaction zone is well localized near the region of greatest temperature change [24] and is not immediately identifiable in temperature profiles as in Figure 2.3. Indeed, the increase in size of the preheat layer with β supports the $\mathcal{O}(1/\beta)$ expectation for the ratio between the reaction and the preheat zones.

2.2. Wavespeed for $1 \ll \mathbf{Le} < \infty$. When the Lewis number is not infinite, but large, ϵ is small, and weak fuel diffusion needs to be permitted. This is a *singular* limit in (1.1) and (2.1), and as a consequence has been hardly examined in the literature. By defining $v = u'$ as before, but now also $z = y'$, the governing equations (2.1) can be represented as a four-dimensional system

$$\begin{cases} u' = v, \\ v' = -c\,v - y\,e^{-1/u}, \\ y' = z, \\ z' = \dfrac{1}{\epsilon}\left(-c\,z + \beta\,y\,e^{-1/u}\right). \end{cases} \tag{2.7}$$

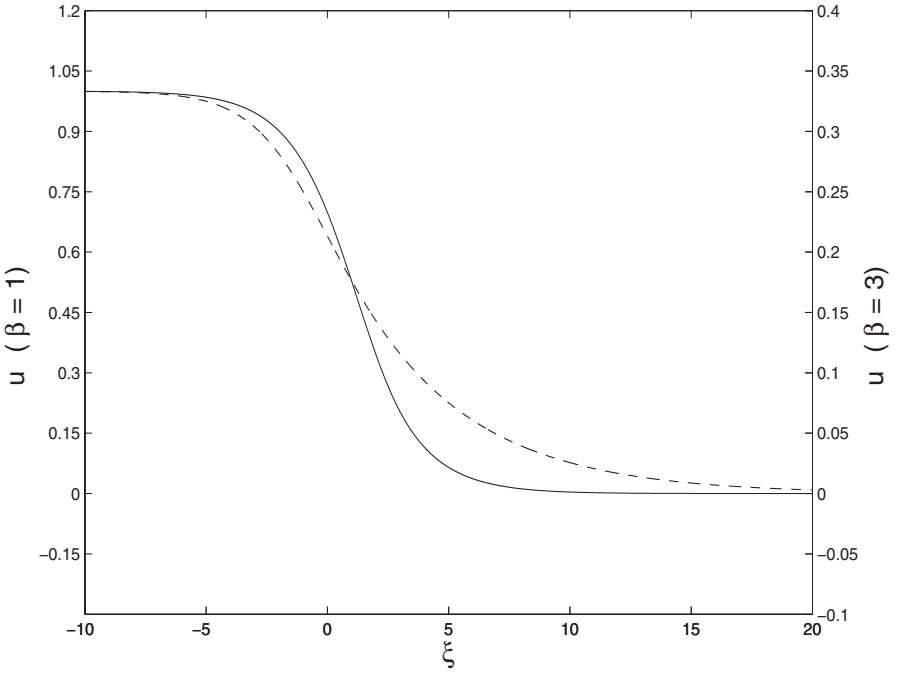


FIG. 2.3. *Temperature front at* $\beta = 1$ (*solid, left scale*) *and* $\beta = 3$ (*dashed, right scale*).

This is reducible to three dimensions: the quantity

$$G_c^\epsilon(u, v, y, z) = \beta\, v + \beta\, c\, u + c\, y + \epsilon\, z$$

can be verified to be a conserved quantity of (2.7). Hence, flow is confined to the invariant three-dimensional surfaces $G_c^\epsilon =$ constant. Now, we seek a wavefront solution which goes from $(u, v, y, z) = (u_B, 0, 0, 0)$ to a value $(0, 0, 1, 0)$, and we find that $G_c^\epsilon(u, v, y, z) = c$, and $u_B = 1/\beta$ as before. The three-dimensional invariant surface on which both points lie is

$$z = \frac{1}{\epsilon}\left(c - \beta\, v - \beta\, c\, u - c\, y\right) .$$

The dynamics of (2.7) on this surface can be projected onto the three variables (u, v, y), such that

$$\begin{cases} u' = v, \\ v' = -c\, v - y\, e^{-1/u}, \\ y' = \dfrac{1}{\epsilon}\left(c - \beta\, v - \beta\, c\, u - c\, y\right). \end{cases} \tag{2.8}$$

We seek the value of c which permits a heteroclinic connection from $(u, v, y) = (1/\beta, 0, 0)$ to $(0, 0, 1)$. The former point (corresponding to $\xi = -\infty$) has only one positive eigenvalue, given by $(-c + \sqrt{c^2 + 4\epsilon\beta e^{-\beta}})/(2\epsilon)$. For small ϵ, we "shoot" in the eigendirection corresponding to this, with an initial guess of the wavespeed given by (2.6). Thereafter, as in the previous section, we adjust c until we obtain a solution

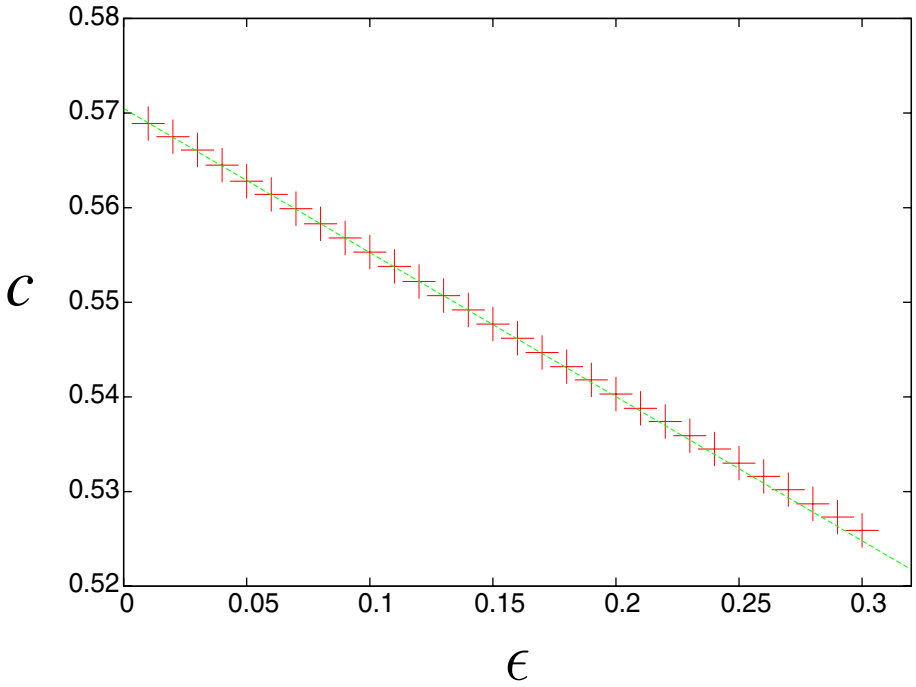


FIG. 2.4. *Numerically computed wavespeed variation with ϵ for $\beta = 1$ (the crosses). The dashed line is the theoretical approximation for $\beta = 1$ obtained from the methods of section* 2.4.

which approaches the point $(0, 0, 1)$ as $\xi \to \infty$. We do this numerically by considering the conditions $c\,y + \epsilon z = c$ and $v + c\,u = 0$, which the front must obey at $\xi = +\infty$, and using a root-finding algorithm to adjust c. For a fixed value $\beta = 1$, we illustrate how the wavespeed c varies with ϵ in Figure 2.4, with the crosses. The dashed curve in Figure 2.4 is a analytical/numerical approximation we obtain for the wavespeed in terms of an explicit formula (2.17). The next two sections describe how we obtain this formula.

We notice that c decreases as we increase ϵ, that is, when we *decrease* the Lewis number. Now, in dimensional form $\mathrm{Le} = \kappa/(\rho\, c_p\, D)$, where ρ, κ, c_p, and D are, respectively, the density, thermal conductivity, specific heat capacity, and molecular diffusivity of the fuel [6, 8, 30, 34, 37]. Increasing ϵ is equivalent to increasing the relative importance of D, ρ, and c_p in relation to κ. Reducing κ obviously decreases the ability of heat to move and hence the combustion speed. Higher densities result in increased fuel mass in each location, which means more heat is needed in a given area to ignite all of the fuel before the wave moves on. Fuels with increased c_p require more heat to increase the temperature by a specified amount. Finally, increasing D increases the transport of burnt fuel into the unburnt region and vice versa, interfering with front propagation.

We computed the changes to the wavefront profile (akin to Figure 2.3) when ϵ is changed (not shown). We verified the obvious physical conclusion that the fuel concentration front becomes less steep when ϵ is increased from zero.

2.3. Slow manifold reduction. We now show that in the limit of small ϵ, it is possible to further reduce the system (2.8) to a *two*-dimensional flow on a *slow*

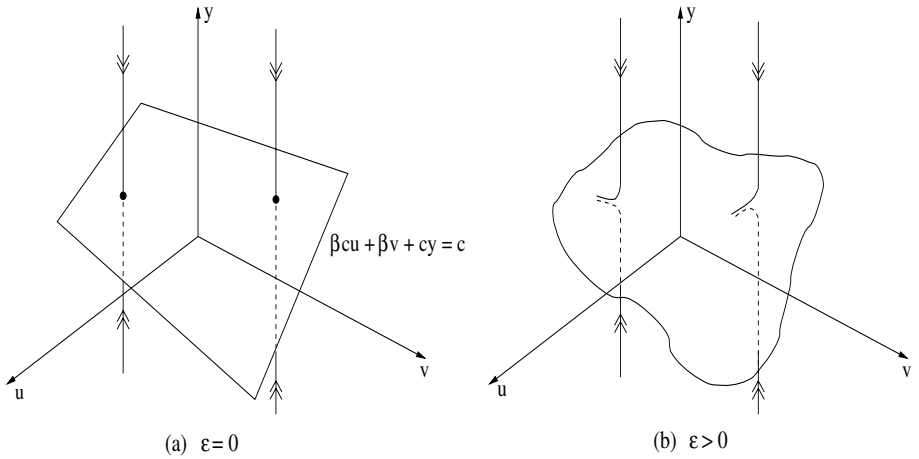


FIG. 2.5. *The hyperbolic invariant manifold* (a) $\mathcal{S}_0$ *for* (2.10) *and* (b) $\mathcal{S}_\epsilon$ *for* (2.9).

manifold. We begin with (2.8) and note that there are two "time"-scales in this singularly perturbed system, where we use "time" loosely to mean the independent variable ξ. We therefore adopt the standard dynamical systems trick of defining a new independent variable $\eta = \xi/\epsilon$ to elucidate motion in the fast "time" η. With a dot denoting the rate of change with respect to η, (2.8) becomes

$$\begin{cases} \dot{u} = \epsilon\, v, \\ \dot{v} = \epsilon \left(-c\, v - y\, e^{-1/u}\right), \\ \dot{y} = c - \beta\, v - \beta\, c\, u - c\, y. \end{cases} \tag{2.9}$$

In the $\epsilon = 0$ limit, the system collapses to

$$\begin{cases} \dot{u} = 0, \\ \dot{v} = 0, \\ \dot{y} = c - \beta\, v - \beta\, c\, u - c\, y, \end{cases} \tag{2.10}$$

in which it is clear that the plane $\mathcal{S}_0$ defined by $c - \beta\, v - \beta\, c\, u - c\, y = 0$ consists entirely of fixed points. This is the same plane as defined through $H_c(u, v, y) = c$ for (2.2), on which the interesting behavior occurred for perfectly solid fuels. Each fixed point has a one-dimensional stable manifold (in the y-direction), and a two-dimensional center manifold, which is $\mathcal{S}_0$. Thus the plane $\mathcal{S}_0$ is invariant and normally hyperbolic with respect to (2.10); there is exponential contraction towards it as illustrated in Figure 2.5(a).

Upon switching on ϵ and considering the dynamics (2.9), $\mathcal{S}_0$ perturbs to an invariant curved entity $\mathcal{S}_\epsilon$, which is order ϵ away from $\mathcal{S}_0$. This is because of the structural stability of normally hyperbolic sets [18], which also implies that normal hyperbolicity is preserved for small ϵ. Therefore, there is exponential decay of trajectories towards $\mathcal{S}_\epsilon$ on time-scales of order η, as shown in Figure 2.5(b). Motion on $\mathcal{S}_\epsilon$ occurs at a slower rate (on time-scales of order ξ), and hence it is termed a "slow manifold." The heteroclinic connection we seek lies on $\mathcal{S}_\epsilon$, from $(u, v, y) = (1/\beta, 0, 0)$ to $(0, 0, 1)$. Since $\mathcal{S}_\epsilon$ is invariant, two-dimensional, and not parallel to the y-axis, it therefore makes sense to project the motion onto the (u, v)-plane in order to describe behavior. To elucidate this motion, we need to once again return to the original time-scale ξ—the slow time associated with motion on the slow manifold.

We return to the relationship $G_c^\epsilon\,(u(\xi), v(\xi), y(\xi), z(\xi)) = c$, which upon differentiation yields

$$\beta\, v' + \beta\, c\, u' + c\, y' + \epsilon\, z' = 0\,,$$

and since $u' = v$ and $y' = z$,

$$z = -\frac{\beta}{c}\, v' - \beta\, v - \frac{\epsilon}{c} z'\,.$$

Substituting back into $G_c^\epsilon(u, v, y, z) = c$, we obtain

$$\beta\, v + \beta\, c\, u + c\, y + \epsilon \left(-\frac{\beta}{c}\, v' - \beta\, v + \mathcal{O}(\epsilon) \right) = c\,,$$

and thus

$$y = 1 - \frac{\beta}{c}\, v - \beta\, u + \epsilon\, \frac{\beta}{c^2}\, v' + \epsilon\, \frac{\beta}{c}\, v + \mathcal{O}(\epsilon^2)\,.$$

Substitution into the v' equation in (2.7) or (2.8) gives

$$v' \left(1 + \epsilon\, \frac{\beta}{c^2} e^{-1/u} \right) = -c\, v - \left(1 - \frac{\beta}{c}\, v - \beta\, u + \epsilon\, \frac{\beta}{c}\, v + \mathcal{O}(\epsilon^2) \right) e^{-1/u}\,.$$

Therefore

$$\begin{aligned} v' &= \left(1 - \epsilon\, \frac{\beta}{c^2} e^{-1/u} \right) \left[-c\, v - \left(1 - \frac{\beta}{c}\, v - \beta\, u + \epsilon\, \frac{\beta}{c}\, v \right) e^{-1/u} \right] + \mathcal{O}(\epsilon^2) \\ &= -c\, v - \left(1 - \frac{\beta}{c}\, v - \beta\, u \right) e^{-1/u} + \epsilon\, \frac{\beta}{c^2} \left(1 - \frac{\beta}{c}\, v - \beta\, u \right) e^{-2/u} + \mathcal{O}(\epsilon^2)\,. \end{aligned}$$

Retaining only $\mathcal{O}(\epsilon)$ terms, we obtain the (u, v)-projected approximate equations on the slow manifold

$$\text{(2.11)} \qquad \begin{cases} u' = v, \\ v' = -c\, v - \left(1 - \dfrac{\beta}{c} v - \beta\, u \right) e^{-1/u} + \epsilon\, \dfrac{\beta}{c^2} \left(1 - \dfrac{\beta}{c}\, v - \beta\, u \right) e^{-2/u}. \end{cases}$$

We will now show how to use these approximate dynamics to predict the correction to the wavespeed resulting from the inclusion of the finiteness of the Lewis number.

2.4. Perturbative formula for wavespeed. Here, we derive and numerically study a formula for the wavespeed correction in going from $\mathrm{Le} = \infty$ to finite Lewis number. Let

$$\text{(2.12)} \qquad c\,(\beta, \epsilon) = c_0\,(\beta) + \epsilon\, c_1\,(\beta) + \mathcal{O}\left(\epsilon^2\right),$$

where c_0 is the wavespeed associated with the infinite Lewis number ($\epsilon = 0$) combustion wavefront. In the spirit of perturbation analysis, we obtain a formula for the correction $c_1(\beta)$ purely in terms of the unperturbed ($\epsilon = 0$) wave, using a nontraditional application of "Melnikov's method" [29] from dynamical systems theory.

Melnikov's method is applied most commonly to area-preserving two-dimensional systems under time-periodic perturbations [4, 21, 38]. (Here, once again, ξ represents

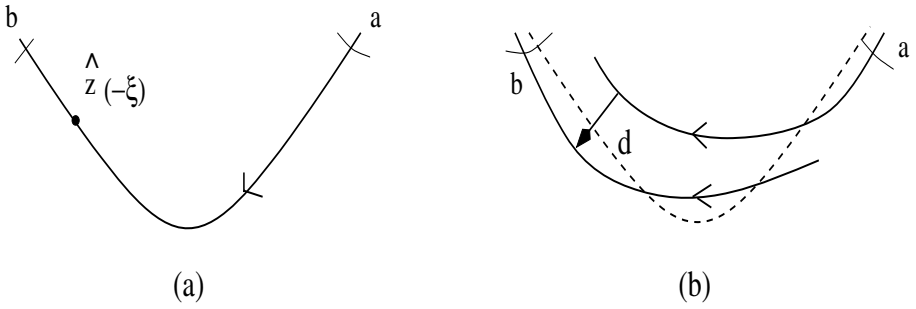


FIG. 2.6. *Manifold structure for the Melnikov approach:* (a) $\epsilon = 0$, (b) $\epsilon \neq 0$.

"time.") Our system (2.11) is not area-preserving and has a perturbation which is independent of the temporal variable. Under these conditions, we describe the method applied to the system

$$\mathbf{z}' = \mathbf{f}(\mathbf{z}) + \epsilon\, \mathbf{g}(\mathbf{z})\,. \tag{2.13}$$

When $\epsilon = 0$, suppose this system possesses a heteroclinic connection between the two saddle fixed points **a** and **b** as shown in Figure 2.6(a). A heteroclinic connection of this sort occurs when a branch of the one-dimensional unstable manifold of **a** coincides with a branch of the stable manifold of **b**. This heteroclinic trajectory can be represented as a solution $\mathbf{z} = \hat{\mathbf{z}}(\xi)$ to (2.13) with $\epsilon = 0$.

Now, for small $\epsilon > 0$ in (2.13), the fixed points **a** and **b** perturb by $\mathcal{O}(\epsilon)$ and retain their stable and unstable manifolds [18]. However, these need no longer coincide. Figure 2.6(b) shows how they can split apart, with the dashed curve showing the original manifold. Let $d(\xi, \epsilon)$ be a distance measure between these manifolds, measured along a perpendicular to the unperturbed heteroclinic drawn at $\hat{\mathbf{z}}(-\xi)$. The variable ξ can thus be used to identify the position along the heteroclinic curve (cf. "heteroclinic coordinates" of Section 4.5 in [38]). Since $d(\xi, 0) = 0$ for all ξ, this distance is Taylor expandable in ϵ in the form

$$d(\xi, \epsilon) = \epsilon\, \frac{M(\xi)}{|\mathbf{f}\,(\hat{\mathbf{z}}(-\xi))|} + \mathcal{O}(\epsilon^2)\,,$$

where the scaling factor $|f\,(\hat{\mathbf{z}}(-\xi))|$ in the denominator represents the unperturbed trajectory's speed at the location ξ. The quantity $M(\xi)$ is the "Melnikov function," for which an expression is

$$M(\xi) = \int_{-\infty}^{\infty} \exp\left[-\int_{-\xi}^{\mu} \nabla \cdot \mathbf{f}(\hat{\mathbf{z}}(s))\, ds\right] \mathbf{f}(\hat{\mathbf{z}}(\mu)) \wedge \mathbf{g}(\hat{\mathbf{z}}(\mu))\, d\mu\,, \tag{2.14}$$

where the wedge product between two vectors is defined by $(a_1, a_2)^T \wedge (b_1, b_2)^T = a_1 b_2 - a_2 b_1$. Obtaining the version (2.14) requires two adjustments to the standard Melnikov approaches [4, 21, 38]: incorporating the nonarea-preserving nature of the unperturbed flow of (2.13), and representing the distance in terms of heteroclinic coordinates. Details are provided in Appendix A. We need to ensure the persistence of a heteroclinic trajectory in (2.11) for $\epsilon > 0$ and thus require $d(\xi, \epsilon) = 0$ for all ξ. For this to happen for all small ϵ, we therefore need to set $M(\xi) \equiv 0$.

To apply this technique to our system, we begin by writing (2.11) in the form (2.13). Using the expansion (2.12), and utilizing binomial expansions for $1/(c_0+\epsilon c_1)$, we get

$$\begin{cases} u' = v, \\ v' = -c_0\, v - e^{-1/u}\Upsilon_{uv} + \epsilon\left(-c_1 v - \dfrac{\beta c_1 e^{-1/u}}{c_0^2} v + \dfrac{\beta e^{-2/u}}{c_0^2}\Upsilon_{uv}\right), \end{cases} \tag{2.15}$$

where higher-order terms in ϵ have been discarded, and

$$\Upsilon_{uv} = 1 - \beta\, u - \frac{\beta}{c_0}\, v\,.$$

By appropriately identifying $\mathbf{f}$ and $\mathbf{g}$ from (2.15) through comparison with (2.13), we see that

$$(\mathbf{f}\wedge\mathbf{g})\,(u,v) = v\left(-c_1 v - \frac{\beta c_1 e^{-1/u} v}{c_0^2} + \frac{\beta e^{-2/u}}{c_0^2}\Upsilon_{uv}\right)$$

and $\nabla\cdot\mathbf{f} = -c_0 + \beta e^{-1/u}/c_0$. Substituting into the Melnikov formula (2.14), and setting it equal to zero, we obtain

$$\int_{-\infty}^{\infty} \exp\left[\int_{-\xi}^{\mu}\left(c_0 - \frac{\beta}{c_0}e^{-1/u(s)}\right) ds\right] v\left(-c_1 v - \frac{\beta v c_1 e^{-1/u}}{c_0^2} + \frac{\beta e^{-2/u}}{c_0^2}\Upsilon_{uv}\right) d\mu = 0\,,$$

where each of $u(\mu)$ and $v(\mu)$ is evaluated along the $\epsilon = 0$ combustion wave. Notice, however, that for this infinite Lewis number combustion wave, (2.4) tells us that the fuel concentration $y(\mu) = \Upsilon_{uv}(\mu)$ for all μ. Therefore

$$c_1(\beta) = \beta\, \frac{\displaystyle\int_{-\infty}^{\infty} \exp\left[\int_{-\xi}^{\mu}\left(c_0 - \frac{\beta}{c_0}e^{-1/u(s)}\right) ds\right] v\, y\, e^{-2/u}\, d\mu}{\displaystyle\int_{-\infty}^{\infty} \exp\left[\int_{-\xi}^{\mu}\left(c_0 - \frac{\beta}{c_0}e^{-1/u(s)}\right) ds\right] v^2(c_0^2 + \beta\, e^{-1/u})\, d\mu}, \tag{2.16}$$

where $u(\mu)$, $v(\mu)$, and $y(\mu)$ in the integrands are obtained from the $\epsilon = 0$ combustion wave discussed in section 2.1. The apparent dependence of c_1 on the wave coordinate ξ is spurious: if I is an antiderivative of the inner integrals in (2.16), a multiplicative term $\exp[-I(-\xi)]$ emerges in both the numerator and denominator, which therefore cancels. Hence, any convenient value for ξ can be chosen in (2.16), for example, 0.

Equation (2.16) is a powerful expression in which the wavespeed correction is expressed purely in terms of the (unperturbed) infinite Lewis number wavefront and system parameters. This correction was obtained through an application of the slow manifold and Melnikov's method (suitably modified). While developed within the current specific context, we note that this technique can in fact be used in a variety of instances which are modeled through coupled reaction-diffusion equations in which the diffusivities are very different.

We note that $v < 0$ for the infinite Lewis number wavefront, as is clear from the phase portrait, Figure 2.1. Alternatively, u is smaller at the front of the wave, where fuel is yet to be burnt, and is therefore a decreasing function of μ, leading to $v = u' < 0$. Based on this, (2.16) immediately displays that $c_1 < 0$, proving the property that the wavespeed decreases when fuel diffusivity is included. This is in agreement with the numerical observations in section 2.2.

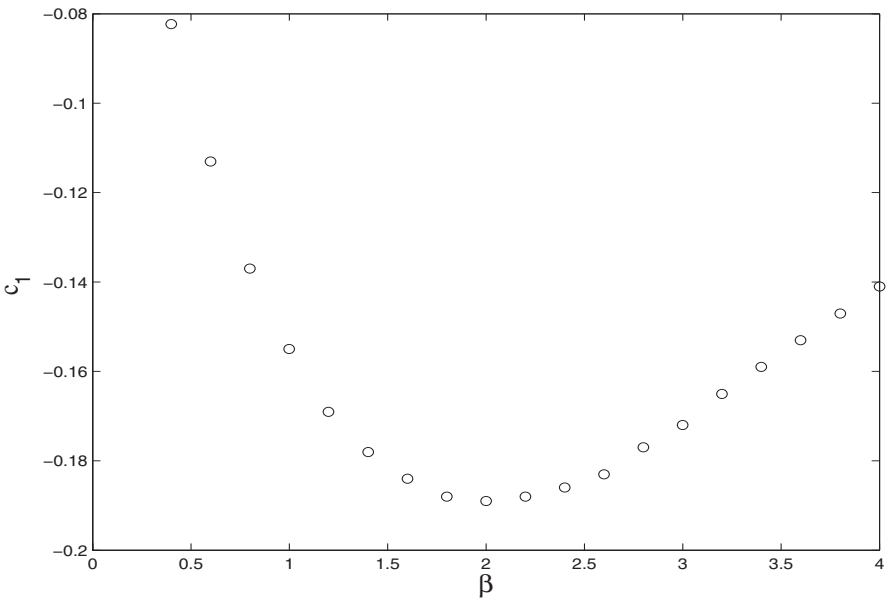


FIG. 2.7. *The perturbing wavespeed as a function of* β.

Equation (2.16) provides an explicit perturbative formula on how the wavespeed varies through the inclusion of the finiteness of the Lewis number, expressed entirely in terms of the infinite Lewis number combustion wave. This result is used to compute the solid line in Figure 2.4, which is the theoretical wavespeed $0.5707 - 0.1552\epsilon$ obtained by using (2.16) and (2.12) when $\beta = 1$. When ϵ is small, it forms an excellent approximation to the numerically obtained wavespeed, as described in section 2.2. Indeed, Figure 2.4 show that the theoretical line is tangential to the curve formed by the closed circles.

The perturbation wavespeed c_1 as a function of β appears in Figure 2.7. There is a value of β (around 2) at which the absolute influence of the finiteness of the Lewis number is greatest. Nevertheless, since c_0 is itself a function of β, it makes sense to investigate the *relative* influence c_1/c_0 of the perturbative term. This is presented in the numerically computed figure, Figure 2.8. The graph is virtually linear and has zero intercept. In other words, the complicated quotient in (2.16) is in fact proportional to the unperturbed wavespeed $c_0(\beta)$, with the proportionality factor *independent of* β. We therefore arrive at the approximation

$$c(\beta, \epsilon) = c_0(\beta)\,[1 - 0.267\,\epsilon\,\beta] = c_0(\beta)\left[1 - 0.267\,\frac{\beta}{\mathrm{Le}}\right], \tag{2.17}$$

for large Lewis numbers, with excellent validity across all β, and with $c_0(\beta)$ also known through (2.6).

Equation (2.17) shows that the wavespeed, as a fraction of the infinite Lewis number wavespeed, acquires a correction linear in the ratio β/Le. We are not aware of any such result being previously reported in the literature of combustion waves. Moreover, the simplicity of this expression is remarkable. For the specific instance $\beta = 1$, we apply this formula in order to arrive at the dashed line in Figure 2.4. Our perturbative theory has clearly given us a very accurate and simple approximation,

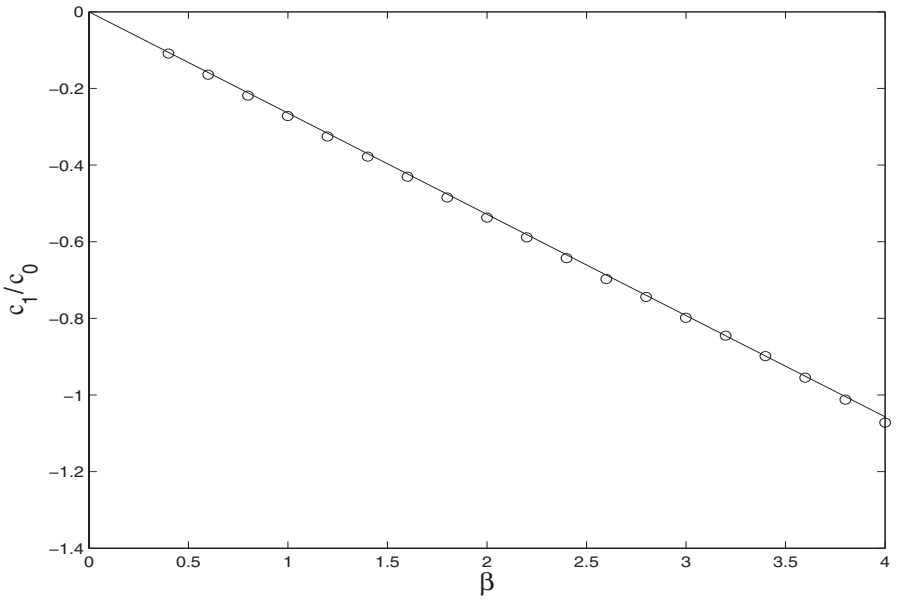


FIG. 2.8. *Relative size of perturbative wavespeed as a function of* β.

and elucidates the straightforward dependence of the wavespeed on the parameters β and Le.

3. Stability analysis. In this section, we test the stability of the combustion wavefront $(u, y) = (u_0(\xi), y_0(\xi))$ we have found as a solution to (1.1) at large Lewis numbers. Consider a perturbation of the form

$$u = u_0(\xi) + U(\xi)\, e^{\lambda t}, \qquad y = y_0(\xi) + Y(\xi)\, e^{\lambda t}. \tag{3.1}$$

At first order, U and Y satisfy an eigenvalue problem

$$\begin{pmatrix} U \\ V \\ Y \\ Z \end{pmatrix}' = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \lambda - \frac{y_0}{u_0^2}\, e^{-1/u_0} & -c & -e^{-1/u_0} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{\beta\, y_0}{\epsilon\, u_0^2}\, e^{-1/u_0} & 0 & \frac{\lambda}{\epsilon} + \frac{\beta}{\epsilon}\, e^{-1/u_0} & -\frac{c}{\epsilon} \end{pmatrix} \begin{pmatrix} U \\ V \\ Y \\ Z \end{pmatrix}. \tag{3.2}$$

Linear instability occurs if there are values of λ in the right half plane for which (3.2) possesses a solution uniformly bounded for all ξ. It turns out that such values of λ can be investigated by analyzing the Evans function [17], which is a complex analytic function $E(\lambda)$ whose zeros correspond to exactly these λ values. If, for example, it can be shown that $E(\lambda)$ has no zeros in the right half plane, the indications from the point spectrum of (3.2) is that the wavefront is stable. If there exist zeros of $E(\lambda)$ in the right half plane, the wavefront is unstable. A description of the Evans function as used in our study is given in Appendix B. This was proposed in [2, 16, 23, 40] and has been used by [20] for a detailed numerical analysis of (1.1). (It must also be mentioned that in the linear stability analysis, it is necessary to consider the essential spectrum associated with (3.2); it turns out that this has no intersection with the right half plane and therefore need not worry us further.)

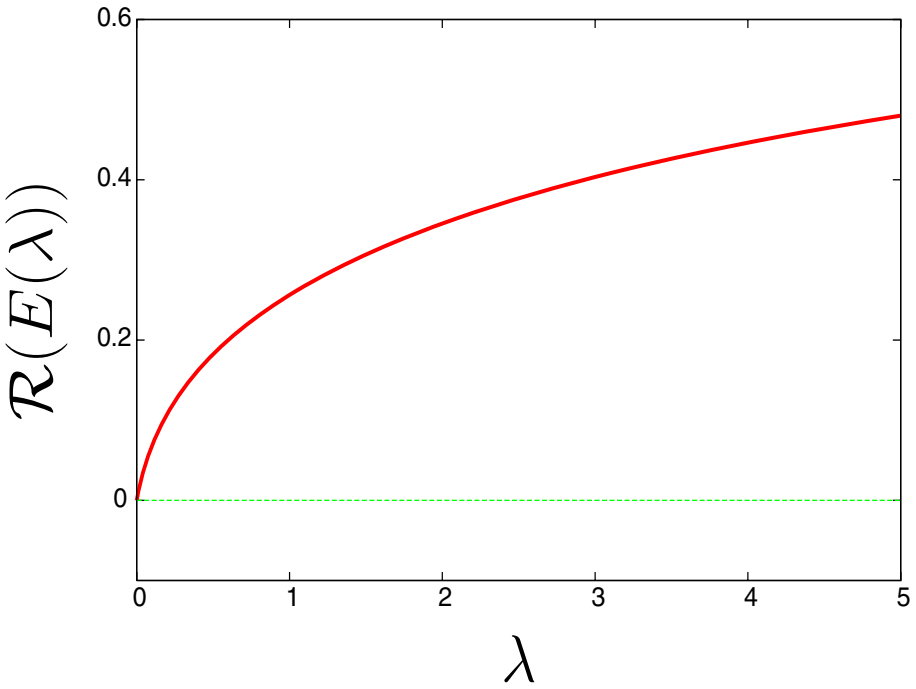


FIG. 3.1. *The real part of the Evans function* $E(\lambda)$ *for* Le=17 *and* $\beta = 1$ *as a function of* λ.

We begin with a traveling wave solution $u_0(\xi)$ and $y_0(\xi)$, obtained using standard shooting methods in section 2, and then compute the Evans function using the procedure outlined in Appendix B. We note that the system is very sensitive due to its stiffness. We found a solution to be accurate enough if we obtain $E(\lambda = 0) \sim \mathcal{O}(10^{-12})$. We are guided in our calculations by the detailed stability analysis of Gubernov et al. [20]. They show, for example, the lack of any eigenvalues of positive real part for small β but show that, for β large enough, two eigenvalues pop into the right half plane exhibiting a Hopf bifurcation. Physically, this corresponds to a pulsating instability in the wavefront, a well-known phenomenon also occurring for $\mathrm{Le} = \infty$ even in slightly different models [5, 14, 28, 37]. Gubernov et al. extend these infinite Lewis number analyses by producing in Figure 5 in [20] the stability boundary in β-ϵ space (their τ is our ϵ). We verify here that our numerically computed wavefronts display the characteristics outlined by them.

In Figure 3.1 we show the Evans function $E(\lambda)$ as it varies with increasing $\lambda \in \mathbb{R}$ for $\mathrm{Le} = 17$ and $\beta = 1$ (this corresponds to a stable regime in Figure 5 of [20]). We find that Evans function does not have any positive real roots. To test for complex roots we vary $\lambda \in i\mathbb{R}$; using Cauchy's theorem we can calculate the winding number to detect possible oscillatory instabilities. In the left panel of Figure 3.2 we show the complex Evans function. Since the system (1.1) is translationally invariant, the Evans function has at least a simple zero at $\lambda = 0$. We checked with a little off-set of the order $\mathcal{O}(10^{-5})$ whether the (real) value of the Evans function at $\lambda = 0$ is shifting towards larger values or smaller values. The off-set allows us to integrate parallel to the imaginary axis of λ and therefore excluding the zero of the Evans function stemming from the root at $\lambda = 0$. This enables us to attribute roots of the Evans function to either the translational mode or to a real instability. For the case depicted

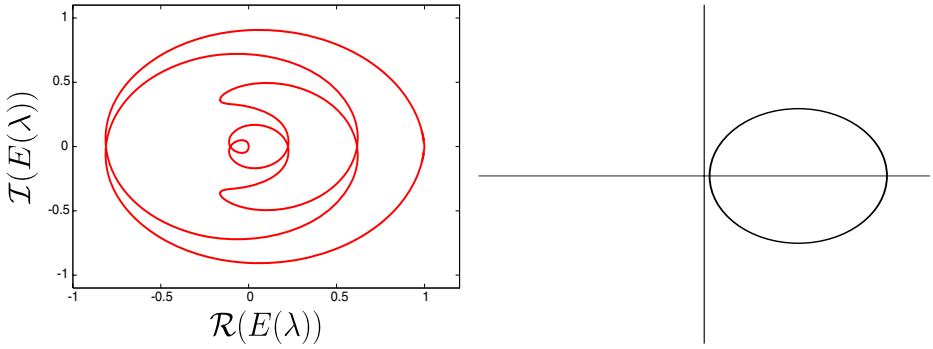


FIG. 3.2. *Left: The real versus the imaginary part of the Evans function* $E(\lambda)$ *for* Le=17 *and* $\beta = 1$. *The spectral parameter* λ *varies along the imaginary axis. Right: A sketch of a topologically equivalent Evans function. The winding number is clearly zero, indicating stability.*

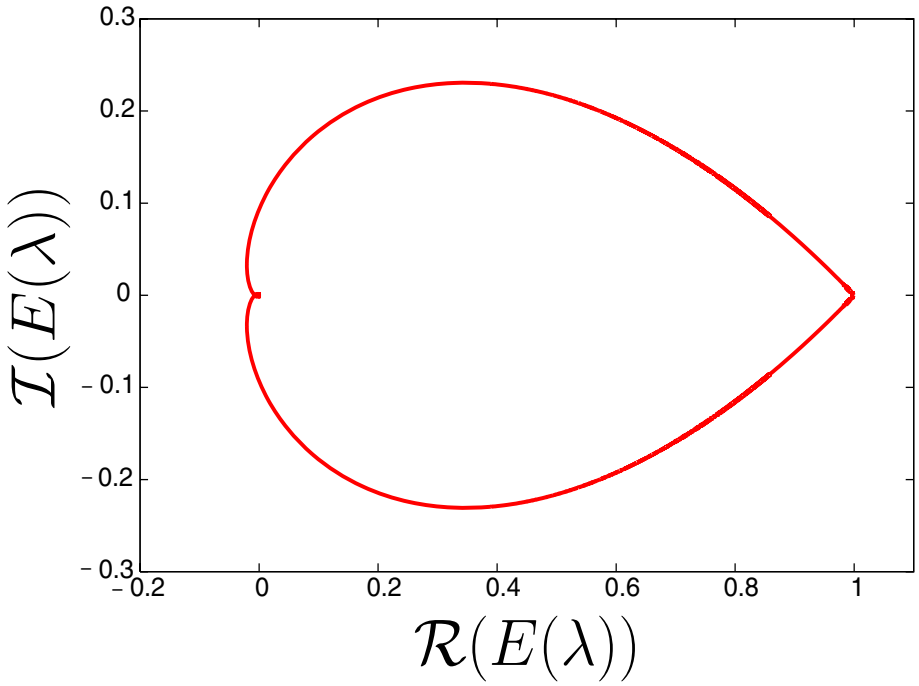


FIG. 3.3. *The real versus the imaginary part of the Evans function* $E(\lambda)$ *for* Le=100 *and* $\beta = 9$. *The spectral parameter* λ *varies along the imaginary axis.*

in Figure 3.2 we find that the Evans function moves to the right. This means that the Evans function can be cast in the topologically equivalent form depicted in the right panel of Figure 3.2 and it clearly has a winding number zero. We therefore find that at these parameter values there are no unstable eigenvalues. (Note that for this argument to work we need our definition of the Evans function to be analytic, which excludes standard methods such as Gram–Schmidt orthogonalizations.)

We next choose Le = 100 and $\beta = 9$, parameters at which (according to Figure 5 in [20]) an oscillatory instability is to be expected. In Figure 3.3 we show the Evans

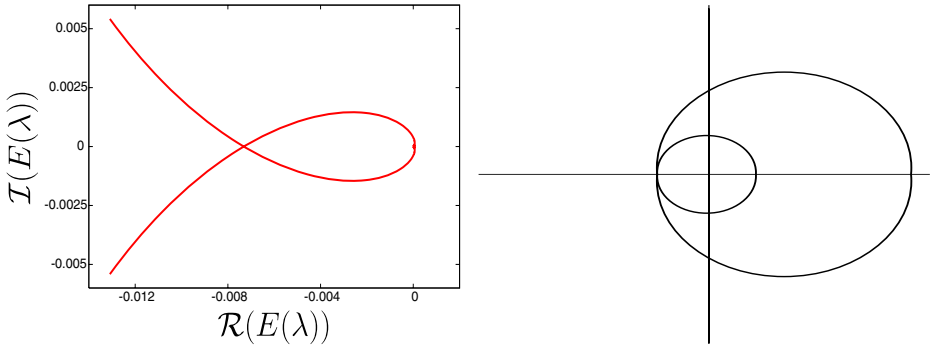


FIG. 3.4. *Left: Closeup of the Evans function depicted in Figure* 3.3 *into the region* $E(\lambda) = 0$. *Right: A sketch of a topologically equivalent Evans function. The winding number is clearly two, indicating an oscillatory instability.*

function in this situation, with a zoom in displayed in Figure 3.4. To determine the winding number we need to check whether or not the small circle of the Evans function in Figure 3.4 includes the zero. We can do so by allowing again for a small off-set of λ. We find that the circle includes zero. Unfolding the behavior of the Evans function then allows us to sketch a topologically equivalent Evans function as in the right panel of Figure 3.4. We verified that the instability is indeed oscillatory by examining the Evans function for $\lambda \in \mathbb{R}^+$, which reveals no zeros. Hence our wavefront displays the predicted characteristics of [20]. Stability properties are not affected unduly by the finiteness of the Lewis number, despite the singularity of this limit.

4. Concluding remarks. In this article, we have studied combustion wavefront in a one-dimensional medium. Our concentration was on very high Lewis numbers relevant to high-density supercritical combustion. We determine the wavespeed as a function of the exothermicity parameter β and the Lewis number Le, by seeking the wavespeed value which establishes a connection between the fixed points corresponding to the fully burnt and the unburnt states. The infinite Lewis number instance reveals an exponential dependence of the wavespeed on β, for which we determine an empirical formula. We then use several suitably modified dynamical systems techniques (slow manifold reduction and Melnikov's method) to compute an explicit formula (2.16) for the correction to the wavespeed when including the effect of a large, but not infinite, Lewis number. We hence obtain a simple formula (2.17) which shows that the relative change in the wavespeed is proportional to β/Le for large Lewis numbers. Our theory is shown to have excellent consistency with the numerically computed wavespeed for large Le, as we show in Figure 2.4.

The stability of the high Lewis number wavefronts is then tested numerically based on the Evans function technique. Our results are in agreement with the stability boundaries presented by Gubernov et al. in Figure 5 in [20].

We remark that the modified Melnikov's method that we have used can in fact be used in more general situations which are described by two coupled reaction-diffusion equations with strongly differing diffusivities. Based on a known wavefront or wave-pulse solution for when the smaller of the diffusivities is zero, our technique can in some instances be used to determine the wavespeed correction resulting from the inclusion of the (previously neglected) diffusivity. Alternatively, it can be adapted to situations in which the wavespeed changes due to some other small parameter. Our analysis of

high Lewis number wavefronts therefore provides a new perturbative methodology for analyzing certain classes of reaction-diffusion equations, pattern formation problems, and combustion waves.

Appendix A. Melnikov function derivation. We briefly outline the modifications needed to the standard Melnikov approaches [4, 21, 38] relevant to section 2.4. Our system is $\dot{\mathbf{z}} = f(\mathbf{z}) + \epsilon\, g(\mathbf{z})$, as given in (2.13). Consider a particular parametrization of the heteroclinic $\hat{\mathbf{z}}(\xi)$. Imagine the perturbed system as embedded in three-dimensional $(\mathbf{z}, s)$ space. In a "time"-slice $s = s_0$, let $\mathcal{T}$ be the normal vector to the heteroclinic drawn at the point $\hat{\mathbf{z}}(0) = \mathbf{z}_0$. The usual approach is to compute the distance between the perturbed manifolds measured along $\mathcal{T}$, and this is expandable as

$$d(s_0, \epsilon) = \epsilon\, \frac{M(s_0)}{|\mathbf{f}(\mathbf{z}_0)|} + \mathcal{O}(\epsilon^2)\,. \tag{A.1}$$

Let $\mathbf{z}^u(s)$ be the trajectory of the perturbed flow which intersects $\mathcal{T}$ and which backwards asymptotes to the perturbed fixed point $\mathbf{a}(\epsilon)$. In other words, $\mathbf{z}^u(s)$ is a trajectory lying on $\mathbf{a}(\epsilon)$'s unstable manifold. The standard approach [4, 21] is to represent

$$\mathbf{z}^\sigma(s) = \hat{\mathbf{z}}(s - s_0) + \epsilon\, \mathbf{z}_1^\sigma(s) + \mathcal{O}(\epsilon^2),$$

where $\sigma = u$ (for "unstable"), valid for $-\infty < s \leq s_0$. A similar expansion on $s_0 \leq s < \infty$ with $\sigma = s$ (for "stable") works for the trajectory $\mathbf{z}^s(s)$, which intersects $\mathcal{T}$ on the time-slice s_0 and which lies on the stable manifold of the perturbed fixed point $\mathbf{b}(\epsilon)$. Then, the standard Melnikov development (see [4, 21]) allows the representation

$$d(s_0, \epsilon) = \epsilon\, \frac{\Delta^u(s_0) - \Delta^s(s_0)}{|\mathbf{f}(\mathbf{z}_0)|} + \mathcal{O}(\epsilon^2)\,, \tag{A.2}$$

where

$$\Delta^\sigma(s) = \mathbf{f}(\hat{\mathbf{z}}(s - s_0)) \wedge \mathbf{z}_1^\sigma(s)$$

for $\sigma = u$ and $\sigma = s$. Now, [4, 21] derive that

$$\dot{\Delta}^\sigma = \nabla \cdot \mathbf{f}(\hat{\mathbf{z}}(s - s_0))\ \Delta^\sigma + \mathbf{f}(\hat{\mathbf{z}}(s - s_0)) \wedge \mathbf{g}(\hat{\mathbf{z}}(s - s_0), s) + \mathcal{O}(\epsilon) \tag{A.3}$$

but, since the unperturbed dynamical system is volume-preserving, have the luxury of ignoring the first term on the right-hand side. We cannot do so here, but we can neglect the second argument in $\mathbf{g}$, since our case is autonomous. To deal with the first term, we multiply (A.3) by the integrating factor

$$\mu(s) = \exp\left[-\int_0^s \nabla \cdot \mathbf{f}(\hat{\mathbf{z}}(r - s_0))\ dr\right]$$

before proceeding. Having done so, we integrate from $-\infty$ to s_0 by choosing $\sigma = u$, then integrate from s_0 to ∞ by choosing $\sigma = s$, and then add the two equations to get

$$\Delta^u(s_0) - \Delta^s(s_0) = \int_{-\infty}^{\infty} \frac{\mu(s)}{\mu(s_0)} \mathbf{f} \wedge \mathbf{g}(\hat{\mathbf{z}}(s - s_0))\ ds\,.$$

(This is an adaptation of the standard process [4, 21].) In conjunction with (A.1) and (A.2), and also employing the shift $s - s_0 \to s$ in the integrand, we obtain the Melnikov function

$$M(s_0) = \int_{-\infty}^{\infty} \exp\left[-\int_0^s \nabla \cdot \mathbf{f}\left(\hat{\mathbf{z}}(r)\right)\, dr\right] \mathbf{f} \wedge \mathbf{g}\left(\hat{\mathbf{z}}(s)\right) ds,$$

which no longer depends on s_0. Having dealt with the nonvolume-preserving instance, the next step is to change our attitude: rather than measuring the distance in a time-slice s but at a *specific* point $\mathbf{z}_0$, we ignore time-slices (since our perturbed system is itself autonomous) and allow the point to vary along the heteroclinic. To do so, choose a *different* parametrization $\hat{\mathbf{w}}(s) = \hat{\mathbf{z}}(s-\xi)$ of the heteroclinic. Thus, the point $\mathbf{w}_0 = \hat{\mathbf{w}}(0) = \hat{\mathbf{z}}(-\xi)$ can be varied along the heteroclinic by choosing different values of ξ. Therefore, ξ will represent different points along the heteroclinic at which the distance measurement is to be made (cf. "heteroclinic coordinates" of [38]). Using the $\mathbf{w}$ trajectory, our earlier results can be expressed as

$$d(s_0, \xi) = \epsilon \frac{M(s_0, \xi)}{|\mathbf{f}\left(\mathbf{w}_0\right)|} + \mathcal{O}(\epsilon^2) = \epsilon \frac{M(\xi)}{|\mathbf{f}\left(\mathbf{z}(-\xi)\right)|} + \mathcal{O}(\epsilon^2), \tag{A.4}$$

where

$$\begin{aligned} M(\xi) &= \int_{-\infty}^{\infty} \exp\left[-\int_0^s \nabla \cdot \mathbf{f}\left(\hat{\mathbf{w}}(r)\right) dr\right] \mathbf{f} \wedge \mathbf{g}\left(\hat{\mathbf{w}}(s)\right) ds \\ &= \int_{-\infty}^{\infty} \exp\left[-\int_0^s \nabla \cdot \mathbf{f}\left(\hat{\mathbf{z}}(r-\xi)\right) dr\right] \mathbf{f} \wedge \mathbf{g}\left(\hat{\mathbf{z}}(s-\xi)\right) ds \\ &= \int_{-\infty}^{\infty} \exp\left[-\int_0^{s+\xi} \nabla \cdot \mathbf{f}\left(\hat{\mathbf{z}}(r-\xi)\right) dr\right] \mathbf{f} \wedge \mathbf{g}\left(\hat{\mathbf{z}}(s)\right) ds \\ &= \int_{-\infty}^{\infty} \exp\left[-\int_{-\xi}^{s} \nabla \cdot \mathbf{f}\left(\hat{\mathbf{z}}(r)\right) dr\right] \mathbf{f} \wedge \mathbf{g}\left(\hat{\mathbf{z}}(s)\right) ds\,. \end{aligned}$$

This, in conjunction with (A.4), is the expression used in section 2.4.

Appendix B. Evans function definition. Here, we describe the Evans function approach for analyzing linear stability. In general, the linear stability of a localized traveling wave solution to a system of PDEs is obtained by studying the eigenvalue problem

$$\mathcal{L}\, w = \lambda\, w, \tag{B.1}$$

where the matrix differential operator $\mathcal{L}$ arises from the linearization of the PDEs. The traveling solution is said to be linearly stable if the spectrum of $\mathcal{L}$ lies in the closed left half plane.

The system (B.1) can be turned into a linear dynamical system of the form

$$U_\xi = \mathbf{A}(\xi, \lambda)\, U, \tag{B.2}$$

where $\mathbf{A}(\xi, \lambda)$ is an $n \times n$ square matrix depending on $\xi = x - c\,t$ and the spectral parameter λ (in our case, $n = 4$). It can be shown that the asymptotic behavior of the solutions to (B.2) is determined by the matrices

$$\mathbf{A}_{\pm\infty}(\lambda) = \lim_{\xi \to \pm\infty} \mathbf{A}(\xi, \lambda)$$

in the following sense (see [11] for details). If μ^+ (resp., μ^-) is an eigenvalue of $\mathbf{A}_{+\infty}$ (resp., $\mathbf{A}_{-\infty}$) with eigenvector v^+ (resp., v^-), then there exists a solution w^+ (resp., w^-) to (B.2) with the property that

$$\lim_{\xi\to\infty} w^+ e^{-\mu^+\xi} = v^+ \quad \left(\text{resp.}, \lim_{\xi\to-\infty} w^- e^{-\mu^-\xi} = v^-\right). \tag{B.3}$$

Note that the superscript "+" refers to exponentially decaying behavior at $\xi = +\infty$, while "−" refers to $\xi = -\infty$.

To study the linear stability, one should consider both the essential and point spectrum of $\mathcal{L}$. The essential spectrum of $\mathcal{L}$ consists of the values of λ for which $\mathbf{A}_\infty$ or $\mathbf{A}_{-\infty}$ has purely imaginary eigenvalues [22]. The point spectrum can be studied by means of the Evans function, first introduced by Evans [16] and later generalized [2]. Roughly speaking, the zeros of this complex-valued function are arranged to coincide with the point spectrum of $\mathcal{L}$.

Let Ω denote a domain of the complex λ plane with no intersection with the essential spectrum and let n_s and n_u denote, respectively, the number of eigenvalues of $\mathbf{A}_\infty$ with negative real part and the number of eigenvalues of $\mathbf{A}_{-\infty}$ with positive real part in Ω. We assume that $n_s + n_u = n$. Let $w_i^+(\lambda,\xi)$, $i = 1, 2, \ldots, n_s$ (resp., $w_i^-(\lambda,\xi)$, $i = 1, 2, \ldots, n_u$), be linearly independent solutions to (B.2) converging to zero as $\xi \to \infty$ (resp., $\xi \to -\infty$) which are analytic of λ in Ω. Clearly, a particular value of λ belongs to the point spectrum of $\mathcal{L}$ if (B.2) admits a solution that is converging to zero for both $\xi \to \pm\infty$, that is, if the space of solutions generated by the w_i^+ intersects with the one generated by the w_i^-. To detect such values of λ in Ω, one can use the definition of the Evans function given in [33],

$$E(\lambda) = \det\left(w_1^+, w_2^+, \ldots, w_{n_s}^+, w_1^-, w_2^-, \ldots, w_{n_u}^-\right),$$

in which the $w_i^\pm$ are evaluated at $\xi = 0$. This function is analytic in Ω and is real for real values of λ, and the locations of the zeros of $E(\lambda)$ correspond to eigenvalues of $\mathcal{L}$.

The first *numerical* computation of the Evans function was by Evans himself in [17] and followed by [32, 35]. However, in all three papers $n_s = 1$, in which case a standard shooting argument can be used. In standard shooting algorithms one follows the stable and/or unstable manifolds at $\xi = \pm\infty$. The Evans function is then given as the intersection of these manifolds. As shown in section 3, our system has $n = 4$ and $n_s = n_u = 2$. This causes the following practical problem: although the n_s (or n_u, respectively) eigenvectors are linear independent solutions of the eigenvalue problem (B.2) at $\xi = \pm\infty$, the numerical integration scheme will lead to an inevitable alignment with the eigendirection corresponding to the largest eigenvalue. This collapse of the eigendirections is usually overcome by using Gram–Schmidt orthogonalization. However, this is not desirable for calculating the Evans function, as it is a nonanalytic procedure which then subsequently prohibits the use of Cauchy's theorem (argument principle) to locate complex zeros of the Evans function. The Evans function is therefore best calculated using exterior algebra [1, 3, 9, 11, 12, 15, 31, 34].

We briefly review the method here, with specific regard to the situation in which $n = 4$ and $n_s = n_u = 2$. For more details the reader is referred to [1, 3, 11, 15] and to the numerical computation in [20]. The main idea behind exterior algebra methods (or compound matrices methods) is that the linear system (B.2) induces a dynamical system on the wedge-space $\bigwedge^2(\mathbb{C}^4)$ for $n_s = n_u = 2$. The wedge-space $\bigwedge^2(\mathbb{C}^4)$ is the

space of all two forms on $\mathbb{C}^n$. This is a space of dimension $\binom{4}{2} = 6$. The induced dynamics on the wedge-space $\bigwedge^2(\mathbb{C}^4)$ can be written as

$$\mathbf{U}_\xi = \mathbf{A}^{(2)}(\xi)\mathbf{U}, \quad \mathbf{U} \in \bigwedge^2(\mathbb{C}^4). \tag{B.4}$$

Here the linear operator (matrix) $\mathbf{A}^{(2)}$ is the restriction of $\mathbf{A}(\xi,\lambda) = \{a_{ij}\}$ to the wedge-space $\bigwedge^2(\mathbb{C}^4)$. Using the standard basis of $\bigwedge^2(\mathbb{C}^4)$

$$\begin{aligned} &\omega_1 = \mathbf{e}_1 \wedge \mathbf{e}_2, \quad \omega_2 = \mathbf{e}_1 \wedge \mathbf{e}_3, \quad \omega_3 = \mathbf{e}_1 \wedge \mathbf{e}_4, \\ &\omega_4 = \mathbf{e}_2 \wedge \mathbf{e}_3, \quad \omega_5 = \mathbf{e}_2 \wedge \mathbf{e}_4, \quad \omega_6 = \mathbf{e}_3 \wedge \mathbf{e}_4, \end{aligned} \tag{B.5}$$

where $\mathbf{e}_{1,2,3,4}$ is the standard basis of $\mathbb{C}^n$, we can find the matrix $\mathbf{A}^{(2)} : \bigwedge^2(\mathbb{C}^4) \to \bigwedge^2(\mathbb{C}^4)$ as a complex 6×6 matrix. With respect to the basis (B.5), $\mathbf{A}^{(2)}$ takes the explicit form

$$\mathbf{A}^{(2)} = \begin{bmatrix} a_{11}+a_{22} & a_{23} & a_{24} & -a_{13} & -a_{14} & 0 \\ a_{32} & a_{11}+a_{33} & a_{34} & a_{12} & 0 & -a_{14} \\ a_{42} & a_{43} & a_{11}+a_{44} & 0 & a_{12} & a_{13} \\ -a_{31} & a_{21} & 0 & a_{22}+a_{33} & a_{34} & -a_{24} \\ -a_{41} & 0 & a_{21} & a_{43} & a_{22}+a_{44} & a_{23} \\ 0 & -a_{41} & a_{31} & -a_{42} & a_{32} & a_{33}+a_{44} \end{bmatrix}.$$

General aspects of the numerical implementation of this theory and details for these constructions in more general systems can be found in [3, 11].

Linearity assures that the induced matrix $\mathbf{A}^{(2)}(\xi,\lambda)$ is also differentiable and analytic. Hence, the limiting matrices,

$$\mathbf{A}^{(2)}_{\pm\infty}(\lambda) = \lim_{\xi\to\pm\infty} \mathbf{A}^{(2)}(\xi,\lambda)\,,$$

will exist. It can readily be shown that the eigenvalues of the matrix $\mathbf{A}^{(2)}_{\pm\infty}(\lambda)$ consist of all possible sums of two eigenvalues of $\mathbf{A}_{\pm\infty}(\lambda)$. Therefore, for $\Re(\lambda) > 0$, the eigenvalue of $\mathbf{A}^{(2)}_{+\infty}(\lambda)$ with the most negative real part is given by $\sigma_+(\lambda) = \mu_1^+ + \mu_2^+$. The eigenvalue $\sigma_+(\lambda)$ has real part strictly less than any other eigenvalue of $\mathbf{A}^{(2)}_{+\infty}(\lambda)$. Analogously, the eigenvalue of $\mathbf{A}^{(2)}_{-\infty}(\lambda)$ with the largest nonnegative real part is given by $\sigma_-(\lambda) = \mu_1^- + \mu_2^-$. The eigenvalue $\sigma_-(\lambda)$ has real part strictly greater than any other eigenvalue of $\mathbf{A}^{(2)}_{-\infty}(\lambda)$. Note that the eigenvalues $\sigma_\pm$ are simple and are analytic functions of λ.

Let $\zeta^\pm(\lambda)$ be the eigenvectors associated with $\sigma_\pm(\lambda)$, defined by

$$\mathbf{A}^{(2)}_{+\infty}(\lambda)\zeta^+(\lambda) = \sigma_+(\lambda)\zeta^+(\lambda) \quad \text{and} \quad \mathbf{A}^{(2)}_{-\infty}(\lambda)\zeta^-(\lambda) = \sigma_-(\lambda)\zeta^-(\lambda)\,. \tag{B.6}$$

These vectors can always be constructed in an analytic way (see [11]) and are readily found to be $\zeta^\pm(\lambda) = v_1^\pm \wedge v_2^\pm$.

Let $\mathbf{U}^\pm(\xi,\lambda) \in \bigwedge^2(\mathbb{C}^4)$ be the solution of the linear system (B.4) which satisfies $\lim_{\xi\to\pm\infty} e^{-\sigma_\pm(\lambda)\xi}\mathbf{U}^\pm(\xi,\lambda) = \zeta^\pm(\lambda)$. This allows us to define the Evans function as

$$E(\lambda) = \mathcal{N}\,\mathbf{U}^-(\xi,\lambda) \wedge \mathbf{U}^+(\xi,\lambda)\,, \quad \lambda \in \Lambda, \tag{B.7}$$

where

$$\mathcal{N} = \mathrm{e}^{-\int_0^\xi \tau(s,\lambda)ds} \quad \text{and} \quad \tau(\xi,\lambda) = \mathrm{Tr}(\mathbf{A}(\xi,\lambda)). \tag{B.8}$$

Expressing $\mathbf{U}^{\pm}(\xi,\lambda)$ as a linear combination with respect to the basis (B.5),

$$\mathbf{U}^{\pm}(\xi,\lambda)=\sum_{j}^{6}U_j^{\pm}\omega_j,$$

the expression for the Evans function (B.8) can be simplified to

$$E(\lambda)=\mathcal{N}\left[\!\left[\mathbf{U}^{-}(\xi,\lambda),\mathbf{\Sigma}\mathbf{U}^{+}(\xi,\lambda)\right]\!\right]_6, \tag{B.9}$$

where $[\![\cdot,\cdot]\!]_6$ is the complex inner product in $\mathbb{C}^4$, and the representation of the Hodge-star operator $\mathbf{\Sigma}$ in the basis (B.5) is

$$\mathbf{\Sigma}=\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 & -1 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & -1 & 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Using the Hodge-star operator, we can relate the most unstable solution $\mathbf{U}^{-}$ of the linearized system at $\xi=-\infty$ with the most unstable solution of the adjoint system of (B.4) at $\xi=-\infty$. Details can be found in [3, 10, 11]. This suggests a normalization of the asymptotic eigenvectors according to

$$[\![\zeta^{-},\mathbf{\Sigma}\zeta^{+}]\!]_6=1, \tag{B.10}$$

which assures that $E(\lambda)\to 1$ for $|\lambda|\to\infty$.

Note that the translational invariance of (1.1) guarantees that the Evans function can be evaluated at any (fixed) spatial location $\xi^{\star}$. However, to avoid unwanted growing of the solutions $\mathbf{U}^{\pm}$ we will consider the scaled solutions

$$\widetilde{\mathbf{U}}^{\pm}(\xi,\lambda)=e^{-\sigma_{\pm}(\lambda)\xi}\mathbf{U}^{\pm}(\xi,\lambda). \tag{B.11}$$

The scaling (B.11) ensures that $\widetilde{\mathbf{U}}^{+}(\xi,\lambda)\big|_{\xi=\xi^{\star}}$ is bounded. The corresponding equation on $\bigwedge^2(\mathbb{C}^4)$,

$$\frac{d}{d\xi}\widetilde{\mathbf{U}}^{\pm}=[\mathbf{A}^{(2)}(\xi,\lambda)-\sigma_{\pm}(\lambda)\mathbf{I}]\widetilde{\mathbf{U}}^{\pm}\,,\quad \widetilde{\mathbf{U}}^{\pm}(\xi,\lambda)\big|_{\xi=L_{\pm\infty}}=\zeta^{\pm}(\lambda)\,, \tag{B.12}$$

is integrated from $\xi=L_{\pm\infty}$ to $\xi=\xi^{\star}$ (where $\xi^{\star}$ is arbitrary but fixed).

The system (B.12) can be integrated using the second-order implicit midpoint method. For a system in the form $\mathbf{U}_{\xi}=\mathbf{B}(\xi,\lambda)\mathbf{U}$, each step of the implicit midpoint rule takes the form

$$\mathbf{U}^{n+1}=[\mathbf{I}-\tfrac{1}{2}\Delta x\mathbf{B}_{n+1/2}]^{-1}[\mathbf{I}+\tfrac{1}{2}\Delta x\mathbf{B}_{n+1/2}]\,\mathbf{U}^{n}, \tag{B.13}$$

where $\mathbf{B}_{n+1/2}=\mathbf{B}(x_{n+1/2},\lambda)$.

Appendix C. The authors wish to thank Harvinder Sidhu, Konstantina Trivisa, Marshall Slemrod, and Joceline Lega for discussions and pointers. Detailed comments and suggestions from two anonymous referees, whose efforts revealed a substantial error in a previous version of this manuscript, are also gratefully acknowledged.

REFERENCES

[1] A. L. Afendikov and T. J. Bridges, *Instability of the Hocking-Stewartson pulse and its implications for the three-dimensional Poiseuille flow*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 457 (2001), pp. 257–272.

[2] J. Alexander, R. Gardner, and C. Jones, *A topological invariant arising in the stability analysis of travelling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[3] L. Allen and T. J. Bridges, *Numerical exterior algebra and the compound matrix method*, Numer. Math., 92 (2002), pp. 197–232.

[4] D. K. Arrowsmith and C. M. Place, *An Introduction to Dynamical Systems*, Cambridge University Press, Cambridge, UK, 1990.

[5] A. Bayliss and B. J. Matkowsky, *Two routes to chaos in condensed phase combustion*, SIAM J. Appl. Math., 50 (1990), pp. 437–459.

[6] A. Bayliss and B. J. Matkowsky, *From traveling waves to chaos in combustion*, SIAM J. Appl. Math., 54 (1994), pp. 147–174.

[7] J. Billingham, *Phase plane analysis of one-dimensional reaction diffusion waves with degenerate reaction terms*, Dyn. Stab. Syst., 15 (2000), pp. 23–33.

[8] J. Billingham and G. N. Mercer, *The effect of heat loss on the propagation of strongly exothermic combustion waves*, Combust. Theory Model., 5 (2001), pp. 319–342.

[9] T. J. Bridges, *The Orr-Sommerfeld equation on a manifold*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 455 (1999), pp. 3019–3040.

[10] T. J. Bridges and G. Derks, *Hodge duality and the Evans function*, Phys. Lett. A, 251 (1999), pp. 363–372.

[11] T. J. Bridges, G. Derks, and G. Gottwald, *Stability and instability of solitary waves of the fifth-order KdV equation: A numerical framework*, Phys. D, 172 (2002), pp. 190–216.

[12] L. Q. Brin, *Numerical testing of the stability of viscous shock waves*, Math. Comp., 70 (2001), pp. 1071–1088.

[13] W. Bush and F. Fendell, *Asymptotic analysis of laminar flame propagation for general Lewis numbers*, Combust. Sci. Tech., 1 (1970), pp. 421–428.

[14] S. A. Cardarelli, D. Golovaty, L. K. Gross, V. T. Gyrya, and J. Zhu, *A numerical study of one-step models of polymerization: Frontal versus bulk mode*, Phys. D, 206 (2005), pp. 145–165.

[15] G. Derks and G. A. Gottwald, *A robust numerical method to study oscillatory instability of gap solitary waves*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 140–158.

[16] J. W. Evans, *Nerve axon equations.* IV. *The stable and the unstable impulse*, Indiana Univ. Math. J., 24 (1974/75), pp. 1169–1190.

[17] J. W. Evans and N. Feroe, *Local stability theory of the nerve impulse*, Math. Biosci., 37 (1977), pp. 23–50.

[18] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971), pp. 193–226.

[19] B. Gray, S. Kalliadasis, A. Lazarovich, C. Macaskill, J. Merkin, and S. Scott, *The suppression of exothermic branched-chain flame through endothermic reaction and radical scavenging*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 458 (2002), pp. 2119–2138.

[20] V. Gubernov, G. N. Mercer, H. S. Sidhu, and R. O. Weber, *Evans function stability of combustion waves*, SIAM J. Appl. Math., 63 (2003), pp. 1259–1275.

[21] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[22] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1981.

[23] C. K. R. T. Jones, *Stability of the travelling wave solution of the FitzHugh-Nagumo system*, Trans. Amer. Math. Soc., 286 (1984), pp. 431–469.

[24] A. Kapila, *Asymptotic Treatment of Chemically Reacting Systems*, Pitman, Boston, 1983.

[25] S. Margolis and S. Johnston, *Multiplicity and stability of supercritical combustion in a nonadiabatic tubular reactor*, Combust. Sci. Tech., 65 (1989), pp. 103–136.

[26] S. B. Margolis and F. A. Williams, *Diffusion/thermal instability of solid propellant flame*, SIAM J. Appl. Math., 49 (1989), pp. 1390–1420.

[27] S. B. Margolis and F. A. Williams, *Flame propagation in solids and high-density fluids with Arrhenius reactant diffusion*, Comb. Flame, 83 (1991), pp. 390–398.

[28] B. J. Matkowsky and G. I. Sivashinsky, *Propagation of a pulsating reaction front in solid fuel combustion*, SIAM J. Appl. Math., 35 (1978), pp. 465–478.

[29] V. K. Melnikov, *On the stability of the centre for time-periodic perturbations*, Trans. Moscow Math. Soc., 12 (1963), pp. 1–56.

[30] G. N. Mercer and R. O. Weber, *Combustion waves in two dimensions and their one-dimensional approximation*, Combust. Theory Model., 1 (1997), pp. 157–165.
[31] B. Ng and W. Reid, *An initial-value method for eigenvalue problems using compound matrices*, J. Comput. Phys., 30 (1979), pp. 125–136.
[32] R. L. Pego, P. Smeraka, and M. I. Weinstein, *Oscillatory instability of solitary waves in a continuum model of lattice vibrations*, Nonlinearity, 8 (1995), pp. 92–941.
[33] B. Sandstede, *Stability of travelling waves*, in Handbook of Dynamical Systems II: Towards Applications, Elsevier, Amsterdam, 2002, pp. 983–1055.
[34] P. Simon, S. Kalliadasis, J. Merkin, and S. Scott, *Evans function analysis of the stability of non-adiabatic flames*, Combust. Theory Model., 7 (2003), pp. 545–561.
[35] J. Swinton and J. Elgin, *Stability of travelling pulse to a laser equation*, Phys. Lett. A, 145 (1990), pp. 428–433.
[36] F. Varas and J. M. Vega, *Linear stability of a plane front in solid combustion at large heat of reaction*, SIAM J. Appl. Math., 62 (2002), pp. 1810–1822.
[37] R. O. Weber, G. N. Mercer, H. S. Sidhu, and B. F. Gray, *Combustion waves for gases (*Le $= 1$*) and solids (*Le $\to \infty$*)*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 453 (1997), pp. 1105–1118.
[38] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, New York, 1990.
[39] J. Xin, *Front propagation in heterogeneous media*, SIAM Rev., 42 (2000), pp. 161–230.
[40] E. Yanagida, *Stability of fast travelling pulse solutions of the FitzHugh-Nagumo equations*, J. Math. Biol., 22 (1985), pp. 81–104.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 487–511

ON THE WAVE STRUCTURE OF TWO-PHASE FLOW MODELS*

STEINAR EVJE† AND TORE FLÅTTEN‡

Abstract. We explore the relationship between two common two-phase flow models, usually denoted as the *two-fluid* and *drift-flux* models. They differ in their mathematical description of momentum transfer between the phases. In this paper we provide a framework in which these two model formulations are unified. The drift-flux model employs a mixture momentum equation and treats interphasic momentum exchange indirectly through the *slip relation*, which gives the relative velocity between the phases as a function of the flow parameters. This closure law is in general highly complex, which makes it difficult to analyze the model algebraically. To facilitate the analysis, we express the quasi-linear formulation of the drift-flux model as a function of three parameters: the derivatives of the slip with respect to the vector of unknown variables. The wave structure of this model is investigated using a perturbation technique. Then we rewrite the drift-flux model with a general slip relation such that it is expressed in terms of the canonical two-fluid form. That is, we replace the mixture momentum equation and the slip relation with equivalent evolution equations for the momentums of each phase. We obtain a mathematically equivalent formulation in terms of a two-fluid model and by this bridge some of the gap between the drift-flux model and the two-fluid model. Finally, the effect of the various exchange terms on the wave structure of the two-fluid model is investigated.

Key words. hyperbolic system of conservation laws, two phase flow, drift-flux model, two-fluid model, perturbation method, eigenvalues, interface friction

AMS subject classifications. 15A90, 35L65, 35P15, 76T10

DOI. 10.1137/050633482

1. Introduction.

In general, multiphase flows exhibit a complex dynamical behavior, where depending on the physical parameters several different *flow regimes* may occur. Flow regimes are commonly divided into *separated* (stratified, annular) and *mixed* (bubbly, dispersed) flows.

There exists no simple model formulation able to describe all these phenomena adequately. Rather, a variety of different models have been suggested with different applications in mind; see, for instance, [6, 8, 23, 24].

A classical way to obtain tractable models is to average in space. Of such models, two particular strategies have attracted considerable interest in the petroleum industry: the *two-fluid* [4, 20] and *drift-flux* [22] models. These models, described in sections 2 and 3, are the focus of the current paper.

The models contain a significant amount of additional closure laws. These closure relations typically depend on the flow structure and represent the main difficulty in the model formulation.

As noted by Bouré [9], the effect of closure relations may be viewed on two different levels:

1. Their *physical magnitude* affects the predicted values of the flow parameters.

*Received by the editors June 13, 2005; accepted for publication (in revised form) October 4, 2006; published electronically February 9, 2007. This work was supported by the Research Council of Norway through the BeMatA program.

http://www.siam.org/journals/siap/67-2/63348.html

†International Research Institute of Stavanger (IRIS), Prof. Olav Hanssensvei 15, NO-4091 Stavanger, Norway (steinar.evje@irisresearch.no).

‡Centre of Mathematics for Applications (CMA), 1053 Blindern, NO-0316 Oslo, Norway (tore.flatten@cma.uio.no).

2. Their *mathematical form* affects the propagation properties of the flow model. That is, differential closure terms affect the velocities and composition of the predicted waves, whereas nondifferential terms do not.

The drift-flux model and its closure relations are commonly formulated to model *mixed* flow regimes. Depending on the closure relations, the two-fluid model has more general validity. In its most basic form, it is nevertheless best suited for a description of *separated* flows. These different domains of applicability manifest themselves through the different wave structures of the common formulation of the two models.

The purpose of this paper is twofold.

(I) Primary aim. To demonstrate how *nondifferential* closure relations for the drift-flux model may be transformed into corresponding *differential* relations in the two-fluid framework. By this transformation, we obtain a two-fluid model whose underlying mathematical structure is identical to the original drift-flux model. Hence it becomes possible to alternate between the two formulations within a unified framework.

(II) Secondary aim. To demonstrate how the wave structure of the drift-flux model may be investigated by a perturbation technique, first applied to two-phase flows by Toumi and coworkers [27, 28], who considered the two-fluid model.

The paper is organized as follows: In sections 2 and 3 we describe the two-fluid and drift-flux models in question. Section 4 is dedicated to the secondary aim of the paper; here we investigate the wave structure of the drift-flux model.

In section 5 we confront the primary aim of our paper, writing the drift-flux model in the framework of a two-fluid model. A main result is equation (118), the explicit form of the interface friction that makes the two-fluid model mathematically equivalent to a general drift-flux model.

Armed with a thorough understanding of the mathematical structure of both models, we demonstrate in section 6 how the wave velocities of the two-fluid model gradually change by addition of the different terms of (118). This illustrates the *physical effects* of the different closure terms on the wave phenomena inherent in the models.

2. Two-fluid model. To be consistent with the dynamical behavior of the flow physics, the two-phase models we consider must describe the following wave phenomena:

- *Sonic waves*, conveying rapid variations in the pressure and the associated velocity fields. They are a consequence of the compressibility of the flow.
- *Material waves*, conveying large scale variations in the volumetric phase fractions and mixture density. They are responsible for the dynamics corresponding to mass transport.
- *Entropy waves*, representing thermodynamic properties transported along the flow.

As noted, for instance, by [9, 27], the entropy waves are uncoupled from the remaining wave structure. Phasic entropies are simply advected with the fluid velocities.

Hence the structure of the sonic and material waves may be studied with no loss of generality by considering only *isentropic* flow models. Such models are based on the physical principle of conservation of the mass and momentum variables, neglecting dynamic energy transfers.

Supplemented by proper closure relations, the models hence consist of mass and momentum balance equations, expressed in the form of *partial differential* equations.

2.1. Model formulation. For a gas (g) and a liquid (ℓ) phase, the isentropic two-fluid model may be written as follows:

- Conservation of mass

$$\frac{\partial}{\partial t}\left(\rho_{\mathrm{g}}\alpha_{\mathrm{g}}\right)+\frac{\partial}{\partial x}\left(\rho_{\mathrm{g}}\alpha_{\mathrm{g}}v_{\mathrm{g}}\right)=0, \tag{1}$$

$$\frac{\partial}{\partial t}\left(\rho_{\ell}\alpha_{\ell}\right)+\frac{\partial}{\partial x}\left(\rho_{\ell}\alpha_{\ell}v_{\ell}\right)=0, \tag{2}$$

- Momentum balances

$$\frac{\partial}{\partial t}\left(\rho_{\mathrm{g}}\alpha_{\mathrm{g}}v_{\mathrm{g}}\right)+\frac{\partial}{\partial x}\left(\rho_{\mathrm{g}}\alpha_{\mathrm{g}}v_{\mathrm{g}}^{2}\right)+\frac{\partial}{\partial x}\left(\alpha_{\mathrm{g}}p_{\mathrm{g}}\right)-p^{\mathrm{i}}\frac{\partial}{\partial x}\left(\alpha_{\mathrm{g}}\right)=Q_{\mathrm{g}}+M_{\mathrm{g}}^{\mathrm{i}}, \tag{3}$$

$$\frac{\partial}{\partial t}\left(\rho_{\ell}\alpha_{\ell}v_{\ell}\right)+\frac{\partial}{\partial x}\left(\rho_{\ell}\alpha_{\ell}v_{\ell}^{2}\right)+\frac{\partial}{\partial x}\left(\alpha_{\ell}p_{\ell}\right)-p^{\mathrm{i}}\frac{\partial}{\partial x}\left(\alpha_{\ell}\right)=Q_{\ell}+M_{\ell}^{\mathrm{i}}. \tag{4}$$

Here α_k is the volume fraction of phase k with

$$\alpha_{\mathrm{g}}+\alpha_{\ell}=1, \tag{5}$$

where ρ_k, p_k, and v_k denote the density, pressure, and fluid velocity of phase k, respectively, and p^{i} is the pressure at the gas-liquid interface. M_k^{i} represents interphasic momentum exchange terms with $M_{\mathrm{g}}^{\mathrm{i}}+M_{\ell}^{\mathrm{i}}=0$. Momentum sources acting on each phase separately, such as wall friction or gravitational forces, are represented by the terms Q_k.

2.2. Closure relations. The closure relations needed to complete the model may be divided into three groups.

2.2.1. Thermodynamic submodels. For each phase k, the thermodynamic *state relation*

$$p_k=p(\rho_k,S_k) \tag{6}$$

must be specified. Here S_k is the entropy of phase k. Furthermore, the interface pressure p^{i} must be expressed as a function of the phasic pressures:

$$p^{\mathrm{i}}=p^{\mathrm{i}}(p_{\mathrm{g}},p_{\ell}). \tag{7}$$

When the flows are separated due to gravitational forces, the relationships between the pressures p^{i}, p_{g}, and p_{ℓ} are commonly chosen to model the effects of *hydrostatics*. In this case, the two-fluid model is able to describe travelling surface waves on the gas-liquid interface; see, for instance, [2].

2.2.2. Phase-specific source terms. The main momentum sources acting on each phase separately are the following:

- *Gravity.*
  The effect of gravitational acceleration is expressed by

$$Q_k=-\rho_k\alpha_k g\sin\theta, \tag{8}$$

  where θ is the angle of the flow direction with respect to the horizontal.

- *Wall friction.*
  For separated flows, the wall friction for each phase is commonly expressed in terms of *friction factors* as follows:

$$Q_k = -f_k \frac{\rho_k \left| v_k \right| v_k}{2}. \tag{9}$$

  The Blasius equation is commonly used for calculating f_k; see, for instance, [1, 25]. According to [7], most *mixed* flow regimes may be modeled to acceptable accuracy by using friction factors corresponding to one-phase liquid flow ($f_{\mathrm{g}} = 0$).

2.2.3. Interphasic momentum exchange terms. The interactions between the phases are highly complex and different in character for each flow regime. Hence these terms are notoriously difficult to derive from theoretical considerations. Nor are they easily determined from experimental data, as their effects are only indirectly visible. We here briefly describe two of the most common approaches for modeling the interphasic momentum exchange, applied to separated and mixed flows, respectively.

- *Stratified flows.*
  For stratified flows it is common [1, 25] to express the interphasic momentum exchange in nondifferential form, as a function of a *friction factor* f_{i}:

$$M_\ell^{\mathrm{i}} = -M_{\mathrm{g}}^{\mathrm{i}} = f_{\mathrm{i}} \frac{\rho_{\mathrm{g}} \left| v_{\mathrm{g}} - v_\ell \right| (v_{\mathrm{g}} - v_\ell)}{2}. \tag{10}$$

  Andritsos and Hanratty [1] noted that waves existing on the gas-liquid interface have a significant effect on the magnitude of f_{i}. They suggested that for sufficiently small gas flow rates $\alpha_{\mathrm{g}} v_{\mathrm{g}} < U_{\mathrm{crit}}$, such that no waves are generated at the interface,

$$f_{\mathrm{i}} \approx f_{\mathrm{g}}. \tag{11}$$

  For $\alpha_{\mathrm{g}} v_{\mathrm{g}} > U_{\mathrm{crit}}$ they developed a correlation where $f_{\mathrm{i}}/f_{\mathrm{g}}$ increases linearly with $\alpha_{\mathrm{g}} v_{\mathrm{g}}$.
- *Bubbly flows.* For a two-phase mixture of gas dispersed within the liquid, the momentum transfer induced by a gas bubble *accelerating* with respect to the surrounding fluid must be taken into account. This effect, denoted as the *virtual mass force*, has been analyzed by Drew, Cheng, and Lahey [10]. By imposing the condition that this interface friction is invariant under a change of reference frame, they derived the expression

$$\begin{aligned} M_{\mathrm{g}}^{\mathrm{i}} = \alpha_{\mathrm{g}} \rho_\ell C_{\mathrm{vm}} \Big(& \partial_t \left(v_{\mathrm{g}} - v_\ell \right) + v_{\mathrm{g}} \partial_x \left(v_{\mathrm{g}} - v_\ell \right) \\ & + \left(v_{\mathrm{g}} - v_\ell \right) \left((\lambda - 2) \partial_x v_{\mathrm{g}} + (1 - \lambda) \partial_x v_\ell \right) \Big), \end{aligned} \tag{12}$$

  where λ and C_{vm} (the coefficient of virtual mass) are volume fraction dependent parameters. The value of C_{vm} is expected to be $1/2$ for noninteracting spheres and smaller for bubbles of other shapes.
  The wave structure of the two-fluid model with virtual mass force included has been analyzed in [18, 19, 28]. In particular, Lahey [19] discusses similarities between such a two-fluid model and the drift-flux model.

2.3. Canonical formulation. The multitude of possible closure relations gives rise to a large class of slightly different models, all falling under the heading of *two-fluid models*. In the following, we will find it useful to base our analyses on some common formulation of these models. By neglecting the phasic pressure difference ($p = p_{\mathrm{g}} = p_\ell$) and writing

$$\tau_{\mathrm{i}} = \left(p - p^{\mathrm{i}}\right) \frac{\partial \alpha_{\mathrm{g}}}{\partial x} - M^{\mathrm{i}}_{\mathrm{g}} = -\left(p - p^{\mathrm{i}}\right) \frac{\partial \alpha_\ell}{\partial x} + M^{\mathrm{i}}_\ell, \tag{13}$$

we arrive at the following *canonical* two-fluid model:

- Conservation of mass

$$\frac{\partial}{\partial t}\left(\rho_{\mathrm{g}} \alpha_{\mathrm{g}}\right) + \frac{\partial}{\partial x}\left(\rho_{\mathrm{g}} \alpha_{\mathrm{g}} v_{\mathrm{g}}\right) = 0, \tag{14}$$

$$\frac{\partial}{\partial t}\left(\rho_\ell \alpha_\ell\right) + \frac{\partial}{\partial x}\left(\rho_\ell \alpha_\ell v_\ell\right) = 0, \tag{15}$$

- Momentum balances

$$\frac{\partial}{\partial t}\left(\rho_{\mathrm{g}} \alpha_{\mathrm{g}} v_{\mathrm{g}}\right) + \frac{\partial}{\partial x}\left(\rho_{\mathrm{g}} \alpha_{\mathrm{g}} v_{\mathrm{g}}^2\right) + \alpha_{\mathrm{g}} \frac{\partial}{\partial x}\left(p\right) + \tau_{\mathrm{i}} = Q_{\mathrm{g}}, \tag{16}$$

$$\frac{\partial}{\partial t}\left(\rho_\ell \alpha_\ell v_\ell\right) + \frac{\partial}{\partial x}\left(\rho_\ell \alpha_\ell v_\ell^2\right) + \alpha_\ell \frac{\partial}{\partial x}\left(p\right) - \tau_{\mathrm{i}} = Q_\ell, \tag{17}$$

where the interfacial momentum exchange term τ_{i} may or may not contain differential operators.

3. Drift-flux model. A strategy to avoid the modeling difficulties associated with the momentum exchange terms, as mentioned in the previous section, is to reformulate the model such that these terms no longer directly appear. This is precisely the idea of the *drift-flux* formulation of two-phase flow. By making the simplifying assumption

$$p = p_{\mathrm{g}} = p_\ell, \tag{18}$$

and adding the two momentum equations (3) and (4), we obtain the conservation equation for the *mixture* momentum:

$$\frac{\partial}{\partial t}(\rho_{\mathrm{g}} \alpha_{\mathrm{g}} v_{\mathrm{g}} + \rho_\ell \alpha_\ell v_\ell) + \frac{\partial}{\partial x}\left(\rho_{\mathrm{g}} \alpha_{\mathrm{g}} v_{\mathrm{g}}^2 + \rho_\ell \alpha_\ell v_\ell^2 + p\right) = Q_{\mathrm{g}} + Q_\ell. \tag{19}$$

Note that (18) is consistent with the assumption of a *mixed* flow regime, which is the situation for which the drift-flux model is commonly applied.

The phasic momentums must satisfy a *slip relation* in the functional form

$$v_{\mathrm{g}} - v_\ell = \Phi(p, \alpha_{\mathrm{g}}, v_{\mathrm{g}}). \tag{20}$$

Hence the two momentum *evolution equations* (16)–(17) of the two-fluid model are replaced by one evolution equation (19) and one *functional relation* (20). Bouré [9] discusses *generalized* drift-flux models where Φ may also contain differential operators.

3.1. Model formulation. In summary, using the nomenclature

$$m_{\mathrm{g}} = \rho_{\mathrm{g}}\alpha_{\mathrm{g}}, \tag{21}$$

$$m_{\ell} = \rho_{\ell}\alpha_{\ell}, \tag{22}$$

$$I_{\mathrm{g}} = m_{\mathrm{g}}v_{\mathrm{g}}, \tag{23}$$

$$I_{\ell} = m_{\ell}v_{\ell}, \tag{24}$$

$$I = I_{\mathrm{g}} + I_{\ell}, \tag{25}$$

$$Q = Q_{\mathrm{g}} + Q_{\ell}, \tag{26}$$

we may express the drift-flux model as

$$\frac{\partial m_{\mathrm{g}}}{\partial t} + \frac{\partial I_{\mathrm{g}}}{\partial x} = 0, \tag{27}$$

$$\frac{\partial m_{\ell}}{\partial t} + \frac{\partial I_{\ell}}{\partial x} = 0, \tag{28}$$

$$\frac{\partial I}{\partial t} + \frac{\partial}{\partial x}\left(I_{\mathrm{g}}v_{\mathrm{g}} + I_{\ell}v_{\ell} + p\right) = Q, \tag{29}$$

supplemented with the following functional relations:

- *Thermodynamics:* $p = p(\rho_{\mathrm{g}}) = p(\rho_{\ell})$.
- *Slip relation:* $v_{\mathrm{g}} - v_{\ell} = \Phi(m_{\mathrm{g}}, m_{\ell}, v_{\mathrm{g}})$.

3.2. Quasi-linear formulation. The model (27)–(29) may be written in the following *quasi-linear form*:

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}(\mathbf{U})\frac{\partial \mathbf{U}}{\partial x} = \mathbf{Q}(\mathbf{U}), \tag{30}$$

where

$$\mathbf{U} = \begin{bmatrix} m_{\mathrm{g}} \\ m_{\ell} \\ I \end{bmatrix} \tag{31}$$

and

$$\mathbf{Q}(\mathbf{U}) = \begin{bmatrix} 0 \\ 0 \\ Q \end{bmatrix}. \tag{32}$$

In the following, we will derive an expression for the Jacobi matrix $\mathbf{A}$. Towards this aim, we will follow the common practice of thermodynamics and take

$$\left(\frac{\partial X}{\partial Y}\right)_{a,b} \tag{33}$$

to mean the partial derivative of X with respect to Y under the assumption of constant a and b.

3.2.1. Some definitions. We now define the following basic abbreviations:

$$\mu_{\mathrm{g}} = \left(\frac{\partial \Phi}{\partial m_{\mathrm{g}}}\right)_{m_\ell, v_{\mathrm{g}}}, \tag{34}$$

$$\mu_\ell = \left(\frac{\partial \Phi}{\partial m_\ell}\right)_{m_{\mathrm{g}}, v_{\mathrm{g}}}, \tag{35}$$

$$\mu_{\mathrm{v}} = \left(\frac{\partial \Phi}{\partial v_{\mathrm{g}}}\right)_{m_{\mathrm{g}}, m_\ell}, \tag{36}$$

$$\zeta = \left(\frac{\partial v_\ell}{\partial v_{\mathrm{g}}}\right)_{m_{\mathrm{g}}, m_\ell}. \tag{37}$$

We further define the *pseudomass* $\hat{\rho}$:

$$\hat{\rho} = m_{\mathrm{g}} + \zeta m_\ell. \tag{38}$$

Remark 1. We observe that by writing (20) as

$$d\Phi = dv_{\mathrm{g}} - dv_\ell, \tag{39}$$

we obtain from (36) and (37) the basic relation

$$\mu_{\mathrm{v}} = 1 - \zeta. \tag{40}$$

We may now derive the following useful differentials.

Differential 1 (gas velocity). *We may expand* dI *as*

$$dI = m_{\mathrm{g}}\, dv_{\mathrm{g}} + v_{\mathrm{g}}\, dm_{\mathrm{g}} + v_\ell\, dm_\ell + m_\ell\, dv_\ell. \tag{41}$$

Using (39) *and*

$$d\Phi = \mu_{\mathrm{g}}\, dm_{\mathrm{g}} + \mu_\ell\, dm_\ell + \mu_{\mathrm{v}}\, dv_{\mathrm{g}}, \tag{42}$$

we obtain

$$dv_{\mathrm{g}} = \frac{1}{\hat{\rho}} \left(dI + (m_\ell \mu_{\mathrm{g}} - v_{\mathrm{g}})\ dm_{\mathrm{g}} + (m_\ell \mu_\ell - v_\ell)\ dm_\ell\right). \tag{43}$$

Differential 2 (gas momentum). *Using*

$$dI_{\mathrm{g}} = m_{\mathrm{g}}\, dv_{\mathrm{g}} + v_{\mathrm{g}}\, dm_{\mathrm{g}}, \tag{44}$$

we obtain from (43)

$$dI_{\mathrm{g}} = \frac{1}{\hat{\rho}} \left(m_{\mathrm{g}}\, dI + (m_{\mathrm{g}} m_\ell \mu_{\mathrm{g}} + \zeta m_\ell v_{\mathrm{g}})\ dm_{\mathrm{g}} + (m_{\mathrm{g}} m_\ell \mu_\ell - m_{\mathrm{g}} v_\ell)\ dm_\ell\right). \tag{45}$$

Differential 3 (liquid momentum). *Using*

$$dI = dI_{\mathrm{g}} + dI_\ell, \tag{46}$$

we obtain from (45)

$$dI_\ell = \frac{1}{\hat{\rho}} \left(\zeta m_\ell\, dI - (m_{\mathrm{g}} m_\ell \mu_{\mathrm{g}} + \zeta m_\ell v_{\mathrm{g}})\ dm_{\mathrm{g}} - (m_{\mathrm{g}} m_\ell \mu_\ell - m_{\mathrm{g}} v_\ell)\ dm_\ell\right). \tag{47}$$

DIFFERENTIAL 4 (pressure). *Writing* $\alpha_{\rm g} + \alpha_\ell = 1$ *as*

$$\frac{m_{\rm g}}{\rho_{\rm g}(p)} + \frac{m_\ell}{\rho_\ell(p)} = 1, \tag{48}$$

we obtain by differentiation

$$dp = \kappa\left(\rho_\ell\, dm_{\rm g} + \rho_{\rm g}\, dm_\ell\right), \tag{49}$$

where

$$\kappa = \frac{1}{(\partial\rho_{\rm g}/\partial p)\,\rho_\ell\alpha_{\rm g} + (\partial\rho_\ell/\partial p)\,\rho_{\rm g}\alpha_\ell}. \tag{50}$$

DIFFERENTIAL 5 (gas momentum convection). *We have*

$$d\left(I_{\rm g}v_{\rm g}\right) = I_{\rm g}\, dv_{\rm g} + v_{\rm g}\, dI_{\rm g}. \tag{51}$$

Hence from (43) *and* (45) *we obtain*

$$\begin{aligned} d\left(I_{\rm g}v_{\rm g}\right) = \frac{1}{\hat\rho}\Big(&2m_{\rm g}v_{\rm g}\, dI + \left(2m_{\rm g}m_\ell v_{\rm g}\mu_{\rm g} + (\zeta m_\ell - m_{\rm g})v_{\rm g}^2\right)\, dm_{\rm g} \\ &+ (2m_{\rm g}m_\ell v_{\rm g}\mu_\ell - 2m_{\rm g}v_{\rm g}v_\ell)\, dm_\ell\Big). \end{aligned} \tag{52}$$

DIFFERENTIAL 6 (liquid momentum convection). *We have*

$$dv_\ell = dv_{\rm g} - d\Phi = \zeta\, dv_{\rm g} - \mu_{\rm g}\, dm_{\rm g} - \mu_\ell\, dm_\ell. \tag{53}$$

From (43) *we obtain*

$$dv_\ell = \frac{1}{\hat\rho}\left(\zeta\, dI - (m_{\rm g}\mu_{\rm g} + \zeta v_{\rm g})\, dm_{\rm g} - (m_{\rm g}\mu_\ell + \zeta v_\ell)\, dm_\ell\right). \tag{54}$$

Hence from

$$d\left(I_\ell v_\ell\right) = I_\ell\, dv_\ell + v_\ell\, dI_\ell \tag{55}$$

we obtain

$$\begin{aligned} d\left(I_\ell v_\ell\right) = \frac{1}{\hat\rho}\Big(&2\zeta m_\ell v_\ell\, dI - (2m_{\rm g}m_\ell v_\ell\mu_{\rm g} + 2\zeta m_\ell v_{\rm g}v_\ell)\, dm_{\rm g} \\ &- \left(2m_{\rm g}m_\ell v_\ell\mu_\ell + (\zeta m_\ell - m_{\rm g})v_\ell^2\right)\, dm_\ell\Big). \end{aligned} \tag{56}$$

3.2.2. The Jacobi matrix. With the aid of these differentials we can more or less directly write down the Jacobi matrix

$$\mathbf{A}(\mathbf{U}) = \frac{1}{\hat\rho}\begin{bmatrix} m_{\rm g}m_\ell\mu_{\rm g} + \zeta m_\ell v_{\rm g} & m_{\rm g}m_\ell\mu_\ell - m_{\rm g}v_\ell & m_{\rm g} \\ -(m_{\rm g}m_\ell\mu_{\rm g} + \zeta m_\ell v_{\rm g}) & m_{\rm g}v_\ell - m_{\rm g}m_\ell\mu_\ell & \zeta m_\ell \\ A_{31} & A_{32} & 2(m_{\rm g}v_{\rm g} + \zeta m_\ell v_\ell) \end{bmatrix}, \tag{57}$$

where

$$A_{31} = \kappa\hat\rho\rho_\ell + 2m_{\rm g}m_\ell\mu_{\rm g}(v_{\rm g} - v_\ell) + (\zeta m_\ell - m_{\rm g})v_{\rm g}^2 - 2\zeta m_\ell v_{\rm g}v_\ell \tag{58}$$

and

$$A_{32} = \kappa\hat\rho\rho_{\rm g} + 2m_{\rm g}m_\ell\mu_\ell(v_{\rm g} - v_\ell) - (\zeta m_\ell - m_{\rm g})v_\ell^2 - 2m_{\rm g}v_{\rm g}v_\ell. \tag{59}$$

4. Wave structure analysis. As is well known from the theory of hyperbolic conservation laws, the velocities of the inherent wave phenomena of the system (30) are given by the eigenvalues of $\mathbf{A}$.

These eigenvalues satisfy the characteristic equation

$$\begin{aligned}(\lambda - v_{\mathrm{g}})(\lambda - v_\ell)(\hat{\rho}\lambda - m_{\mathrm{g}} v_{\mathrm{g}} - \zeta m_\ell v_\ell) + m_{\mathrm{g}} m_\ell \left(\mu_\ell (\lambda - v_{\mathrm{g}})^2 - \mu_{\mathrm{g}} (\lambda - v_\ell)^2\right) \\ + \kappa \rho_{\mathrm{g}} \rho_\ell \left(\alpha_{\mathrm{g}} \alpha_\ell (\rho_{\mathrm{g}} \mu_{\mathrm{g}} - \rho_\ell \mu_\ell) - \alpha_{\mathrm{g}} (\lambda - v_\ell) - \zeta \alpha_\ell (\lambda - v_{\mathrm{g}})\right) = 0.\end{aligned} \tag{60}$$

Remark 2 (eigenvectors). The eigenvector equation for $\mathbf{A}$ is

$$\mathbf{A}\omega = \lambda\omega. \tag{61}$$

From (57) we obtain

$$\omega = \begin{bmatrix} m_{\mathrm{g}} \left(m_\ell \mu_\ell + (\lambda - v_\ell)\right) \\ \zeta m_\ell (\lambda - v_{\mathrm{g}}) - m_{\mathrm{g}} m_\ell \mu_{\mathrm{g}} \\ \lambda \left(\hat{\rho}\lambda - m_{\mathrm{g}} m_\ell (\mu_{\mathrm{g}} - \mu_\ell) - m_{\mathrm{g}} v_\ell - \zeta m_\ell v_{\mathrm{g}}\right) \end{bmatrix}. \tag{62}$$

The eigenvalue equation (60), being a third-order polynomial, can in principle be solved exactly to yield algebraic expressions for the eigenvalues λ. However, as tools for understanding the wave structure of the drift-flux model, these exact solutions are of limited value due to their high degree of complexity. In practice, one would often prefer making some simplifying assumptions and study the resulting *approximate* eigenvalues.

4.1. The Zuber–Findlay relation. A very important special case is the Zuber–Findlay slip relation [30], which can be written in the following simple analytical form:

$$v_{\mathrm{g}} = K \left(\alpha_{\mathrm{g}} v_{\mathrm{g}} + \alpha_\ell v_\ell\right) + S \tag{63}$$

or equivalently

$$\Phi = \frac{(K-1)v_{\mathrm{g}} + S}{K\alpha_\ell}. \tag{64}$$

This expression was derived from continuity considerations by Zuber and Findlay [30], where two different effects are taken into account:

1. The effect of nonuniform velocity and concentration profiles. The *shape factor* K is defined as

$$K = \frac{\langle (\alpha_{\mathrm{g}} v_{\mathrm{g}} + \alpha_\ell v_\ell)\alpha_{\mathrm{g}} \rangle}{\langle \alpha_{\mathrm{g}} v_{\mathrm{g}} + \alpha_\ell v_\ell \rangle \langle \alpha_{\mathrm{g}} \rangle}, \tag{65}$$

   where

$$\langle \cdot \rangle = \frac{1}{A} \int_A (\cdot)(x, y, z) dA. \tag{66}$$

   Here A is the cross-sectional area in the (y, z)-plane.
2. The effect of local relative velocity. The *drift velocity* S is defined as the terminal velocity of a single gas bubble rising through the liquid.

The Zuber–Findlay relation (63) has been experimentally established for a broad range of parameters for both bubbly and slug flows [3, 15].

This particular drift-flux model has been extensively studied by Théron [26] and Benzoni-Gavage [5]. By making some simplifying assumptions (most notably constant K and S as well as an incompressible liquid phase) they obtained the eigenvalues

- *sonic waves*

$$\lambda_{\mathrm{s}} = v_\ell \pm \sqrt{\frac{p}{\rho_\ell \alpha_{\mathrm{g}} \left(1 - K\alpha_{\mathrm{g}}\right)}}, \tag{67}$$

- *material wave*

$$\lambda_{\mathrm{m}} = v_{\mathrm{g}}. \tag{68}$$

Benzoni-Gavage [5] demonstrated that the sonic characteristic fields are genuinely nonlinear, whereas the material field is linearly degenerate. Provided the liquid phase is incompressible, Gavrilyuk and Fabre [16] have demonstrated that under a suitable variable transformation, the drift-flux model with slip relation (63) is mathematically similar to the Euler equations of gas dynamics.

In the following sections, we demonstrate how the drift-flux model may be analyzed more generally using a perturbation technique suggested by Toumi and coworkers [27, 28, 29]. In particular, we allow the liquid to be compressible and recover the above results of [26, 5] as the low-order limit in the perturbation parameter.

4.2. A simplifying assumption. In the following, we will assume that the slip relation can be expressed in the Zuber–Findlay form (63). Here we allow the parameters K and S to be expressed as general functions:

$$K = K(p, v_{\mathrm{g}}), \tag{69}$$

$$S = S(p, v_{\mathrm{g}}). \tag{70}$$

Equivalently, this can be expressed as a differential equation:

$$\alpha_\ell \left(\frac{\partial \Phi}{\partial \alpha_\ell}\right)_p + \Phi = 0. \tag{71}$$

From (34), (35), and (48) we may derive the following identity:

$$\left(\frac{\partial \Phi}{\partial \alpha_\ell}\right)_p \equiv \rho_\ell \mu_\ell - \rho_{\mathrm{g}} \mu_{\mathrm{g}}. \tag{72}$$

Hence from (71) we obtain

$$\mu_\ell = \frac{\rho_{\mathrm{g}}}{\rho_\ell}\mu_{\mathrm{g}} + \frac{v_\ell - v_{\mathrm{g}}}{m_\ell} \tag{73}$$

and the eigenvalue equation (60) simplifies to

$$\begin{aligned}(\lambda - v_{\mathrm{g}})(\lambda - v_\ell)(\hat{\rho}\lambda - m_{\mathrm{g}} v_{\mathrm{g}} - \zeta m_\ell v_\ell) + m_{\mathrm{g}} m_\ell \left(\mu_\ell (\lambda - v_{\mathrm{g}})^2 - \mu_{\mathrm{g}} (\lambda - v_\ell)^2\right) \\ - \kappa \rho_{\mathrm{g}} \rho_\ell \left(\alpha_{\mathrm{g}} + \zeta \alpha_\ell\right)(\lambda - v_{\mathrm{g}}) = 0.\end{aligned} \tag{74}$$

4.3. Dimensionless formulation. By making the substitution

$$\lambda = v_{\mathrm{g}} + a\sigma \tag{75}$$

we will achieve some simplification, where a now plays the role of the unknown. We may now write (74) as

$$\begin{aligned} &a\sigma(v_{\mathrm{g}} - v_\ell + a\sigma)\left(\hat{\rho}a\sigma + \zeta m_\ell(v_{\mathrm{g}} - v_\ell)\right) + m_{\mathrm{g}}m_\ell\left(\mu_\ell a^2\sigma^2 - \mu_{\mathrm{g}}(v_{\mathrm{g}} - v_\ell + a\sigma)^2\right) \\ &- \kappa\rho_{\mathrm{g}}\rho_\ell\left(\alpha_{\mathrm{g}} + \zeta\alpha_\ell\right)a\sigma = 0. \end{aligned} \tag{76}$$

Now defining σ as

$$\sigma^2 = \kappa\hat{\rho}\left(\alpha_{\mathrm{g}} + \zeta\alpha_\ell\right) \tag{77}$$

and introducing the *dimensionless* variables

$$\varepsilon = \frac{v_{\mathrm{g}} - v_\ell}{\sigma}, \tag{78}$$

$$z = \frac{m_{\mathrm{g}}\alpha_\ell}{\sigma}\mu_{\mathrm{g}}, \tag{79}$$

$$\psi = \frac{\rho_{\mathrm{g}}}{\hat{\rho}}, \tag{80}$$

$$\varphi = \frac{\rho_\ell}{\hat{\rho}}, \tag{81}$$

the eigenvalue equation (76) may correspondingly be written in dimensionless form

$$a(\varepsilon + a)(a + \zeta\alpha_\ell\varphi\varepsilon) + z\psi a^2 - z\varphi(\varepsilon + a)^2 - \alpha_{\mathrm{g}}\psi\varepsilon a^2 - \varphi\psi a = 0. \tag{82}$$

Now introducing the *pseudoliquid fraction*

$$\hat{\alpha} = \zeta\alpha_\ell \tag{83}$$

and noting that

$$\alpha_{\mathrm{g}}\psi + \zeta\alpha_\ell\varphi = 1, \tag{84}$$

the eigenvalue equation (82) simplifies to

$$a^3 + \left(2\hat{\alpha}\varphi\varepsilon - z(\varphi - \psi)\right)a^2 + \varphi\left(\hat{\alpha}\varepsilon^2 - 2z\varepsilon - \psi\right)a - z\varphi\varepsilon^2 = 0. \tag{85}$$

4.4. A power series approximation. We may now write a as a power series expansion

$$a = \sum_{i=0}^{\infty}\beta_i\chi^i \tag{86}$$

for some perturbation parameter χ. Now several choices for χ are available through (78)–(81), depending on the values of the physical variables. In the following, we will use as our starting point the *incompressible* limit and obtain eigenvalues accurate to the lowest orders of compressibility.

Towards this aim, we observe that σ given by (77) will have a magnitude in the order of the phasic sound velocities (which tend to infinity in the incompressible limit). Hence, for subsonic flows, we expect

$$\varepsilon \ll 1. \tag{87}$$

Consequently we write

$$a = \sum_{i=0}^{\infty} \beta_i \varepsilon^i \tag{88}$$

and obtain the coefficients β_i by repeatedly solving (85) to the corresponding order in ε.

4.4.1. Material wave. From (85) we obtain

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -z/\psi \\ 2z^2/\psi^2 \\ \vdots \end{bmatrix}, \tag{89}$$

which translates into the eigenvalue

$$\lambda^{\mathrm{m}} = v_{\mathrm{g}} - \frac{\alpha_{\mathrm{g}}\alpha_\ell}{\alpha_{\mathrm{g}} + \zeta\alpha_\ell}\mu_{\mathrm{g}}\frac{(v_{\mathrm{g}} - v_\ell)^2}{\kappa} + \mathcal{O}(\varepsilon^3) \tag{90}$$

by the relations of section 4.3.

4.4.2. Sonic waves. We will find it convenient to introduce the shorthand

$$w = \sqrt{z^2(\psi - \varphi)^2 + 4\varphi\psi}. \tag{91}$$

From (85) we obtain

- *downstream pressure wave*

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} \left((\varphi - \psi)z + w\right)/2 \\ 2\varphi\left(z - \hat{\alpha}\beta_0\right)/w \\ \beta_1\left(4\varphi\psi(1 - \hat{\alpha}\varphi) - z^2(\varphi^2 - \psi^2) - 2\varphi w z\right)/(2\beta_0 w^2) \\ \vdots \end{bmatrix}, \tag{92}$$

- *upstream pressure wave* (obtained from (85))

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} \left((\varphi - \psi)z - w\right)/2 \\ -2\varphi\left(z - \hat{\alpha}\beta_0\right)/w \\ \beta_1\left(4\varphi\psi(1 - \hat{\alpha}\varphi) - z^2(\varphi^2 - \psi^2) + 2\varphi w z\right)/(2\beta_0 w^2) \\ \vdots \end{bmatrix}. \tag{93}$$

Now by writing the sonic eigenvalues in the form

$$\lambda^{\mathrm{p}} = \bar{v}^{\mathrm{p}} \pm c, \tag{94}$$

the coefficients (92)–(93) yield after some manipulation

$$\bar{v}^{\mathrm{p}} = \frac{m_{\mathrm{g}}v_{\mathrm{g}} + \zeta m_\ell v_\ell}{m_{\mathrm{g}} + \zeta m_\ell} + m_{\mathrm{g}}\alpha_\ell\mu_{\mathrm{g}}\frac{\rho_\ell - \rho_{\mathrm{g}}}{2\hat{\rho}} + \frac{\alpha_{\mathrm{g}}\alpha_\ell}{\alpha_{\mathrm{g}} + \zeta\alpha_\ell}\mu_{\mathrm{g}}\frac{(v_{\mathrm{g}} - v_\ell)^2}{2\kappa} + \mathcal{O}(\varepsilon^3), \tag{95}$$

as well as the sonic velocity c:

$$c = \frac{1}{2}w\sigma + \frac{z\varphi}{w}\left(2 - \hat{\alpha}(\varphi - \psi)\right)(v_{\mathrm{g}} - v_\ell) + \mathcal{O}(\varepsilon^2). \tag{96}$$

Remark 3. Given that

$$\operatorname{trace}(\mathbf{A}) = \sum_i \lambda_i, \tag{97}$$

the following *exact* relation between $\bar{v}^{\mathrm{p}}$ and λ^{m} is satisfied:

$$2\bar{v}^{\mathrm{p}} + \lambda^{\mathrm{m}} = \frac{1}{\hat{\rho}}\left(m_{\mathrm{g}} m_\ell \mu_{\mathrm{g}}\left(1 - \frac{\rho_{\mathrm{g}}}{\rho_\ell}\right)\right) + v_{\mathrm{g}} + 2\frac{m_{\mathrm{g}} v_{\mathrm{g}} + \zeta m_\ell v_\ell}{\hat{\rho}}. \tag{98}$$

Remark 4. Although these eigenvalue expressions have been obtained under the assumption (71), similar techniques may be applied to solve (60) for other slip relations not satisfying (71). However, some knowledge of the relationship between the parameters μ_{g}, μ_ℓ, and μ_{v} will be useful for simplifying the calculations and determining a good choice of perturbation parameter.

4.5. Zuber–Findlay revisited. We now revisit the special case of the Zuber–Findlay slip relation (63)

$$v_{\mathrm{g}} = K(\alpha_{\mathrm{g}} v_{\mathrm{g}} + \alpha_\ell v_\ell) + S, \tag{99}$$

but we now consider K and S to be *constants*, which depend on the flow regime. This further simplification of (71) is often used for practical calculations [11, 15, 30].

4.5.1. Slip derivatives. By differentiation, we obtain the following explicit expressions for the slip parameters (34)–(37):

$$\mu_{\mathrm{v}} = \frac{K-1}{K\alpha_\ell}, \tag{100}$$

$$\mu_{\mathrm{g}} = (v_{\mathrm{g}} - v_\ell)\kappa\frac{\partial \rho_\ell}{\partial p}, \tag{101}$$

$$\mu_\ell = -(v_{\mathrm{g}} - v_\ell)\kappa\frac{\alpha_{\mathrm{g}}}{\alpha_\ell}\frac{\partial \rho_{\mathrm{g}}}{\partial p}, \tag{102}$$

and

$$\zeta = \frac{1 - K\alpha_{\mathrm{g}}}{K\alpha_\ell}. \tag{103}$$

Asymptotic expressions for the eigenvalues could now be obtained by substituting (100)–(103) into the previously calculated expressions (90) and (95)–(96). Equivalently, we may also substitute (100)–(103) into (76) and repeat the power series analysis. This will greatly simplify the calculations, a point that will be demonstrated in the following.

4.5.2. Eigenvalue equation. From (101) we note that z, as defined by (79), may be written as

$$z = \eta\varepsilon, \tag{104}$$

where

$$\eta = m_{\mathrm{g}}\alpha_\ell\kappa\frac{\partial \rho_\ell}{\partial p}. \tag{105}$$

Substituting in (85) we obtain the eigenvalue equation

$$a^3 + \varepsilon\left(2\hat{\alpha}\varphi - \eta(\varphi - \psi)\right)a^2 + \varphi\left(\hat{\alpha}\varepsilon^2 - 2\eta\varepsilon^2 - \psi\right)a - \eta\varphi\varepsilon^3 = 0. \tag{106}$$

4.5.3. Material wave. Solving (106) to powers of ε yields the following coefficients:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \vdots \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -\eta/\psi \\ 0 \\ \eta(2\eta - \hat{\alpha})/\psi^2 \\ \vdots \end{bmatrix}. \tag{107}$$

Hence

$$\lambda_{\mathrm{m}} = v_{\mathrm{g}} - K\alpha_{\mathrm{g}}\alpha_\ell \frac{\partial \rho_\ell}{\partial p}(v_{\mathrm{g}} - v_\ell)^3 + \mathcal{O}(\varepsilon^5). \tag{108}$$

Changes in the material composition are consequently propagated by the velocity of the gas bubbles, plus small correction terms representing volumetric changes due to compression. Note that $\lambda_{\mathrm{m}} = v_{\mathrm{g}}$ becomes an *exact* eigenvalue for $\eta = 0$, the limit of incompressible liquid [16].

4.5.4. Sonic waves. For the sonic waves, we obtain the following coefficients:

- *Downstream pressure wave*

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} \sqrt{\varphi\psi} \\ \eta(\phi - \psi)/2 - \hat{\alpha}\varphi \\ \left(2\varphi(4\eta + 2\hat{\alpha}\eta\psi - \eta^2\psi - 4\hat{\alpha}) + (2\hat{\alpha} - \eta)^2\varphi^2 + \eta^2\psi^2\right)/(8\sqrt{\varphi\psi}) \\ \vdots \end{bmatrix}, \tag{109}$$

- *Upstream pressure wave*

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} -\sqrt{\varphi\psi} \\ \eta(\phi - \psi)/2 - \hat{\alpha}\varphi \\ -\left(2\varphi(4\eta + 2\hat{\alpha}\eta\psi - \eta^2\psi - 4\hat{\alpha}) + (2\hat{\alpha} - \eta)^2\varphi^2 + \eta^2\psi^2\right)/(8\sqrt{\varphi\psi}) \\ \vdots \end{bmatrix}. \tag{110}$$

Hence

$$\bar{v}^{\mathrm{p}} = \frac{m_{\mathrm{g}}v_{\mathrm{g}} + \zeta m_\ell v_\ell}{m_{\mathrm{g}} + \zeta m_\ell} + m_{\mathrm{g}}\alpha_\ell\kappa\frac{\partial \rho_\ell}{\partial p}(v_{\mathrm{g}} - v_\ell)\frac{\rho_\ell - \rho_{\mathrm{g}}}{2\hat{\rho}} + \mathcal{O}(\varepsilon^3), \tag{111}$$

and the sound velocity c may be written as

$$c = \sqrt{\frac{\kappa\rho_{\mathrm{g}}\rho_\ell}{K\alpha_{\mathrm{g}}(\rho_{\mathrm{g}} - \rho_\ell) + \rho_\ell}} + \mathcal{O}(\varepsilon)^2. \tag{112}$$

Remark 5. Note that $\rho_{\mathrm{g}} \ll \rho_\ell$ implies $c \ll \sigma$, and the requirement $\varepsilon \ll 1$ (87) has a significantly broader range of validity than the assumption of subsonic slip, $|v_{\mathrm{g}} - v_\ell| \ll c$.

Remark 6. The sonic eigenvalues may be written as

$$\lambda_{\mathrm{s}} = v_\ell \pm \sqrt{\frac{(\partial p/\partial \rho_{\mathrm{g}})\rho_{\mathrm{g}}}{(1 - K\alpha_{\mathrm{g}})\rho_\ell\alpha_{\mathrm{g}}}} + \mathcal{O}(\eta) + \mathcal{O}(\psi) + \mathcal{O}(\varepsilon^2), \tag{113}$$

which reduces to the result (67) when $p(\rho_{\mathrm{g}})$ satisfies the ideal gas law.

5. Two-fluid formulation. In this section, we perform the transformation required to write the *general* drift-flux model of section 3.1 in *canonical two-fluid form* as described in section 2.3. In other words, we replace the conservation equation (19), together with the slip relation (20), with equivalent evolution equations for the momentums of each phase.

5.1. Momentum evolution equations. We first derive an explicit gas momentum evolution equation for the general drift-flux model with slip relation (20). Our starting point is the previously derived differential (45), which becomes

$$\frac{\partial I_{\mathrm{g}}}{\partial t} = \frac{1}{\hat{\rho}} \left(m_{\mathrm{g}} \frac{\partial I}{\partial t} + (m_{\mathrm{g}} m_\ell \mu_{\mathrm{g}} + \zeta m_\ell v_{\mathrm{g}}) \frac{\partial m_{\mathrm{g}}}{\partial t} + (m_{\mathrm{g}} m_\ell \mu_\ell - m_{\mathrm{g}} v_\ell) \frac{\partial m_\ell}{\partial t} \right), \tag{114}$$

when written as a partial derivative with respect to t.

By using the conservation equations (27)–(29), we obtain the gas momentum evolution equation, written in terms of spatial derivatives

$$\begin{aligned} &\frac{\partial I_{\mathrm{g}}}{\partial t} + \frac{m_{\mathrm{g}}}{\hat{\rho}} \frac{\partial}{\partial x} (I_{\mathrm{g}} v_{\mathrm{g}} + I_\ell v_\ell + p) + \frac{\zeta m_\ell}{\hat{\rho}} v_{\mathrm{g}} \frac{\partial I_{\mathrm{g}}}{\partial x} \\ &- \frac{m_{\mathrm{g}}}{\hat{\rho}} v_\ell \frac{\partial I_\ell}{\partial x} + \frac{m_{\mathrm{g}} m_\ell}{\hat{\rho}} \left(\mu_{\mathrm{g}} \frac{\partial I_{\mathrm{g}}}{\partial x} + \mu_\ell \frac{\partial I_\ell}{\partial x} \right) = \frac{m_{\mathrm{g}}}{\hat{\rho}} Q. \end{aligned} \tag{115}$$

Further manipulation of derivatives yields

$$\begin{aligned} &\frac{\partial I_{\mathrm{g}}}{\partial t} + \frac{\partial}{\partial x} (I_{\mathrm{g}} v_{\mathrm{g}}) + \frac{m_{\mathrm{g}}}{\hat{\rho}} \frac{\partial p}{\partial x} \\ &+ \frac{m_{\mathrm{g}} m_\ell}{\hat{\rho}} \left(v_\ell \frac{\partial v_\ell}{\partial x} - \zeta v_{\mathrm{g}} \frac{\partial v_{\mathrm{g}}}{\partial x} + \mu_{\mathrm{g}} \frac{\partial I_{\mathrm{g}}}{\partial x} + \mu_\ell \frac{\partial I_\ell}{\partial x} \right) = \frac{m_{\mathrm{g}}}{\hat{\rho}} Q. \end{aligned} \tag{116}$$

5.1.1. Canonical form. Writing (116) under the canonical two-fluid form of section 2.3,

$$\frac{\partial I_{\mathrm{g}}}{\partial t} + \frac{\partial}{\partial x} (I_{\mathrm{g}} v_{\mathrm{g}}) + \alpha_{\mathrm{g}} \frac{\partial p}{\partial x} + \tau_{\mathrm{i}} = Q_{\mathrm{g}}, \tag{117}$$

where $Q = Q_{\mathrm{g}} + Q_\ell$, we find that the interface friction τ_{i} is given by

$$\begin{aligned} \tau_{\mathrm{i}} = {} & \alpha_{\mathrm{g}} \alpha_\ell \frac{\rho_{\mathrm{g}} - \zeta \rho_\ell}{\hat{\rho}} \frac{\partial p}{\partial x} + \frac{\zeta m_\ell}{\hat{\rho}} Q_{\mathrm{g}} - \frac{m_{\mathrm{g}}}{\hat{\rho}} Q_\ell \\ &+ \frac{m_{\mathrm{g}} m_\ell}{\hat{\rho}} \left(v_\ell \frac{\partial v_\ell}{\partial x} - \zeta v_{\mathrm{g}} \frac{\partial v_{\mathrm{g}}}{\partial x} + \mu_{\mathrm{g}} \frac{\partial I_{\mathrm{g}}}{\partial x} + \mu_\ell \frac{\partial I_\ell}{\partial x} \right). \end{aligned} \tag{118}$$

5.1.2. Liquid momentum evolution. By inserting (118) into the canonical liquid momentum equation (17), we obtain

$$\begin{aligned} &\frac{\partial I_\ell}{\partial t} + \frac{\partial}{\partial x} (I_\ell v_\ell) + \frac{\zeta m_\ell}{\hat{\rho}} \frac{\partial p}{\partial x} \\ &- \frac{m_{\mathrm{g}} m_\ell}{\hat{\rho}} \left(v_\ell \frac{\partial v_\ell}{\partial x} - \zeta v_{\mathrm{g}} \frac{\partial v_{\mathrm{g}}}{\partial x} + \mu_{\mathrm{g}} \frac{\partial I_{\mathrm{g}}}{\partial x} + \mu_\ell \frac{\partial I_\ell}{\partial x} \right) = \frac{\zeta m_\ell}{\hat{\rho}} Q. \end{aligned} \tag{119}$$

5.2. Quasi-linear formulation. We may now express this rewritten drift-flux model in quasi-linear form:

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{U}}{\partial x} = \mathbf{Q}, \tag{120}$$

similar to section 3.2. However, the matrix $\mathbf{A}$ is now 4×4 and $\mathbf{U}$ is given by

$$\mathbf{U} = \begin{bmatrix} \rho_{\mathrm{g}}\alpha_{\mathrm{g}} \\ \rho_{\ell}\alpha_{\ell} \\ \rho_{\mathrm{g}}\alpha_{\mathrm{g}}v_{\mathrm{g}} \\ \rho_{\ell}\alpha_{\ell}v_{\ell} \end{bmatrix}, \tag{121}$$

whereas the vector of sources is

$$\mathbf{Q} = \frac{1}{\hat{\rho}} \begin{bmatrix} 0 \\ 0 \\ m_{\mathrm{g}}Q \\ \zeta m_{\ell}Q \end{bmatrix}. \tag{122}$$

We now split (118) into four parts:

$$\tau_{\mathrm{i}} = \tau_p + \tau_v + \tau_{\alpha} + \tau_Q, \tag{123}$$

where

$$\tau_p = \alpha_{\mathrm{g}}\alpha_{\ell}\frac{\rho_{\mathrm{g}} - \zeta\rho_{\ell}}{\hat{\rho}}\frac{\partial p}{\partial x}, \tag{124}$$

$$\tau_v = \frac{m_{\mathrm{g}}m_{\ell}}{\hat{\rho}}\left(v_{\ell}\frac{\partial v_{\ell}}{\partial x} - \zeta v_{\mathrm{g}}\frac{\partial v_{\mathrm{g}}}{\partial x}\right), \tag{125}$$

$$\tau_{\alpha} = \frac{m_{\mathrm{g}}m_{\ell}}{\hat{\rho}}\left(\mu_{\mathrm{g}}\frac{\partial I_{\mathrm{g}}}{\partial x} + \mu_{\ell}\frac{\partial I_{\ell}}{\partial x}\right), \tag{126}$$

and

$$\tau_Q = \frac{\zeta m_{\ell}}{\hat{\rho}}Q_{\mathrm{g}} - \frac{m_{\mathrm{g}}}{\hat{\rho}}Q_{\ell}. \tag{127}$$

This defines a natural decomposition of the Jacobi matrix as follows:

$$\mathbf{A}(\mathbf{U}) = \mathbf{A}_0 + \mathbf{A}_p + \mathbf{A}_v + \mathbf{A}_{\alpha}, \tag{128}$$

i.e., one additional contribution for each differential term of the interface friction.

5.2.1. $\mathbf{A_0}$. The Jacobi matrix for the canonical two-fluid model with $\tau_{\mathrm{i}} = 0$ is [12]

$$\mathbf{A}_0 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \kappa\rho_{\ell}\alpha_{\mathrm{g}} - v_{\mathrm{g}}^2 & \kappa\rho_{\mathrm{g}}\alpha_{\mathrm{g}} & 2v_{\mathrm{g}} & 0 \\ \kappa\rho_{\ell}\alpha_{\ell} & \kappa\rho_{\mathrm{g}}\alpha_{\ell} - v_{\ell}^2 & 0 & 2v_{\ell} \end{bmatrix}. \tag{129}$$

5.2.2. $\mathbf{A}_{\boldsymbol{p}}$. From (49) we obtain

$$\mathbf{A}_p(\mathbf{U}) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \kappa\rho_\ell\alpha_{\mathrm{g}}\alpha_\ell\frac{\rho_{\mathrm{g}}-\zeta\rho_\ell}{\hat{\rho}} & \kappa\rho_{\mathrm{g}}\alpha_{\mathrm{g}}\alpha_\ell\frac{\rho_{\mathrm{g}}-\zeta\rho_\ell}{\hat{\rho}} & 0 & 0 \\ -\kappa\rho_\ell\alpha_{\mathrm{g}}\alpha_\ell\frac{\rho_{\mathrm{g}}-\zeta\rho_\ell}{\hat{\rho}} & -\kappa\rho_\ell\alpha_{\mathrm{g}}\alpha_\ell\frac{\rho_{\mathrm{g}}-\zeta\rho_\ell}{\hat{\rho}} & 0 & 0 \end{bmatrix}. \tag{130}$$

5.2.3. $\mathbf{A}_{\boldsymbol{v}}$. From

$$dv_{\mathrm{g}} = \frac{1}{m_{\mathrm{g}}}\,dI_{\mathrm{g}} - \frac{v_{\mathrm{g}}}{m_{\mathrm{g}}}\,dm_{\mathrm{g}} \tag{131}$$

and

$$dv_\ell = \frac{1}{m_\ell}\,dI_\ell - \frac{v_\ell}{m_\ell}\,dm_\ell \tag{132}$$

we obtain

$$\mathbf{A}_v(\mathbf{U}) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{\zeta m_\ell}{\hat{\rho}}v_{\mathrm{g}}^2 & -\frac{m_{\mathrm{g}}}{\hat{\rho}}v_\ell^2 & -\frac{\zeta m_\ell}{\hat{\rho}}v_{\mathrm{g}} & \frac{m_{\mathrm{g}}}{\hat{\rho}}v_\ell \\ -\frac{\zeta m_\ell}{\hat{\rho}}v_{\mathrm{g}}^2 & \frac{m_{\mathrm{g}}}{\hat{\rho}}v_\ell^2 & \frac{\zeta m_\ell}{\hat{\rho}}v_{\mathrm{g}} & -\frac{m_{\mathrm{g}}}{\hat{\rho}}v_\ell \end{bmatrix}. \tag{133}$$

5.2.4. $\mathbf{A}_{\boldsymbol{\alpha}}$. We directly obtain

$$\mathbf{A}_\alpha(\mathbf{U}) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{m_{\mathrm{g}}m_\ell}{\hat{\rho}}\mu_{\mathrm{g}} & \frac{m_{\mathrm{g}}m_\ell}{\hat{\rho}}\mu_\ell \\ 0 & 0 & -\frac{m_{\mathrm{g}}m_\ell}{\hat{\rho}}\mu_{\mathrm{g}} & -\frac{m_{\mathrm{g}}m_\ell}{\hat{\rho}}\mu_\ell \end{bmatrix}. \tag{134}$$

5.2.5. End result. Adding all contributions we obtain from (128)

(135)

$$\mathbf{A}(\mathbf{U}) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \frac{m_{\mathrm{g}}}{\hat{\rho}}\left(\kappa\rho_\ell - v_{\mathrm{g}}^2\right) & \frac{m_{\mathrm{g}}}{\hat{\rho}}\left(\kappa\rho_{\mathrm{g}} - v_\ell^2\right) & \left(2-\frac{\zeta m_\ell}{\hat{\rho}}\right)v_{\mathrm{g}} + \frac{m_{\mathrm{g}}m_\ell}{\hat{\rho}}\mu_{\mathrm{g}} & \frac{m_{\mathrm{g}}}{\hat{\rho}}v_\ell + \frac{m_{\mathrm{g}}m_\ell}{\hat{\rho}}\mu_\ell \\ \frac{\zeta m_\ell}{\hat{\rho}}\left(\kappa\rho_\ell - v_{\mathrm{g}}^2\right) & \frac{\zeta m_\ell}{\hat{\rho}}\left(\kappa\rho_{\mathrm{g}} - v_\ell^2\right) & \frac{\zeta m_\ell}{\hat{\rho}}v_{\mathrm{g}} - \frac{m_{\mathrm{g}}m_\ell}{\hat{\rho}}\mu_{\mathrm{g}} & \left(2-\frac{m_{\mathrm{g}}}{\hat{\rho}}\right)v_\ell - \frac{m_{\mathrm{g}}m_\ell}{\hat{\rho}}\mu_\ell \end{bmatrix}.$$

5.2.6. Eigenvalues. The eigenvalues of the matrix $\mathbf{A}$ are the roots of the polynomial equation

$$\begin{aligned} &\lambda\Big[(\lambda - v_{\mathrm{g}})(\lambda - v_\ell)(\hat{\rho}\lambda - m_{\mathrm{g}}v_{\mathrm{g}} - \zeta m_\ell v_\ell) \\ &\quad + m_{\mathrm{g}}m_\ell\left(\mu_\ell(\lambda - v_{\mathrm{g}})^2 - \mu_{\mathrm{g}}(\lambda - v_\ell)^2\right) \\ &\quad + \kappa\rho_{\mathrm{g}}\rho_\ell\left(\alpha_{\mathrm{g}}\alpha_\ell(\rho_{\mathrm{g}}\mu_{\mathrm{g}} - \rho_\ell\mu_\ell) - \alpha_{\mathrm{g}}(\lambda - v_\ell) - \zeta\alpha_\ell(\lambda - v_{\mathrm{g}})\right)\Big] = 0. \end{aligned} \tag{136}$$

By direct comparison with (60), we see that this may be written as

$$\lambda P(\lambda) = 0, \tag{137}$$

where $P(\lambda)$ is the eigenvalue polynomial for the original drift-flux model.

Remark 7. We have written the drift-flux model as a quasi-linear system of four equations by deriving two momentum equations which replace the mixed momentum equation and the slip law. As a consequence, the characteristic speeds of this system are given by (137) showing that a new characteristic speed $\lambda = 0$, representing the slip relation, has been added to the characteristic speeds already given by the drift-flux model.

This situation is similar to what is observed for a much simpler problem. Consider the scalar equation

$$u_t + f(u)_x = k'(x)g(u), \tag{138}$$

where k, f, and g are given functions. A common approach for deriving numerical schemes for the model (138) is to first write the model as a quasi-linear system of two equations, by adding the trivial equation $k_t = 0$, which gives

$$U_t + A(U)U_x = 0, \qquad U = \begin{pmatrix} u \\ k \end{pmatrix}, \qquad A(U) = \begin{pmatrix} f'(u) & -g(u) \\ 0 & 0 \end{pmatrix}. \tag{139}$$

The characteristic speeds of this system are given by $\lambda_1 = f'(u)$ and $\lambda_2 = 0$. If $f'(u) = 0$ for some u, then the eigenvalues coincide, and we have so-called resonance; see, for instance, [17] and the references therein for more on this. Note that this phenomenon might well also occur for our system (120)–(126), since one of the solutions of $P(\lambda) = 0$ corresponding to the slow material wave (see below for more details) can be zero. This happens when $v_{\mathrm{g}} = v_\ell = 0$.

It is interesting to note that the form (139) often is used as the starting point for designing numerical schemes for solving (138). In a similar manner we could imagine to use the above two-fluid form (120)–(126) as a starting point for developing a numerical scheme for the drift-flux model, e.g., by using the numerical schemes more recently proposed in [13, 14] for the two-fluid model.

6. Interface friction and wave velocities. In this section, we investigate how the wave structure of the two-fluid model gradually changes as it is transformed into a drift-flux model by addition of the various terms of (123). Our starting point is the *canonical model* with $\tau_i = 0$:

$$\frac{\partial}{\partial t}\left(\rho_{\mathrm{g}}\alpha_{\mathrm{g}}\right) + \frac{\partial}{\partial x}\left(\rho_{\mathrm{g}}\alpha_{\mathrm{g}}v_{\mathrm{g}}\right) = 0, \tag{140}$$

$$\frac{\partial}{\partial t}\left(\rho_\ell\alpha_\ell\right) + \frac{\partial}{\partial x}\left(\rho_\ell\alpha_\ell v_\ell\right) = 0, \tag{141}$$

$$\frac{\partial}{\partial t}\left(\rho_{\mathrm{g}}\alpha_{\mathrm{g}}v_{\mathrm{g}}\right) + \frac{\partial}{\partial x}\left(\rho_{\mathrm{g}}\alpha_{\mathrm{g}}v_{\mathrm{g}}^2\right) + \alpha_{\mathrm{g}}\frac{\partial}{\partial x}\left(p\right) = Q_{\mathrm{g}}, \tag{142}$$

$$\frac{\partial}{\partial t}\left(\rho_\ell\alpha_\ell v_\ell\right) + \frac{\partial}{\partial x}\left(\rho_\ell\alpha_\ell v_\ell^2\right) + \alpha_\ell\frac{\partial}{\partial x}\left(p\right) = Q_\ell. \tag{143}$$

6.1. Wave structure of the canonical model. For different choices of τ_i, Toumi and coworkers [27, 28, 29] investigated the wave structure of the model with a perturbation technique. For $\tau_i = 0$, the wave velocities are precisely the eigenvalues of the matrix $\mathbf{A}_0$ given by (129). Now defining

$$\varepsilon = \frac{v_{\mathrm{g}} - v_\ell}{\hat{c}}, \tag{144}$$

where $\hat{c}$ is a *mixture* sonic velocity given by

$$\hat{c} = \sqrt{\frac{\rho_\ell \alpha_{\mathrm{g}} + \rho_{\mathrm{g}} \alpha_\ell}{(\partial \rho_{\mathrm{g}}/\partial p)\rho_\ell \alpha_{\mathrm{g}} + (\partial \rho_\ell/\partial p)\rho_{\mathrm{g}} \alpha_\ell}} = \sqrt{(\rho_\ell \alpha_{\mathrm{g}} + \rho_{\mathrm{g}} \alpha_\ell)\kappa}, \tag{145}$$

approximate eigenvalues for (140)–(143) were presented by Evje and Flåtten [12] as described below.

6.1.1. Material waves. Writing

$$\lambda_{\mathrm{m}} = \bar{v}^{\mathrm{v}} \pm \gamma, \tag{146}$$

we obtain

$$\bar{v}^{\mathrm{v}} = \frac{\rho_{\mathrm{g}} \alpha_\ell v_{\mathrm{g}} + \rho_\ell \alpha_{\mathrm{g}} v_\ell}{\rho_{\mathrm{g}} \alpha_\ell + \rho_\ell \alpha_{\mathrm{g}}} + \mathcal{O}(\varepsilon^3) \tag{147}$$

and

$$\gamma = i \frac{\sqrt{\rho_{\mathrm{g}} \rho_\ell \alpha_{\mathrm{g}} \alpha_\ell}(v_{\mathrm{g}} - v_\ell)}{\rho_{\mathrm{g}} \alpha_\ell + \rho_\ell \alpha_{\mathrm{g}}} + \mathcal{O}(\varepsilon^3). \tag{148}$$

Remark 8. Note that unless $v_{\mathrm{g}} = v_\ell$, γ is imaginary and the canonical two-fluid model with $\tau_i = 0$ loses hyperbolicity. Hence the inclusion of a differential interface friction τ_i is essential for obtaining a well-behaved mathematical solution.

Remark 9. Note that if $\rho_{\mathrm{g}} \ll \rho_\ell$, $\bar{v}^{\mathrm{v}} \approx v_\ell$ and the material waves travel with the velocity of the liquid phase. This is quite the opposite of the drift-flux model, where the velocity of the material wave corresponds to the gas velocity v_{g} (section 4.4).

6.1.2. Sonic waves. Writing

$$\lambda_{\mathrm{s}} = \bar{v}^{\mathrm{p}} \pm c, \tag{149}$$

we obtain

$$\bar{v}^{\mathrm{p}} = \frac{\rho_{\mathrm{g}} \alpha_\ell v_\ell + \rho_\ell \alpha_{\mathrm{g}} v_{\mathrm{g}}}{\rho_{\mathrm{g}} \alpha_\ell + \rho_\ell \alpha_{\mathrm{g}}} + \mathcal{O}(\varepsilon^3) \tag{150}$$

and

$$c = \hat{c}\left(1 + \mathcal{O}(\varepsilon^2)\right). \tag{151}$$

Remark 10. If $\rho_{\mathrm{g}} \ll \rho_\ell$, $\bar{v}^{\mathrm{p}} \approx v_{\mathrm{g}}$ and the part of the sonic wave that is transported along the flow travels with the velocity of the gas phase. Again this contrasts the drift-flux model, where the corresponding result of section 4.4 yields $\bar{v}^{\mathrm{p}} \approx v_\ell$.

6.2. Numerical investigations. In the framework of the canonical two-fluid model, the eigenvalues of the previous section correspond to $\tau_i = 0$, whereas the eigenvalues of section 4.4 correspond to the interface friction (123),

$$\tau_{\mathrm{i}} = \tau_p + \tau_v + \tau_\alpha + \tau_Q, \tag{152}$$

which was derived in section 5.1.1. We now study the relation between the interface friction and the wave velocities more closely, by looking at a specific example. More precisely, we consider a two-phase flow satisfying the Zuber–Findlay slip relation (63) with phasic properties roughly representing an air-water mixture.

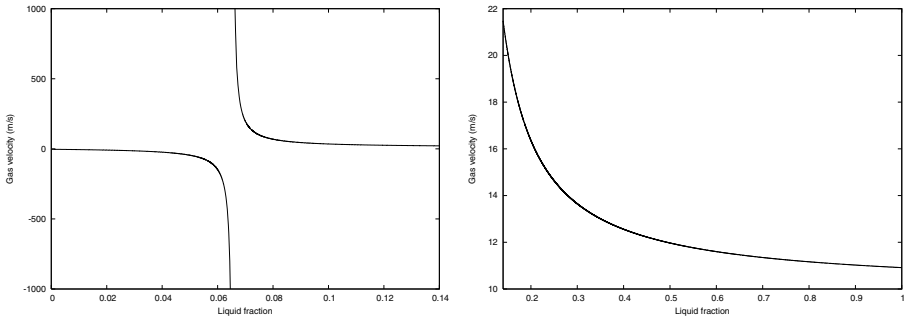


FIG. 1. *The Zuber–Findlay gas velocity as a function of liquid fraction. Left: Near singularity. Right: Physical region.*

6.2.1. Model parameters. In the following, we assume that the phasic velocities are related by the Zuber–Findlay slip relation

$$v_{\mathrm{g}} = K\left(\alpha_{\mathrm{g}} v_{\mathrm{g}} + \alpha_\ell v_\ell\right) + S, \tag{153}$$

where we choose

$$K = 1.07 \tag{154}$$

and

$$S = 0.216 \text{ m/s}. \tag{155}$$

Furthermore, we assume the flow conditions

$$v_\ell = 10 \text{ m/s}, \tag{156}$$

$$\rho_{\mathrm{g}} = 1.0 \text{ kg/m}^3, \tag{157}$$

$$\rho_\ell = 1000 \text{ kg/m}^3, \tag{158}$$

$$\partial p/\partial \rho_{\mathrm{g}} = 10^5 \text{ m}^2/\text{s}^2, \tag{159}$$

$$\partial p/\partial \rho_\ell = 10^6 \text{ m}^2/\text{s}^2. \tag{160}$$

6.2.2. Gas velocity. By inspecting the slip expression (153) we find there is a singularity corresponding to

$$\hat{\alpha} = \zeta \alpha_\ell = \frac{1 - K\alpha_{\mathrm{g}}}{K} = 0, \tag{161}$$

which with our choice of parameters occurs at

$$\alpha_\ell^{\mathrm{crit}} \approx 0.0654. \tag{162}$$

The gas velocity v_{g} changes sign from $-\infty$ to $+\infty$ across the singularity, as shown in Figure 1. However, $\alpha_\ell < \alpha_\ell^{\mathrm{crit}}$ implies large gas bubbles corresponding more or less to the *annular* flow regime, where the drift-flux model is not applicable [16]. Hence we discard the corresponding results as unphysical and base our further investigations on the assumption $\alpha_\ell > \alpha_\ell^{\mathrm{crit}}$.

6.2.3. Wave velocities. We now investigate the effect of the different terms of

$$\tau_{\mathrm{i}} = \tau_p + \tau_v + \tau_\alpha + \tau_Q \tag{163}$$

on the wave velocities of the canonical two-fluid model. Note that τ_Q, as given by (124), is purely nondifferential, and hence has no effect on the wave structure of the model. In the following plots we use the labels

- *two-fluid:* $\tau_i = 0$,
- *drift-flux:* $\tau_i = \tau_p + \tau_v + \tau_\alpha$

to denote the special choices of interface friction yielding the basic two-fluid and drift-flux wave structures, respectively, as described in sections 4 and 6.1.

Remark 11. Note that with our choice of slip relation (153), the expression (126) may by use of (101) and (102) be rewritten as

$$\tau_\alpha = (v_{\mathrm{g}} - v_\ell)\frac{m_{\mathrm{g}} m_\ell}{\hat{\rho}\alpha_\ell}\frac{\partial \alpha_\ell}{\partial t}. \tag{164}$$

In the following, wave velocities corresponding to different choices of τ_i (163) are calculated as the eigenvalues of the corresponding matrix $\mathbf{A}(\mathbf{U})$ as described in section 5.2. A numerical algorithm was used to calculate the eigenvalues, sorted in ascending order by their real parts as

$$\mathrm{Re}(\lambda_1) < \mathrm{Re}(\lambda_2) < \mathrm{Re}(\lambda_3) < \mathrm{Re}(\lambda_4). \tag{165}$$

Here λ_1 and λ_4 are sonic waves, whereas λ_2 and λ_3 represent slow waves.

6.2.4. Slip wave. As noted in Remark 7, the slip relation manifests itself as a stationary wave for the *drift-flux* interface friction ($\tau_i = \tau_p + \tau_v + \tau_\alpha$). Hence one of the two material waves described in section 6.1.1, corresponding to $\tau_i = 0$, will gradually transform into this stationary "slip wave" as the terms (124)–(126) are added to the interface friction.

The effect of this is illustrated in Figure 2, where $|\lambda_2|$ is plotted as a function of liquid fraction. Already for $\tau_i = \tau_p + \tau_v$, we obtain λ_2=0, which is left unchanged by the addition of τ_α. Note that $\tau_\alpha = 0$ corresponds to a special case of the drift-flux model, where the slip relation satisfies $\mu_{\mathrm{g}} = \mu_\ell = 0$. Hence the "drift-flux" character of the system ($\lambda_2 \equiv 0$) is fully manifest in the τ_p and τ_v components of the interface friction.

6.2.5. Material wave. As seen by the analyses of section 4.5 and 6.1, one material wave is gradually transformed from (146) ($\lambda_{\mathrm{m}} \approx v_\ell$) into (108) ($\lambda_{\mathrm{m}} \approx v_{\mathrm{g}}$).

This is illustrated in Figure 3, where $\mathrm{Re}(\lambda_3)$ is plotted as a function of liquid fraction. Note that without the inclusion of τ_α, the wave velocity is constant. This demonstrates the fact that $\tau_\alpha = 0$ implies that the slip is independent of volume fraction.

6.2.6. Sound velocity. Following sections 4.4.2 and 6.1.2, we write the sonic waves as a combination of two components as follows:

$$\lambda_{\mathrm{s}} = \bar{v}^{\mathrm{p}} \pm c, \tag{166}$$

where $\bar{v}^{\mathrm{p}}$ represents the part of the sonic wave that is transported with the flow, whereas c is the sonic velocity with respect to $\bar{v}^{\mathrm{p}}$. Hence we get

$$c = \frac{\lambda_4 - \lambda_1}{2} \tag{167}$$

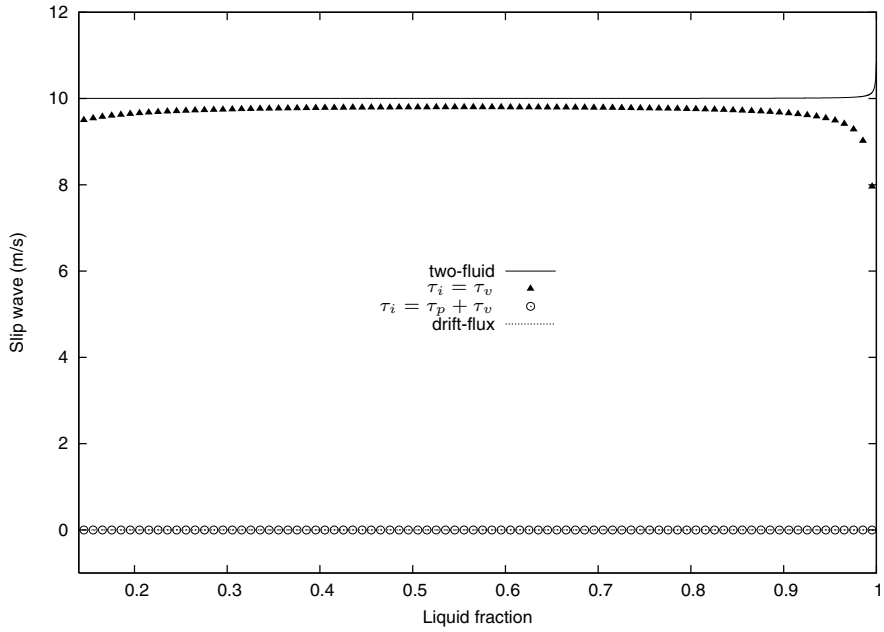


FIG. 2. *Slip wave velocity as a function of liquid fraction.*

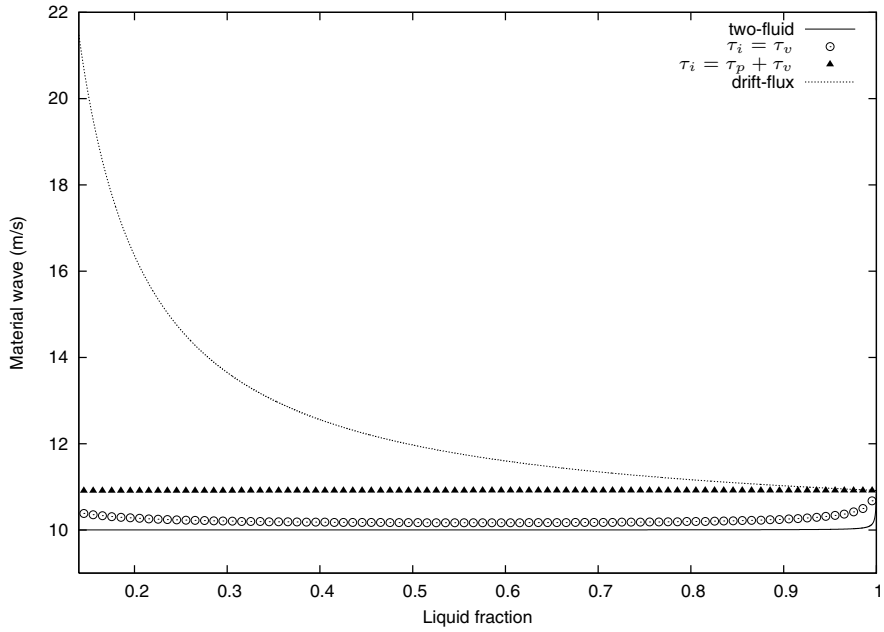


FIG. 3. *Material wave velocity as a function of liquid fraction.*

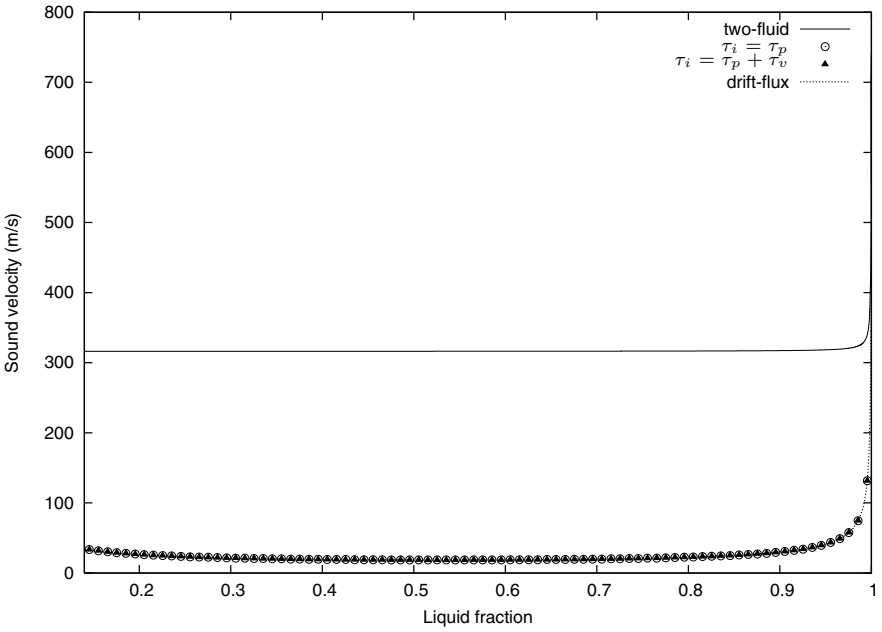


FIG. 4. *Mixture sound velocity as a function of liquid fraction.*

and

$$\bar{v}^{\mathrm{p}} = \frac{\lambda_1 + \lambda_4}{2}. \tag{168}$$

In Figure 4, the sound velocity c is plotted as a function of liquid fraction. We observe that c is transformed from the two-fluid sound velocity (145) into the drift-flux sound velocity (112) by the action of τ_p alone; the terms τ_v and τ_α have no additional effect.

Remark 12. This plot illustrates the fact that whereas for the two-fluid model

$$c_{\mathrm{tf}} \approx c_{\mathrm{g}}, \tag{169}$$

the drift-flux sonic velocity satisfies

$$c_{\mathrm{df}} \ll \min(c_{\mathrm{g}}, c_\ell). \tag{170}$$

A similar parabolic-like shape for $c_{\mathrm{mix}}(\alpha_\ell)$ was also derived by Nguyen, Winther, and Greiner [21] by considering the interface as an elastic wall. They also pointed out that this shape is consistent with experimental data for *mixed* flows.

6.2.7. Sonic transport velocity. The sonic transport velocity $\bar{v}^{\mathrm{p}}$ is plotted in Figure 5. We get more or less the inverse of Figure 3; now $\bar{v}^{\mathrm{p}} \approx v_{\mathrm{g}}$ (two-fluid model) is transformed into $\bar{v}^{\mathrm{p}} \approx v_\ell$ (drift-flux model) by the action of the interface friction (118).

7. Summary. A quasi-linear form of the *drift-flux* two-phase flow model has been derived. The wave structure of this model has been investigated by a perturbation technique, extending previous results of Théron [26] and Benzoni-Gavage [5].

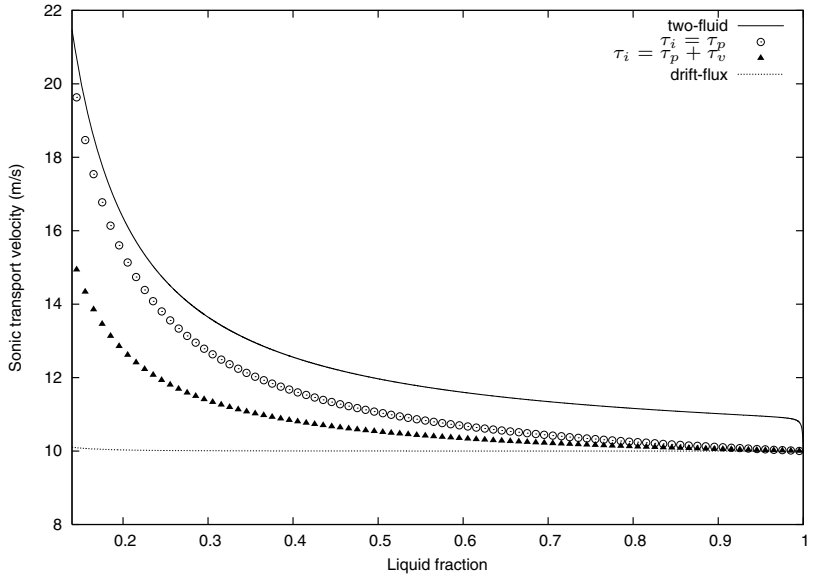


FIG. 5. *Sonic transport velocity* $\bar{v}^{\mathrm{p}}$ *as a function of liquid fraction.*

The drift-flux model has further been rewritten within the framework of a more general *two-fluid* model, by derivation of the proper form of the interface friction τ_i. Here the slip relation is represented as a stationary wave.

The interface friction τ_i may be split into four parts

$$\tau_i = \tau_p + \tau_v + \tau_\alpha + \tau_Q, \tag{171}$$

where the following hold:

- The terms τ_p and τ_v make up the *drift-flux* nature of the system (stationary slip wave).
- The term τ_p is almost exclusively associated with the mixture sound velocity c.
- The term τ_α is associated with the slow waves, imposing a dependency of volume fraction on the material wave.

The drift-flux and two-fluid formulations are often considered to be different modeling strategies with different domains of applicability. The unification presented in this paper may facilitate the implementation of both models within a single computer code. Furthermore, the link presented between the *observable* slip velocity and the underlying interface friction may serve as an aid for developing better physical models for two-phase flows.

Acknowledgments. We thank the editor and reviewers for their comments, which helped us improve the first version of this paper.

REFERENCES

[1] N. ANDRITSOS AND T. HANRATTY, *Influence of interfacial waves in stratified gas-liquid flows*, AIChE J., 33 (1987), pp. 444–454.

[2] D. BARNEA AND Y. TAITEL, *Kelvin-Helmholtz stability criteria for stratified flow: Viscous versus non-viscous (inviscid) approaches*, Int. J. Multiphase Flow, 19 (1993), pp. 639–649.

[3] K. H. Bendiksen, *An experimental investigation of the motion of long bubbles in inclined tubes*, Int. J. Multiphase Flow, 10 (1984), pp. 467–483.
[4] K. H. Bendiksen, D. Malnes, R. Moe, and S. Nuland, *The dynamic two-fluid model OLGA: Theory and application*, SPE Prod. Eng., 6 (1991), pp. 171–180.
[5] S. Benzoni-Gavage, *Analyse numérique des modèles hydrodynamiques d'écoulements diphasiques instationnaires dans les réseaux de production pétrolière*, Thèse, ENS Lyon, Lyon cedex, France, 1991.
[6] P. A. Berthelsen, *Calculations of stratified wavy two-phase flow in pipes*, Int. J. Multiphase Flow, 31 (2005), pp. 571–592.
[7] D. Bestion, *The physical closure laws in the CATHARE code*, Nucl. Eng. Des., 124 (1990), pp. 229–245.
[8] F. Bouchut, Y. Brenier, J. Cortes, and J.-F. Ripoll, *A hierarchy of models for two-phase flows*, J. Nonlinear Sci., 10 (2000), pp. 639–660.
[9] J. A. Bouré, *Wave phenomena and one-dimensional two-phase flow models*, Multiphase Sci. Tech., 9 (1997), pp. 1–107.
[10] D. Drew, L. Cheng, and R. T. Lahey, Jr., *The analysis of virtual mass effects in two-phase flow*, Int. J. Multiphase Flow, 5 (1979), pp. 233–242.
[11] S. Evje and K. K. Fjelde, *Hybrid flux-splitting schemes for a two-phase flow model*, J. Comput. Phys., 175 (2002), pp. 674–701.
[12] S. Evje and T. Flåtten, *Hybrid flux-splitting schemes for a common two-fluid model*, J. Comput. Phys., 192 (2003), pp. 175–210.
[13] S. Evje and T. Flåtten, *Hybrid central-upwind schemes for numerical resolution of two-phase flows*, Math. Model. Numer. Anal., 39 (2005), pp. 253–273.
[14] S. Evje and T. Flåtten, *Weakly implicit numerical schemes for a two-fluid model*, SIAM J. Sci. Comput., 26 (2005), pp. 1449–1484.
[15] F. França and R. T. Lahey, Jr., *The use of drift-flux techniques for the analysis of horizontal two-phase flows*, Int. J. Multiphase Flow, 18 (1992), pp. 787–801.
[16] S. L. Gavrilyuk and J. Fabre, *Lagrangian coordinates for a drift-flux model of a gas–liquid mixture*, Int. J. Multiphase Flow, 22 (1996), pp. 453–460.
[17] K. H. Karlsen, N. H. Risebro, and J. D. Towers, *Front tracking for scalar balance laws*, J. Hyperbolic Differ. Equ., 1 (2004), pp. 115–148.
[18] R. T. Lahey, Jr., L. Y. Cheng, D. A. Drew, and J. E. Flaherty, *The effect of virtual mass on the numerical stability of accelerating two-phase flows*, Int. J. Multiphase Flow, 6 (1980), pp. 281–294.
[19] R. T. Lahey, Jr., *Void wave propagation phenomena in two-phase flow*, AIChE J., 37 (1991), pp. 123–135.
[20] M. Larsen, E. Hustvedt, P. Hedne, and T. Straume, *PeTra: A novel computer code for simulation of slug flow*, in Proceedings of the SPE Annual Technical Conference and Exhibition, SPE 38841, San Antonio, TX, 1997, pp. 1–12.
[21] D. L. Nguyen, E. R. F. Winther, and M. Greiner, *Sonic velocity in two-phase systems*, Int. J. Multiphase Flow, 7 (1981), pp. 311–320.
[22] C. Pauchon, H. Dhulesia, D. Lopez, and J. Fabre, *TACITE: A comprehensive mechanistic model for two-phase flow*, in Proceedings of the BHRG Conference on Multiphase Production, Cannes, France, 1993.
[23] V. H. Ransom and D. L. Hicks, *Hyperbolic two-pressure models for two-phase flow*, J. Comput. Phys., 53 (1984), pp. 124–151.
[24] H. B. Stewart and B. Wendroff, *Review article; two-phase flow: Models and methods*, J. Comput. Phys., 56 (1984), pp. 363–409.
[25] Y. Taitel and A. E. Dukler, *A theoretical approach to the Lockhart-Martinelli correlation for stratified flow*, Int. J. Multiphase Flow, 2 (1976), pp. 591–595.
[26] B. Théron, *Ecoulement diphasique instationnaires en conduite horizontale*, Thèse, INP Toulouse, Toulouse cedex, France, 1989.
[27] I. Toumi, *An upwind numerical method for two-fluid two-phase flow models*, Nuc. Sci. Eng., 123 (1996), pp. 147–168.
[28] I. Toumi and A. Kumbaro, *An approximate linearized Riemann solver for a two-fluid model*, J. Comput. Phys., 124 (1996), pp. 286–300.
[29] I. Toumi, A. Kumbaro, and H. Paillère, *Approximate Riemann solvers and flux vector splitting schemes for two-phase flow*, Commisariat a l'Energie Atomique, Paris cedex, France, 1999. Notes presented at the 30th VKI Computational Fluid Dynamics Lecture Series, Rhode-Saint-Genèse, Belgium, 1999.
[30] N. Zuber and J. A. Findlay, *Average volumetric concentration in two-phase flow systems*, J. Heat Transfer, 87 (1965), pp. 453–468.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 512–529

PATTERN FORMATION IN AN ARRAY OF OSCILLATORS WITH ELECTRICAL AND CHEMICAL COUPLING*

FATMA GUREL KAZANCI† AND BARD ERMENTROUT†‡

Abstract. Weak coupling theory is applied to a model for firing waves in the procerebral lobe of the slug. Inhibitory synapses and electrical synapses have different synchronizing properties. We show that, in concert, these two types of coupling can cause a bifurcation to a patterned state from synchrony which ultimately develops into traveling waves. Normal forms for the bifurcation are computed, and the results are compared to numerical simulations of the phase models.

Key words. electrical coupling, inhibition, oscillations, waves, synchrony

AMS subject classifications. 37G05, 92C20

DOI. 10.1137/060661041

1. Introduction. Networks of coupled neural oscillators exhibit a variety of activity patterns according to the properties of the coupling. There is clear experimental evidence for the existence of electrical and chemical synapses in neocortical inhibitory networks [11]. The effect of each type of coupling in isolation is well studied [4, 17, 19]. Depending on the nature of the neural oscillation, inhibition can be either synchronizing or desynchronizing [19, 12]. Electrical coupling between oscillators is established via gap junctions. In numerous computational and theoretical studies, it has been shown that electrical coupling can promote either synchrony or antisynchrony [18, 1, 4, 3], depending on the shape of the action potential and the nature of the oscillator. Recently, the combined effects of these couplings have been an area of theoretical interest [16, 13, 2, 17]. In these papers, both the inhibition and the gap junctions encouraged synchronization. Coupling is between pairs of oscillators or in all-to-all coupled networks.

In a recent paper [6], Ermentrout et al. explored a biophysical model for the olfactory lobe of the garden slug. Under resting conditions the slug lobe produces slow periodic traveling waves of electrical activity. The oscillations are generated by a class of inhibitory bursting neurons, which are coupled via gap junctions and chemical inhibitory synapses. Experimentally, the wave of activity is biased to move in one direction because of an intrinsic gradient in the frequency of the bursting cells. In the above paper, the authors developed and simulated a biophysical model for the waves with both inhibition and gap junctions. They found that even in the absence of a gradient in frequency, it was possible to generate waves in an otherwise homogeneous network. With large enough gap junction coupling (or small enough inhibition), the network synchronized. However, with weak electrical coupling the network becomes desynchronized, breaking into clusters of cells with different phase-lags. At intermediate coupling strengths, the network produces waves.

Our broad goal in this paper is to explore what happens in spatially distributed

*Received by the editors May 26, 2006; accepted for publication (in revised form) November 7, 2006; published electronically February 9, 2007. The authors were supported by National Science Foundation grant DMS05135.
http://www.siam.org/journals/siap/67-2/66104.html
†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (fag4@pitt.edu, bard@math.pitt.edu).
‡Corresponding author.

networks when one form of coupling (here, gap junctions) encourages synchrony but the other form of coupling (chemical inhibition) encourages (at least pairwise) anti-phase (half cycle apart) locking. More specifically, we suppose that the synchronous coupling is local and the desynchronizing coupling is long range. Since electrical junctions require that membranes of the cells be in direct contact, we expect that gap junction coupling is spatially localized. In contrast, chemical inhibition might be expected to have longer range. In the slug brain model, the inhibition is global, in that each cell inhibited all the other cells in the network, while the gap junctions were only between nearest neighbors. We show below that inhibition is desynchronizing for the slug model and that gap junctions synchronize, so the slug model serves as an example of a spatially distributed network in which the two types of coupling work in opposition. Ermentrout and Kopell [10] explored the effects of one or two long range desynchronizing interactions between cells that were coupled with local synchronizing interactions. Various types of waves were found via direct analytic calculations which were possible due to the simple form of the coupling.

In this paper, we explore the bifurcation to patterns in a general network of oscillators in which there is long range desynchronizing coupling and short range synchronizing coupling. The strength of the former coupling is a parameter which when increased causes the synchronous state to lose stability. We determine the critical values for this parameter via linear stability analysis, and the direction of the bifurcation via a normal form calculation. To make the analysis possible and to avoid the confound of boundary effects, we forgo the linear chain and work on a circular domain. Numerical results of the chain produce similar behavior, but the analysis is considerably more difficult. The normal form calculation is made somewhat more difficult by the presence of a zero eigenvalue arising from translation invariance. Our method is to first reduce the biophysical model to a chain of phase-coupled oscillators on which we can apply the general theory. Thus, in the first section, after introducing the biophysical model, we compute the interaction functions under the assumption of weak coupling. We show that for this model, gap junctions are synchronizing, while chemical inhibition is desynchronizing. Next, we analyze the bifurcation of patterned states from synchrony in a continuum chain of phase-oscillators. We find a novel phase-locked state which is patterned but not a traveling wave. We numerically illustrate the transition to traveling waves as predicted in the reduced system and provide conditions for the stability of the traveling wave.

2. The model and reduction. Ermentrout et al. introduced a biophysical model for a network of bursting and nonbursting cells in the procerebral lobe of Limax [6]. The bursting cells oscillate at about 1 Hz and are responsible for the electrical wave observed in the lobe. The nonbursting cells fire only in the presence of extrinsic stimuli. Thus, since we are interested only in the genesis of the wave, we focus on the bursting cells. Each cell is an intrinsic oscillator, and, in the model, two types of synapses couple the oscillating neurons: chemical inhibition and electrical or gap junctions. The membrane potential for each bursting cell obeys the following equations:

$$C\frac{dV}{dt} = -I_L - I_K - I_{Ca} - I_{gap} - I_{syn},$$

where each term is a current due respectively to the leak, the potassium channels, the calcium channels, the gap junction coupling, and the synaptic inhibition. We used the parameters given in Appendix B. The gap junction coupling is over nearest neighbors

and depends on the voltage difference between the pre- and postsynaptic cells:

$$I_{gap} = g_{gap}(V_{post} - V_{pre}).$$

Here, "post" refers to the cell receiving the connection from the "pre" cell. The inhibition, I_{syn}, is global—every cell inhibits every other cell. Each synaptic interaction adds a current of the form

$$I_{syn} = g_{syn} s_{pre}(V_{post} - E_{syn}),$$

where $E_{syn} = -78$ mV, and the synaptic conductance obeys an equation of the form

$$\frac{ds_{pre}}{dt} = \frac{0.1}{(1 + \exp(-(V_{pre} + 45)/5))} - \frac{s_{pre}}{100}.$$

Networks of coupled oscillators are generally difficult to analyze. However, the method of averaging has proven to be very useful for studying synchronization between oscillators [17]. That is, if we assume that the conductances g_{gap}, g_{syn} are sufficiently small, it is possible to reduce a network of coupled oscillators to a system of phase models where each oscillator is represented by its scalar phase and interactions are through the differences in the phases [21, 15, 9]. Let V_i be the membrane potential of the ith cell and s_i be the synaptic component of the ith cell. If

$$-I_{syn,i} = -g_{syn} \sum_j w_{ij} s_j (V_i - E_{syn})$$

is the synaptic current into the ith cell and w_{ij} is the weight of the connection between cell i and j, which is taken to be $1/N$, where N is the number of oscillators, then with the weak coupling assumption, the phase interactions will take the form

$$-\bar{I}_{syn,i} = g_{syn} \sum_j w_{ij} H_{syn}(\theta_j - \theta_i), \tag{2.1}$$

where

$$H_{syn}(\phi) = \frac{1}{T} \int_0^T V^*(t) s(t + \phi)(E_{syn} - V(t))\, dt.$$

$V^*(t)$ is the voltage component for the T-periodic solution to the adjoint equation for the stable limit cycle. $V(t)$, $s(t)$ are the voltage and synaptic components, respectively. For the gap junction coupling, we find

$$-\bar{I}_{gap,i} = g_{gap} \sum_j z_{ij} H_{gap}(\theta_j - \theta_i), \tag{2.2}$$

where

$$H_{gap}(\phi) = \frac{1}{T} \int_0^T V^*(t)(V(t + \phi) - V(t))\, dt.$$

The weights, z_{ij}, satisfy $z_{ij} = f(|i - j|)$, where f is a decreasing function in its argument. The phase of each oscillator, θ_i, obeys the reduced dynamics

$$\theta_i' = 1 - \bar{I}_{syn,i} - \bar{I}_{gap,i},$$

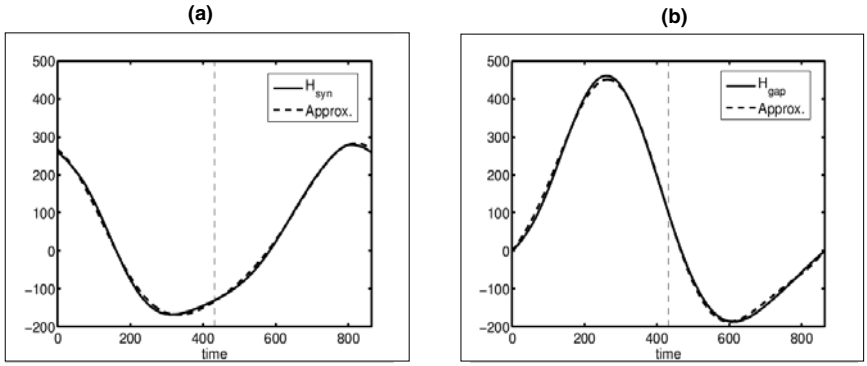


FIG. 2.1. *Coupling functions are computed using XPPAUT. Approximations are estimated from the Fourier expansion.* (a) *shows* H_{syn} *and its approximation, and* (b) *shows* H_{gap} *and its approximation. The functions are plotted over a period of the oscillations, and the dashed line marks the half period.*

where the two currents are given by (2.1) and (2.2). The phase of each oscillator maps directly onto the potential (or other variable) of each bursting cell once the zero phase is chosen. A standard choice of zero phase is the peak of the membrane potential.

Figure 2.1 shows both H_{syn} and H_{gap} for the Limax model evaluated numerically using XPPAUT [5], along with their approximations using the 0, 1, 2 order terms of the Fourier expansion. The Fourier approximations of these functions are used in bifurcation calculations in section 3, the stability arguments in section 4, and the numerical simulations of the phase model in section 5. Their values are given in Appendix B. Note that $H_{syn}(0) \neq 0$, $H_{gap} = 0$, $H'_{syn}(0) < 0$, and $H'_{gap}(0) > 0$.

To interpret the meaning of these inequalities, consider a pair of identical cells:

$$\theta'_1 = 1 + g_{syn}H_{syn}(\theta_2 - \theta_1) + g_{gap}H_{gap}(\theta_2 - \theta_1),$$
$$\theta'_2 = 1 + g_{syn}H_{syn}(\theta_1 - \theta_2) + g_{gap}H_{gap}(\theta_1 - \theta_2).$$

Let $\phi = \theta_2 - \theta_1$. Then

$$\phi' = g_{syn}[H_{syn}(-\phi) - H_{syn}(\phi)] + g_{gap}[H_{gap}(-\phi) - H_{gap}(\phi)] \equiv F(\phi).$$

Clearly, $F(0) = 0$, so synchrony, $\theta_2 = \theta_1$, is a solution. Synchrony is stable if $F'(0) < 0$ or

$$g_{syn}H'_{syn}(0) + g_{gap}H'_{gap}(0) > 0.$$

Since the conductances, g_{syn}, g_{gap} are nonnegative and $H'_{syn}(0) < 0$, $H'_{gap}(0) > 0$, synchrony is stable if the gap junctions dominate. Since $F(\phi)$ is an odd T-periodic function, $F(T/2) = 0$. This *antiphase* solution will be stable for the coupled pair if $F'(T/2) < 0$ or, equivalently,

$$g_{syn}H'_{syn}(T/2) + g_{gap}H'_{gap}(T/2) > 0.$$

As seen in Figure 2.1 by the dashed vertical lines at $T/2$, antiphase is stable for synaptic and unstable for gap junction coupling. In the models considered by Lewis and Rinzel, both synaptic and electrical coupling encourage stable synchrony [16]. Thus, the interaction of networks will lead to synchronous behavior. In contrast, for

the intrinsic dynamics in the Limax model, electrical coupling encourages synchrony, but synaptic inhibition opposes it. Our goal in the rest of this paper is to explore the consequences of these differences in a one-dimensional spatially organized array of N oscillators.

2.1. The spatial equations. We introduce a discrete model where we have all-to-all synaptic coupling and local gap junction coupling. The equations can be written down as

(2.3)
$$\frac{d\theta_j}{dt} = \omega + \frac{g_{syn}}{N}\sum_{k=1}^{N} H_{syn}(\theta_k - \theta_j) + g_{gap}\sum_{l=-m}^{m} J_l H_{gap}(\theta_{j+l} - \theta_j), \quad j = 1, \ldots, N,$$

where θ_j represents the phase of oscillator j, ω is the intrinsic frequency for all the oscillators, g_{syn} is the synaptic coupling strength, g_{gap} is the strength of electrical coupling, and H_{syn}, H_{gap} are the functions describing synaptic and gap junction coupling, respectively. We note that the key point in the weak coupling assumption is that the effects of different types of coupling are linear and additive. Thus, only the ratio of g_{syn} and g_{gap} in the phase model matter. The oscillators are arranged in a ring to avoid boundary effects. W.l.o.g., we assume that the ring has length 2π. The factor $\frac{1}{N}$ in front of the synaptic coupling contribution guarantees that the model also works when we allow $N \to \infty$. The weights J_l are taken to be nonnegative, and we assume that $J_{-l} = J_l$. m represents the scope of gap junction coupling. We also note that $m \ll N$, since we assume that gap junction coupling is local. Henceforth, we assume that the period of the oscillators (and thus of the coupling functions) is 2π. The function H_{syn} favors the antiphase state for pairwise interactions so that π is a stable fixed point for a pair of oscillators coupled with only synaptic coupling. The function H_{gap} favors the in-phase state for pairwise interactions so that 0 is a stable fixed point for a pair of oscillators coupled with only gap junction coupling. This is equivalent to saying the following:

A1. $H'_{syn}(0) < 0$.
A2. $H'_{gap}(0) > 0$.

For simplicity, we also need the following:

A3. $H_{syn}(0) = 0$.
A4. $H_{gap}(0) = 0$.

Note that A4 holds automatically for gap junctions, since a cell cannot be coupled to itself via gap junctions. If $H_{syn}(0) = \kappa \neq 0$, then let $\theta_j = \hat{\theta}_j + (\omega + g_{syn}\kappa)t$. We write

$$\frac{d\hat{\theta}_j}{dt} = \frac{g_{syn}}{N}\sum_{k=1}^{N} \hat{H}_{syn}(\hat{\theta}_k - \hat{\theta}_j) + g_{gap}\sum_{l=-m}^{m} J_l H_{gap}(\hat{\theta}_{j+l} - \hat{\theta}_j), \quad j = 1, \ldots, N,$$

where $\hat{H}_{syn}(\phi) = H_{syn}(\phi) - \kappa$. We can see that $\hat{H}_{syn}(0) = 0$, so w.l.o.g., we assume $H_{syn}(0) = 0$.

We also make a normalization assumption on J_l by taking the following:

A5*. $\sum_{l=-m}^{m} J_l = 1$.

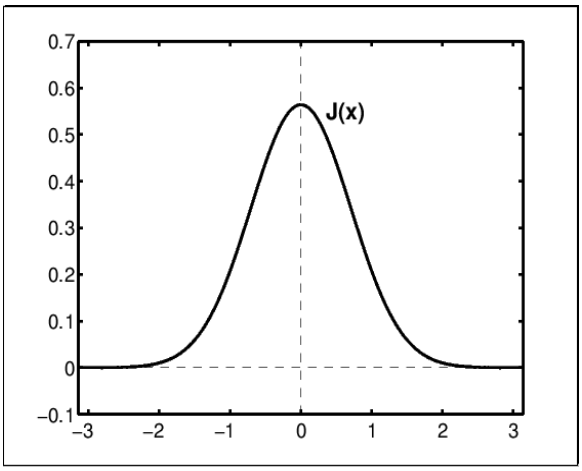


FIG. 2.2. *A plot of $J(x)$ where the sum was taken over $n = -2, -1, 0, 1, 2$.*

For the purposes of calculations, it is much easier to work with the continuum analogue of (2.3), so our analysis will be on a continuum version of the network. Hence, from now on, we study this model

$$\begin{aligned}\frac{\partial\theta}{\partial t} &= \omega + \frac{g_{syn}}{2\pi}\int_0^{2\pi} H_{syn}(\theta(y)-\theta(x))\,dy \\ &\quad + g_{gap}\int_0^{2\pi} J(x-y)\,H_{gap}(\theta(y)-\theta(x))\,dy,\end{aligned} \tag{2.4}$$

where analogous assumptions are made as for the discrete model. We remark that the continuum model can be derived from the discrete model in the limit as $N \to \infty$ with a suitable normalization assumption on the function J_ℓ. One difference is that the discrete model, θ_j, was a function of time and the discrete index j, whereas it is now a function of time and space. We assume that $J(x)$ is a nonnegative, symmetric kernel around 0 and that the normalization condition is

A5. $\int_0^{2\pi} J(y)\,dy = 1$.

In our numerical simulations, we assumed $J(x) = \sum_{n=-\infty}^{\infty} \exp(-(x+2\pi n)^2)/\sqrt{\pi}$. (See Figure 2.2.)

3. Linear stability analysis for synchronous solution. We want to study the spatial interactions between synchronizing and antisynchronizing influences. We start with the synchronous state and study its stability. The synchronous state is where all of the oscillators have the same phase. Note that if we assume heterogeneity in the intrinsic frequencies, synchrony is not a solution to the system. If we have homogeneity, $\theta(x,t) = \Omega t$ is a solution to (2.4), where $\Omega = \omega + g_{syn}H_{syn}(0)$ represents the frequency of the network. To determine the stability of synchrony, we let $\theta(x,t) = \Omega t + \psi(x,t)$ and write

$$\begin{aligned}\frac{\partial\psi}{\partial t} &= \frac{g_{syn}}{2\pi}\int_0^{2\pi} H'_{syn}(0)\,[\psi(y)-\psi(x)]\,dy \\ &\quad + g_{gap}\int_0^{2\pi} J(x-y)\,H'_{gap}(0)\,[\psi(y)-\psi(x)]\,dy + O\left(|\psi|^2\right).\end{aligned} \tag{3.1}$$

If we keep only the linear terms above, we can see that $\psi(x,t) = e^{inx}\ e^{\lambda_n t}$ solves (3.1) with the appropriate choice of λ_n. Let $I_n = \int_0^{2\pi} J(s)\ e^{-ins}\ ds$, and substitute $\psi(x,t)$ into (3.1) to get

$$\lambda_n = -g_{syn}\ H'_{syn}(0) + g_{gap}\ H'_{gap}(0)\ [I_n - 1] \tag{3.2}$$

for $n \neq 0$. For $n = 0$, $\lambda_0 = 0$. We choose $J(x)$ so that we have $I_1 \geq 1$ and $I_1 > I_2 > \cdots > I_n > I_{n+1} > \cdots$. This means that the first Fourier mode dominates. The Gaussian kernel shown in Figure 2.2 satisfies this criterion, as does, for example, the periodic version of an exponential kernel, $\exp(-|x|)$. With this assumption, it is easy to see that the first eigenvalue to cross over to positive values would be λ_1. We call $n = 1$ the most unstable node. To find the critical value of g_{syn}, we solve for $\lambda_1 = 0$, which gives us

$$g^*_{syn} = \frac{g_{gap}\ H'_{gap}(0)\ [I_1 - 1]}{H'_{syn}(0)}. \tag{3.3}$$

Here * is used to denote the value of g_{syn} at the bifurcation point. To study the stability of the bifurcating solutions we need to find the normal form for the bifurcation. We prove the following theorem.

THEOREM 3.1. *The system* (2.4) *with the assumptions* A1–A5 *has a pitchfork bifurcation at* g^*_{syn}*, and the corresponding normal form is*

$$0 = \zeta z^2 \bar{z} + \eta z.$$

The coefficients ζ *and* η *are*

$$\begin{aligned} \zeta &= 12B_{1,3} - 3B_{2,3} - 9B_{0,3} + 2C\ B_{0,2} - 2CB_{2,2}, \\ \eta &= -g_2\alpha_1, \end{aligned}$$

where

$$B_{n,j} = \int_0^{2\pi} A_j(y')\ e^{iny'}\ dy',$$

$$A_j(x) = \frac{g^*_{syn}}{2\pi}\ \alpha_j + g_{gap}\ \beta_j\ J(x),$$

$$C = \frac{2B_{1,2} - B_{2,2} - B_{0,2}}{B_{2,1} - B_{0,1}},$$

with $\alpha_j = \frac{H^j_{syn}(0)}{j!}$ *and* $\beta_j = \frac{H^j_{gap}(0)}{j!}$ *for* $j = 1, 2, \ldots$. *Here* $f^j(x_0)$ *represents the* i*th derivative at* x_0 *for* $f = H_{syn}, H_{gap}$.

Proof. We use a perturbation expansion for the solution ψ and g_{syn} as

$$\theta(x,t) = \Omega(\epsilon)\ t + \hat{\psi}(x,\epsilon), \tag{3.4}$$

where

$$\Omega(\epsilon) = \epsilon\ \Omega_1 + \epsilon^2\ \Omega_2 + \epsilon^3\ \Omega_3 + \cdots,$$

$$\hat{\psi}(x,\epsilon) = \epsilon\ \psi_1(x) + \epsilon^2\ \psi_2(x) + \epsilon^3\ \psi_3(x) + \cdots,$$

and

$$g_{syn} = g^*_{syn} + \epsilon\ g_1 + \epsilon^2\ g_2.$$

We define a linear operator $\mathcal{L}$ as follows:

$$\begin{aligned}
\mathcal{L}\psi &= \frac{g^*_{syn}}{2\pi}\int_0^{2\pi} H'_{syn}(0)[\hat{\psi}(y,\epsilon) - \hat{\psi}(x,\epsilon)]dy \\
&\quad + g_{gap}\int_0^{2\pi} J(x-y)H'_{gap}(0)[\hat{\psi}(y,\epsilon) - \hat{\psi}(x,\epsilon)]dy \\
&= \frac{g^*_{syn}}{2\pi}\int_0^{2\pi} H'_{syn}(0)[\hat{\psi}(x-y',\epsilon) - \hat{\psi}(x,\epsilon)]dy' \\
&\quad + g_{gap}\int_0^{2\pi} J(y')H'_{gap}(0)[\hat{\psi}(x-y',\epsilon) - \hat{\psi}(x,\epsilon)]dy' \qquad (3.5)
\end{aligned}$$

with the substitution $y' = x - y$. Note that $e^{\pm ix}$, $\mathbf{1}$ are in the null space of $\mathcal{L}$ and that $\mathcal{L}$ is self-adjoint. (Here, we use $\mathbf{1}$ to denote the constant function which is 1 for all x.) We need to find the Taylor expansions of H_{syn} and H_{gap} around 0 for the full system

$$\begin{aligned}
H_{syn}(x) &= H_{syn}(0) + H'_{syn}(0)\ x + \frac{H''_{syn}(0)}{2}\ x^2 + \frac{H'''_{syn}(0)}{6}\ x^3 + \cdots \\
&= \alpha_1\ x + \alpha_2\ x^2 + \alpha_3\ x^3 + \cdots, \\
H_{gap}(x) &= H_{gap}(0) + H'_{gap}(0)\ x + \frac{H''_{gap}(0)}{2}\ x^2 + \frac{H'''_{gap}(0)}{6}\ x^3 + \cdots \\
&= \beta_1\ x + \beta_2\ x^2 + \beta_3\ x^3 + \cdots.
\end{aligned}$$

Substituting θ in the form given in (3.4) into (2.4),

$$\begin{aligned}
\Omega(\epsilon) &= \frac{(g^*_{syn} + \epsilon g_1 + \epsilon^2 g_2)}{2\pi}\int_0^{2\pi} \alpha_1\ [\hat{\psi}(x-y',\epsilon) - \hat{\psi}(x,\epsilon)]\ dy' \\
&\quad + \frac{(g^*_{syn} + \epsilon g_1 + \epsilon^2 g_2)}{2\pi}\int_0^{2\pi} \alpha_2\ [\hat{\psi}(x-y',\epsilon) - \hat{\psi}(x,\epsilon)]^2\ dy' \\
&\quad + \frac{(g^*_{syn} + \epsilon g_1 + \epsilon^2 g_2)}{2\pi}\int_0^{2\pi} \alpha_3\ [\hat{\psi}(x-y',\epsilon) - \hat{\psi}(x,\epsilon)]^3\ dy' \\
&\quad + g_{gap}\int_0^{2\pi} J(y')\ \beta_1\ [\hat{\psi}(x-y',\epsilon) - \hat{\psi}(x,\epsilon)]\ dy' \\
&\quad + g_{gap}\int_0^{2\pi} J(y')\ \beta_2\ [\hat{\psi}(x-y',\epsilon) - \hat{\psi}(x,\epsilon)]^2\ dy' \\
&\quad + g_{gap}\int_0^{2\pi} J(y')\ \beta_3\ [\hat{\psi}(x-y',\epsilon) - \hat{\psi}(x,\epsilon)]^3\ dy' + O\left(|\hat{\psi}|^4\right). \qquad (3.6)
\end{aligned}$$

Let $A_j(x) = \frac{g^*_{syn}}{2\pi}\ \alpha_j + g_{gap}\ \beta_j J(x)$ for $j = 1, 2, 3$ and $Q(x) = \int_0^{2\pi}[\alpha_1(\hat{\psi}(x-y',\epsilon) - \hat{\psi}(x,\epsilon)) + \alpha_2(\hat{\psi}(x-y',\epsilon) - \hat{\psi}(x,\epsilon))^2 + \alpha_3(\hat{\psi}(x-y',\epsilon) - \hat{\psi}(x,\epsilon))^3]\ dy'$, which lets us to

rewrite (3.6) as

$$\begin{aligned}\Omega(\epsilon) = &\int_0^{2\pi} A_1(y')\ [\hat{\psi}(x-y',\epsilon) - \hat{\psi}(x,\epsilon)]\ dy' \\ &+ \int_0^{2\pi} A_2(y')\ [\hat{\psi}(x-y',\epsilon) - \hat{\psi}(x,\epsilon)]^2\ dy' \\ &+ \int_0^{2\pi} A_3(y')\ [\hat{\psi}(x-y',\epsilon) - \hat{\psi}(x,\epsilon)]^3\ dy' \\ &+ \epsilon\ \frac{g_1}{2\pi}\ Q(x) + \epsilon^2\ \frac{g_2}{2\pi}\ Q(x) + O\left(|\hat{\psi}|^4\right). \end{aligned} \tag{3.7}$$

We match the coefficients of powers of ϵ terms from both sides of (3.7). This allows us to compute the coefficients for the normal form. The rest of the calculations are given in the appendix. The normal form for the bifurcation is

$$0 = \zeta z^2 \bar{z} + \eta z,$$

where $\zeta = 12B_{1,3} - 3B_{2,3} - 9B_{0,3} + 2CB_{0,2} - 2CB_{2,2} - 2CB_{2,2}$ and $\eta = -g_2\alpha_1$. Note that η is positive since $\alpha_1 < 0$ from our assumptions. Thus, depending on the sign of ζ, we can determine the stability of the new solutions. □

In our case, we compute $\zeta = -210.09$ and $\eta = 105g_2$, which tells us that we have a supercritical pitchfork bifurcation. The new solution bifurcating from synchrony is stable.

4. Existence and stability of the traveling wave. Next, we look at the existence and stability of the traveling wave, $\theta(x,t) = \Omega t + x$. Substituting θ back into (2.4) gives us

$$\Omega = \omega + \frac{g_{syn}}{2\pi}\int_0^{2\pi} H_{syn}(y)dy + g_{gap}\int_0^{2\pi} J(y)H_{gap}(y)dy. \tag{4.1}$$

W.l.o.g., we can assume that the average of $H_{syn}(y)$ is zero and so let $I = \int_0^{2\pi} J(y)H_{gap}(y)dy$; (4.1) reduces to $\Omega = \omega + g_{gap}I$. (The value of the frequency is irrelevant to the stability calculation since the right-hand sides always involve terms of the form $\theta(x,t) - \theta(y,t)$ so that adding Ct to θ, where C is any constant, has no effect.) We prove the following theorem about the stability of the traveling wave.

THEOREM 4.1. *The traveling wave solution, $\theta(x,t) = \Omega t + x$, is a stable solution to* (2.4) *if we have*

$$\begin{aligned} Re(\lambda_n) = &-\frac{1}{2}g_{syn}nb_n + 2\pi g_{gap}\sum_{m=-\infty}^{-n-1}\frac{m}{4}[(c_m f_{n+m} - d_m e_{n+m}) - (c_m f_m - d_m e_m)] \\ &- 2\pi g_{gap}\sum_{m=-n}^{-1}\frac{m}{4}[(c_m f_{n+m} + d_m e_{n+m}) + (c_m f_m - d_m e_m)] \\ &- 2\pi g_{gap}\sum_{m=1}^{\infty}\frac{m}{4}[(c_m f_{n+m} - d_m e_{n+m}) - (c_m f_m - d_m e_m)] \\ &\leq 0 \end{aligned}$$

for $n > 0$, where

$$H_{syn}(y) = \sum_{n=0}^{\infty} a_n \cos ny + b_n \sin ny,$$

$$H_{gap}(y) = \sum_{n=0}^{\infty} c_n \cos ny + d_n \sin ny,$$

$$J(y) = \sum_{n=0}^{\infty} e_n \cos ny + f_n \sin ny.$$

Proof. Letting $\theta(x,t) = \Omega t + x + \psi(x,t)$, we write

$$\begin{aligned} \frac{\partial \psi}{\partial t} = \frac{g_{syn}}{2\pi} &\int_0^{2\pi} H'_{syn}(y)\ [\psi(x+y) - \psi(x)]\ dy \\ &+ g_{gap} \int_0^{2\pi} J(-y)\ H'_{gap}(y)\ [\psi(x+y) - \psi(x)]\ dy + O\left(|\psi|^2\right). \end{aligned} \tag{4.2}$$

$\psi(x,t) = e^{inx} e^{\lambda_n t}$ solves (4.2) up to linear order. We solve for λ_n to get

$$\lambda_n = \frac{g_{syn}}{2\pi} \int_0^{2\pi} H'_{syn}(y)[e^{iny} - 1]dy + g_{gap} \int_0^{2\pi} J(-y)\ H'_{gap}(y)[e^{iny} - 1]dy. \tag{4.3}$$

We look at the real part of λ_n, for H_{syn}, H_{gap}, and J real-valued,

$$Re(\lambda_n) = \frac{g_{syn}}{2\pi} \int_0^{2\pi} H'_{syn}(y)[\cos ny - 1]dy + g_{gap} \int_0^{2\pi} J(-y)\ H'_{gap}(y)[\cos ny - 1]dy.$$

When $n = 0$, $Re(\lambda_0) = 0$, which corresponds to translation invariance. We need $Re(\lambda_n) \leq 0$ for $n \neq 0$. Also, we want to make our analysis as general as possible. For this reason we use the complex Fourier series expansion for H_{syn}, H_{gap}, and J. Let $J(y) = \sum_{k=-\infty}^{\infty} \alpha_k e^{iky}$, $H_{syn}(y) = \sum_{l=-\infty}^{\infty} \beta_l e^{ily}$, $H_{gap}(y) = \sum_{m=-\infty}^{\infty} \gamma_m e^{imy}$, where $\alpha_{-k} = \alpha_k$, $\beta_{-l} = \beta_l$, and $\gamma_{-m} = \gamma_m$. Substituting these into (4.3) gives

$$\lambda_n = -ig_{syn} n \beta_{-n} + 2\pi i g_{gap} \sum_{m=-\infty}^{\infty} [m\gamma_m(\alpha_{-(n+m)} - \alpha_{-m})].$$

If we look at the real part of λ_n, we see that

$$\begin{aligned} Re(\lambda_n) = &-\frac{1}{2} g_{syn} n b_n + 2\pi g_{gap} \sum_{m=-\infty}^{-n-1} \frac{m}{4}[(c_m f_{n+m} - d_m e_{n+m}) - (c_m f_m - d_m e_m)] \\ &- 2\pi g_{gap} \sum_{m=-n}^{-1} \frac{m}{4}[(c_m f_{n+m} + d_m e_{n+m}) + (c_m f_m - d_m e_m)] \\ &- 2\pi g_{gap} \sum_{m=1}^{\infty} \frac{m}{4}[(c_m f_{n+m} - d_m e_{n+m}) - (c_m f_m - d_m e_m)], \end{aligned}$$

where H_{syn}, H_{gap}, and J are given with the Fourier expansion with coefficients $a_0 = 2\beta_0$, $c_0 = 2\gamma_0$, $e_0 = 2\alpha_0$, $a_n = \beta_n + \beta_{-n}$, $b_n = i(\beta_n - \beta_{-n})$, $c_n = \gamma_n + \gamma_{-n}$, $d_n = i(\gamma_n - \gamma_{-n})$, $e_n = \alpha_n + \alpha_{-n}$, and $e_n = i(\alpha_n - \alpha_{-n})$. ∎

In our case, the traveling wave is always stable. Substituting our parameters into the eigenvalue equation, we see that $Re(\lambda_n) \leq -50g_{syn} - 824.27g_{gap} \leq 0$ for all positive values of g_{syn} and g_{gap}.

We close this section with some comments on the existence and stability of traveling waves in the discrete system for local gap junction coupling. Consider a discrete ring,

$$\frac{d\theta_j}{dt} = \omega + \sum_{i=-m}^{m} a_i H(\theta_{j+m} - \theta_j),$$

where $m \ll N$ and N is the number of oscillators. The coupling constants a_i are nonnegative. Suppose that H is 2π-periodic and that $H'(x) > 0$ for $-r < x < r$ and $r > 0$. Then, it follows from [7] that the synchronous state is asymptotically stable. Now, consider a traveling wave:

$$\theta_j = \Omega t + 2\pi j/N.$$

This satisfies the discrete model if and only if

$$\Omega = \omega + \sum_{i=-m}^{m} a_i H(2\pi i/N).$$

If m/N is sufficiently small, then

$$H'(\pm 2\pi m/M) > 0$$

since $H'(x)$ is positive in some neighborhood of 0. Thus, again from [7], the traveling wave is asymptotically stable. Figure 2.1(b) shows that $H'_{gap}(x) > 0$ over more than half the cycle surrounding the origin. Thus, we can pick m as large as $N/4$ and still be assured that the traveling wave is stable. This shows that there is bistability between traveling waves and synchrony in the discrete model with small enough synaptic coupling.

5. Numerical results. In this section we (i) show that the bifurcation theory developed for the continuum model appears to hold for the discrete model by numerically simulating the latter, (ii) numerically extend the local bifurcation analysis to get the full picture for the discrete phase-model, (iii) numerically simulate the conductance-based model and show patterns similar to those found via our analysis, and (iv) compute the bifurcation diagram for a line of 20 oscillators which are *not* connected in a ring.

Figure 5.1 depicts the steady-state relative phases for a ring of 20 phase-oscillators using the interaction functions shown in Figure 2.1. The strength of the gap junction coupling is fixed at .01, and g_{syn} is varied along the vertical axis. Simulations are done by starting the relative phases close to synchrony and then letting them evolve until a steady state is reached. Figure 5.1(a) shows this steady state (color-coded) for each value of g_{syn} examined. (We remark that in the phase model, the absolute value of the coupling parameters is irrelevant, and only their ratio matters.) Figure 5.1(b) shows vertical cross sections from part (a) to more clearly illustrate different types of solutions observed for various $\rho \equiv g_{syn}/g_{gap}$ values. For example, when $\rho = 0.3$, there is no difference in phases of the oscillators, indicating that the system is synchronized. In contrast, between $\rho \approx 0.35$ and $\rho \approx 0.87$, the solution is the

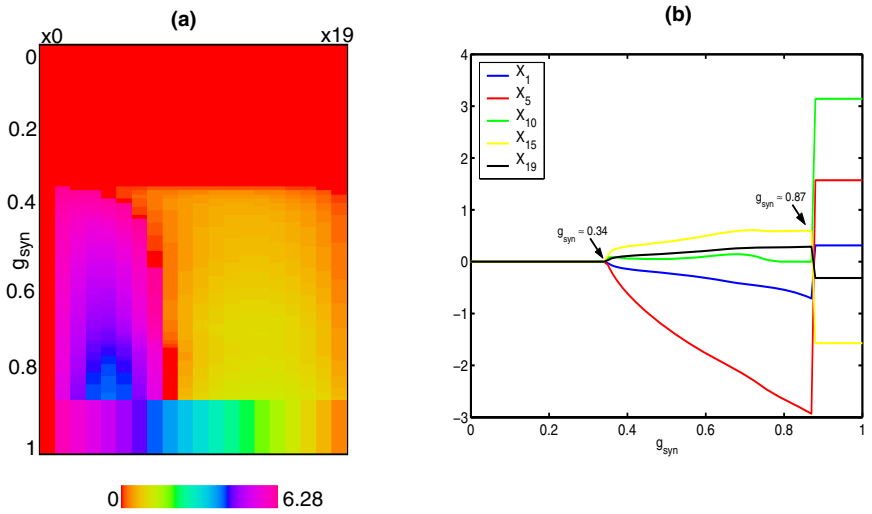


FIG. 5.1. *Transition from synchrony to intermediate state and then to traveling wave.* (a) *illustrates an array plot of the relative phases of the oscillators as* $\rho \equiv g_{syn}/g_{gap}$ *is increased;* (b) *shows the phases of some oscillators as* ρ *is increased.*

patterned state which bifurcates from the synchronous state as described in section 3. As ρ increases beyond 0.87, the patterned state (which qualitatively resembles a cosine wave) disappears and leaves a traveling wave as the only solution. The traveling wave is, in fact, stable for all ρ shown in the diagram, so that for $\rho < 0.87$ there is bistability. The loss of stability of the synchronous state occurs at $\rho \approx 0.35$, which is very close to the value of 0.3476 predicted in section 2.

To give the reader some intuition for the patterns, we depict the spatio-temporal patterns in terms of their absolute phase in Figure 5.2. As we increase the relative coupling strength, we see the transition from synchrony to a stable patterned state (compare 5.2(a) and (b)). This is the state which arises via the pitchfork bifurcation calculated in section 2. As we further increase g_{syn}, the patterned state disappears and produces traveling waves; the transition from the patterned state to the waves is shown in Figure 5.2(c). Finally, for larger g_{syn}, only the traveling wave remains.

The analytic calculations along with the numerical calculations of the phase reduced model show that as the inhibition increases, the synchronous state loses stability to a patterned state in which the relative phases are close to a cosine wave. Further increases in the inhibition result in a deepening of this pattern, followed by a transition to a traveling wave. In Figure 5.3, we show the result of a simulation of the biophysical model as the synaptic inhibition increases. To match the theory, we have made the connections periodic, so that the last cell is coupled to the first. Figure 5.3(a) shows a clear phase pattern in which the oscillators at the end lag the ones in the middle. This corresponds to the patterned state shown in Figures 5.2(b) and 5.1(a), when $\rho \approx 0.4$. For a larger amount of inhibition, the behavior becomes quite complicated and, after a long transient, begins a transition to traveling waves, as shown in Figure 5.3(b). Thus, the phase model provides a very good description of the full biophysical model and has the advantage of being simple enough to analyze.

We conclude this section with some comments on the simplification to a ring of oscillators instead of a line as in the original model. The main reason for assuming a

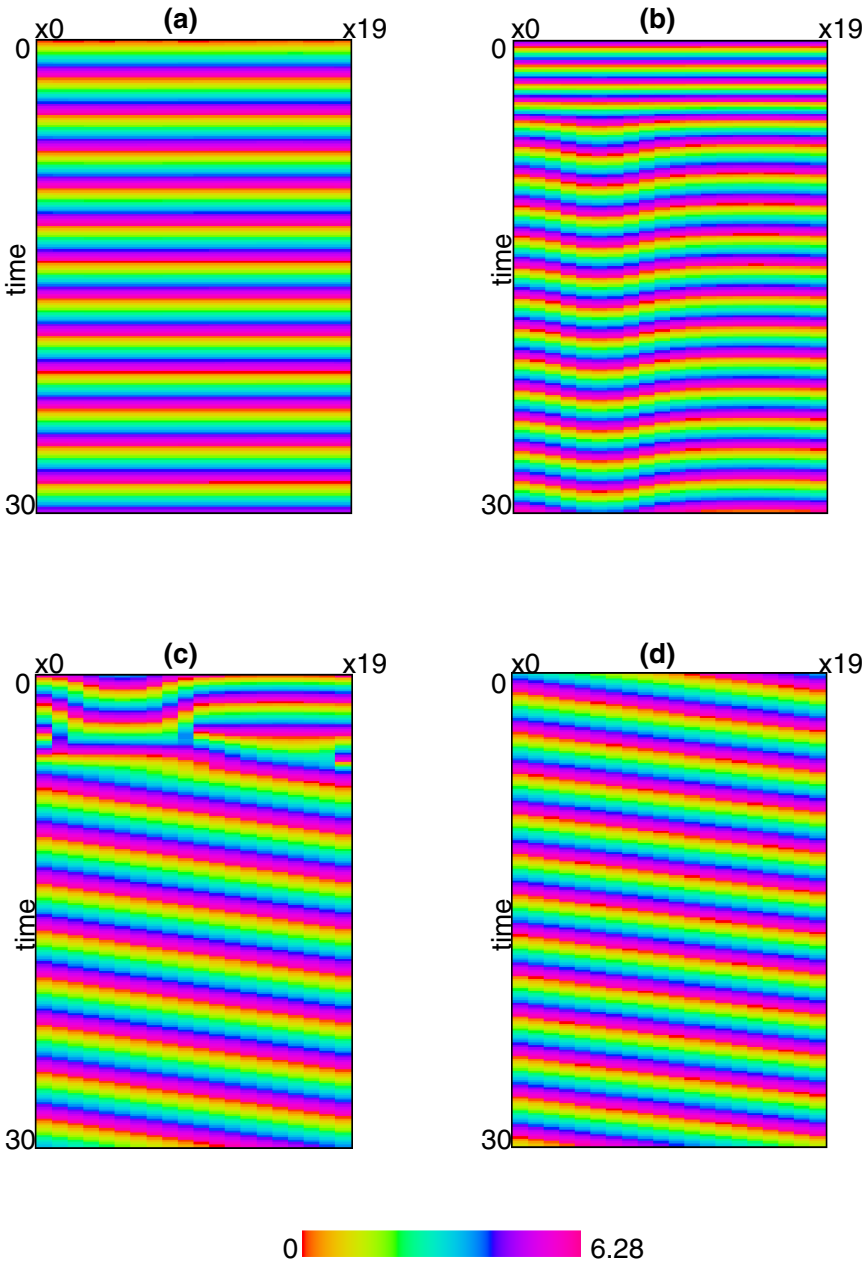


FIG. 5.2. *Evolution of the solution to the discrete phase model is shown as we change* g_{syn} *while* g_{gap} *is kept at the value* 0.03. (a) *shows synchronous solution when* $g_{syn} = 0.01$, (b) *shows the intermediate state when* $g_{syn} = 0.02$, (c) *shows the transition to the traveling wave solution when* $g_{syn} = 0.03$, *and* (d) *shows the traveling wave solution when* $g_{syn} = 0.04$.

ring is that the analytic calculations are then possible. If, instead of a ring, we consider a line of oscillators and choose the coupling functions so that the synchronous state exists, we can explore the stability and bifurcations as the antisynchronous (synaptic inhibition) coupling increases. Rather than attempt these calculations analytically, we instead display numerical simulations for the phase model with all-to-all synaptic

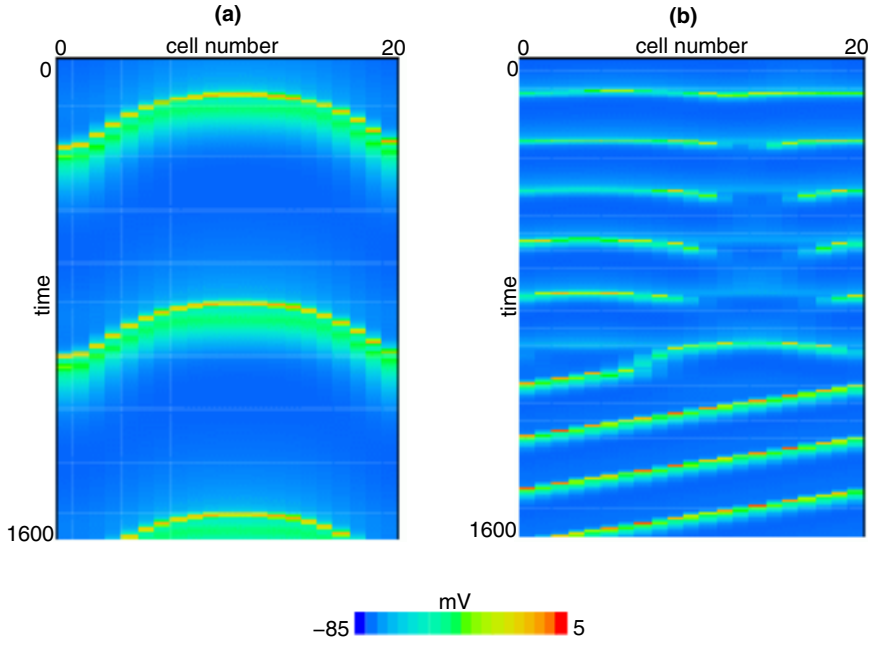


FIG. 5.3. *Behavior of the conductance-based model for* $g_{syn} = 0.03$ *and* $g_{syn} = 0.07$. *Voltage is plotted for each oscillator.*

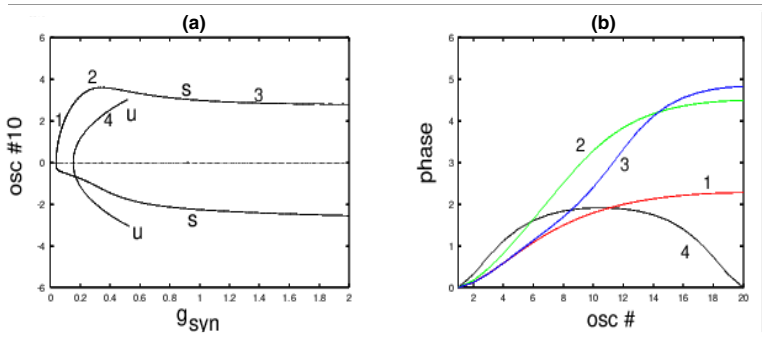


FIG. 5.4. *Behavior of* 20 *phase oscillators in a* line *as the synaptic coupling increases.* (a) *The relative phase of oscillator* 10; $g_{gap} = 1$ *and the interaction functions are the same as in the ring model.* (b) *Spatial profiles with various solutions from* (a). *u refers to the solution being unstable, and s means the solution is stable.*

coupling and nearest neighbor gap junction coupling.

Figure 5.4 shows the behavior for a line of 20 oscillators. By considering a linear array, the symmetry in the ring model was broken, and we are able to use the AUTO bifurcation package [5]. We depict two pitchfork bifurcations. The first emerges as a stable supercritical bifurcation. The pattern is like a half of a cosine wave, as opposed to the full cycle seen in a ring. The ring of oscillators can be imagined as a pair of lines joined symmetrically through the midline. Thus, we expect that the first bifurcation would be "half" of that seen in the ring (see curve 1 in Figure 5.4(b)). As g_{syn} increases, this branch seems to approach a solution which looks like a traveling

wave (Figure 5.4(b) curves 2 and 3). There is no true traveling wave in the line due to the boundary conditions; however, the solutions in the figure look like traveling waves. A second branch bifurcates supercritically but it inherits the instability of the synchronous branch, so that it is unstable. The shape of this solution is shown by curve 4 in Figure 5.4(b). As these solutions were unstable, they were not continued beyond $g_{syn} = 0.5$. Thus, while the details are somewhat different, the ultimate result is the same for both a ring and a linear array: as g_{syn} increases, synchrony loses stability, and for large enough g_{syn} there is a traveling wave. The traveling wave exists for all values of g_{syn} in the ring model but not for the linear array.

6. Discussion. In this paper, we have shown that the combination of long range inhibitory synaptic coupling with local gap junction coupling was sufficient to induce a destabilization of the synchronous state. A new state which is not a traveling wave but rather a spatially organized phase shift stably appears and is lost as the amount of long range inhibitory coupling increases. Numerical solutions indicate that the only remaining attracting state is a traveling wave. Our mathematical results concern a network on a ring; the original motivation for this problem is the slug olfactory lobe, which is actually a line of oscillators. However, it is known from our earlier work [14] that boundary effects are enough to induce patterns of phases that depend very strongly on the choices of boundary conditions at the edges. To avoid this difficulty, we have considered periodic boundary conditions which eliminate questions about the behavior at the edges. In spite of this simplifying assumption, we see that the linear array and the ring behave similarly, at least when the inhibition is sufficiently large compared to electrical coupling.

A number of studies have investigated interactions between electrical coupling and synaptic coupling between neural oscillators. This problem is important since inhibitory interneurons in the mammalian neocortex appear to be coupled with both types of interactions. These networks may act as the "pacemakers" for 40 Hz oscillations observed in the cortex during various cognitive tasks [20]. Most theoretical explorations involve either pairs of cells or globally coupled networks. In most instances, both the synaptic and the electrical coupling encourage synchrony, so that there is not a chance for pattern formation. However, [4] has shown that gap junctions can either stabilize or destabilize synchrony, depending on the shape of the action potential, while [17] has shown that the intrinsic currents also affect whether or not electrical coupling is synchronizing. Combining coupling that destabilizes with coupling that stabilizes synchrony can be expected to produce other patterns of activity besides waves. Such patterns may play some role in cortical processing of information and may confer certain computational advantages [8].

Appendix A. To calculate the normal form for the bifurcation, we match the "ϵ" terms from (3.6):

$$\Omega_1 = \int_0^{2\pi} A_1(y')\ [\psi_1(x-y') - \psi_1(x)]\ dy' \equiv \mathcal{L}\ \psi_1.$$

We integrate both sides of the equation with respect to x to get $\Omega_1 = 0$. If we solve $\mathcal{L}\psi_1 = 0$, we get that $\psi_1(x) = ze^{ix} + \bar{z}e^{-ix}$ w.l.o.g. Next, we look at ϵ^2 terms:

$$\Omega_2 = \mathcal{L}\ \psi_2 + \int_0^{2\pi} A_2(y')\ [\psi_1(x-y') - \psi_1(x)]^2\ dy' + \frac{g_1\alpha_1}{2\pi} \int_0^{2\pi} [\psi_1(x-y') - \psi_1(x)]\ dy'. \tag{A.1}$$

Substituting ψ_1 into (A.1) and integrating with respect to x,

$$\int_0^{2\pi} \Omega_2 \, dx = 8\pi \; z \; \bar{z} \; [B_0(A_2) - B_1(A_2)],$$
$$\Omega_2 = 4z \; \bar{z} \; [B_0(A_2) - B_1(A_2)],$$

where $B_n(A_j) = \int_0^{2\pi} A_j(y') \; e^{\pm iny'} \; dy'$, $j = 1, 2, 3$. Now, we multiply (A.1) by e^{-ix},

$$0 = -2\pi \; g_1 \; \alpha_1 \; z,$$

which implies $g_1 = 0$. So, we can write $g_{syn} = g^*_{syn} + \epsilon^2 g_2$, and we solve for ψ_2:

$$0 = \mathcal{L} \; \psi_2 + (z^2 e^{2ix} + \bar{z}^2 e^{-2ix}) \; [B_2(A_2) + B_0(A_2) - 2B_1(A_2)]. \tag{A.2}$$

We now propose that $\psi_2 = C \; z^2 e^{2ix} + \bar{C} \; \bar{z}^2 e^{-2ix}$ and substitute back into (A.1) to get

$$0 = [B_2(A_1) - B_0(A_1)] \; (C \; z^2 e^{2ix} + \bar{C} \; \bar{z}^2 e^{-2ix})$$
$$+ [B(A_2) + B_0(A_2) - 2B_1(A_2)] \; (z^2 e^{2ix} + \bar{z}^2 e^{-2ix}).$$

Looking at the coefficients of the z^2 term gives

$$C = \frac{2B_1(A_2) - B_2(A_2) - B_0(A_2)}{B_2(A_1) - B_0(A_1)}. \tag{A.3}$$

We have to make sure here that the denominator is nonzero. This is easy to see, since $B_2(A_1) - B_0(A_1) = 0$ would imply that $g^*_{syn} = \frac{g_S \beta_1 (I_2 - 1)}{\alpha_2}$, which is not true since $g^*_{syn} = \frac{g_{gap} \beta_1 (I_1 - 1)}{\alpha_2}$ and $I_1 > I_2$.

Next, we look at ϵ^3 terms:

$$\Omega_3 = \mathcal{L} \; \psi_3 + \int_0^{2\pi} A_3(y') \; [\psi_1(x - y') - \psi_1(x)]^3 \; dy'$$
$$+ 2 \int_0^{2\pi} A_2(y') \; [\psi_1(x - y')\psi_2(x - y') + \psi_1(x)\psi_2(x)] \; dy'$$
$$- 2 \int_0^{2\pi} A_2(y') \; [\psi_1(x - y')\psi_2(x) + \psi_1(x)\psi_2(x - y')] \; dy'$$
$$+ \frac{g_2 \alpha_1}{2\pi} \int_0^{2\pi} [\psi_1(x - y') - \psi_1(x)] \; dy'. \tag{A.4}$$

Let us look at the terms in (A.4) closely:

$$[\psi_1(x - y') - \psi_1(x)]^3 = [z e^{ix} e^{-iy'} + \bar{z} e^{-ix} e^{iy'} - z e^{ix} - \bar{z} e^{-ix}]^3$$
$$= z^3 e^{3ix} e^{-3iy'} + 3z^2 \bar{z} e^{ix} e^{-iy'} + 3z \bar{z}^2 e^{-ix} e^{iy'} + \bar{z}^3 e^{-3ix} e^{3iy'}$$
$$- 3z^3 e^{3ix} e^{-2iy'} - 3z^2 \bar{z} e^{ix} e^{-2iy'} - 6z^2 \bar{z} e^{ix}$$
$$- 6z \bar{z}^2 e^{-ix} - 3z \bar{z}^2 e^{-ix} e^{2iy'} - 3\bar{z}^3 e^{-3ix} e^{2iy'}$$
$$+ 3z^3 e^{3ix} e^{iy'} + 3z^2 \bar{z} e^{ix} e^{iy'} + 6z^2 \bar{z} e^{ix} e^{-iy'}$$
$$+ 6z \bar{z}^2 e^{-ix} e^{iy'} + 3z \bar{z}^2 e^{-ix} e^{-iy'} + 3\bar{z}^3 e^{-3ix} e^{iy'}$$
$$- z^3 e^{3ix} - 3z^2 \bar{z} e^{ix} - 3z \bar{z}^2 e^{-ix} - \bar{z}^3 e^{-3ix}.$$

Let $T = [\psi_1(x-y')\psi_2(x-y') + \psi_1(x)\psi_2(x) - \psi_1(x-y')\psi_2(x) - \psi_1(x)\psi_2(x-y')]$; then

$$\begin{aligned}
T &= (ze^{ix}e^{-iy'} + \bar{z}e^{-ix}e^{iy'})(Cz^2e^{2ix}e^{-2iy'} + \bar{C}\bar{z}^2e^{-2ix}e^{2iy'}) \\
&\quad + (ze^{ix} + \bar{z}e^{-ix})(Cz^2e^{2ix} + \bar{C}\bar{z}^2e^{-2ix}) \\
&\quad - (ze^{ix}e^{-iy'} + \bar{z}e^{-ix}e^{iy'})(Cz^2e^{2ix} + \bar{C}\bar{z}^2e^{-2ix}) \\
&\quad - (ze^{ix} + \bar{z}e^{-ix})(Cz^2e^{2ix}e^{-2iy'} + \bar{C}\bar{z}^2e^{-2ix}e^{2iy'}) \\
&= Cz^3e^{3ix}e^{-3iy'} + \bar{C}z\bar{z}^2e^{-ix}e^{iy'} + Cz^2\bar{z}e^{ix}e^{-iy'} \\
&\quad + \bar{C}\bar{z}^3e^{-3ix}e^{3iy'} + Cz^3e^{3ix} + \bar{C}z\bar{z}^2e^{-ix} \\
&\quad + Cz^2\bar{z}e^{ix} + \bar{C}\bar{z}^3e^{-3ix} - Cz^3e^{3ix}e^{-iy'} - \bar{C}z\bar{z}^2e^{-ix}e^{-iy'} \\
&\quad - Cz^2\bar{z}e^{ix}e^{iy'} - \bar{C}\bar{z}^3e^{-3ix}e^{iy'} \\
&\quad - Cz^3e^{3ix}e^{-2iy'} - \bar{C}z\bar{z}^2e^{-ix}e^{2iy'} \\
&\quad - Cz^2\bar{z}e^{ix}e^{-2iy'} - \bar{C}\bar{z}^3e^{-3ix}e^{2iy'}.
\end{aligned}$$

Substituting ψ_1 and ψ_2 into (A.4) and using the expansions for $[\psi_1(x-y')-\psi_1(x)]^3$ and T, we then integrate with respect to x to get $\Omega_3 = 0$. Next, multiply both sides by e^{-ix} and integrate with respect to x to get

$$\begin{aligned}
0 &= z^2\bar{z}\int_0^{2\pi} A_3(y')\ [9e^{-iy'} + 3e^{iy'} - 3e^{-2iy'} - 9]\ dy' \\
&\quad + 2C\ z^2\bar{z}\int_0^{2\pi} A_2(y')\ [e^{-iy'} - e^{iy'} - e^{-2iy'} + 1]\ dy' - g_2\alpha_1 z.
\end{aligned}$$

We can simplify this as follows:

$$0 = z^2\ \bar{z}[12B_1(A_3) - 3B_2(A_3) - 9B_0(A_3) + 2C\ B_0(A_2) - 2CB_2(A_2)] - g_2\alpha_1 z.$$

By letting $\zeta = 12B_1(A_3) - 3B_2(A_3) - 9B_0(A_3) + 2C\ B_0(A_2) - 2CB_2(A_2)$ and $\eta = -g_2\alpha_1$, we have the normal form at the bifurcation point as

$$0 = \zeta z^2\bar{z} + \eta z.$$

Appendix B. We use the biophysical model given in [6]. Each uncoupled bursting cell in the Limax model has the form

$$\begin{aligned}
C\frac{dV}{dt} &= -I_L - I_K - I_{Ca} \\
&= -g_L(V - E_L) - g_K n^4(V - E_K) - g_{Ca}m^2h(V - E_{Ca}),
\end{aligned}$$

where n, h obey the equations

$$\begin{aligned}
\frac{dn}{dt} &= .075[a_n(V)(1-n) - b_n(V)n], \\
\frac{dh}{dt} &= \frac{1.125(h_\infty(V) - h)}{\tau_h(V)},
\end{aligned}$$

with

$$\begin{aligned}
a_n(V) &= .032(-48 - V)/(\exp(-(48+V)/5) - 1), \\
b_n(V) &= .5\exp(-(43+V)/40), \\
h_\infty(V) &= 1/(1 + \exp((V+86)/4)), \\
\tau_h(V) &= \begin{cases} \text{if } (V < (-80)), \text{then } (\exp((V+470)/66.6)), \\ \text{else } (28 + \exp((V+25)/-10.5)). \end{cases}
\end{aligned}$$

The activation gate for the T-type calcium current has the form

$$m(V) = 1/(1 + \exp(-(V + 60))).$$

The parameters are $C = 2.66$, $g_K = 5$, $g_L = 0.024$, $g_{Ca} = 2$, $E_K = -90$, $E_{Ca} = 140$, $E_L = -82$, and $E_{syn} = -78$. Using this model, we compute the approximations of the coupling functions as follows:

$$H_{syn}(x) = 35 + 200\cos(x) + 32\cos(2x) - 95\sin(x) - 5\sin(2x),$$

$$H_{gap}(x) = 87 - 50\cos(x) - 37\cos(2x) + 295\sin(x) - 65\sin(2x).$$

REFERENCES

[1] T. Bem, Y. LeFeuvre, J. Simmers, and P. Meyrand, *Electrical coupling can prevent expression of adult-like properties in an embryonic neural circuit*, J. Neurophysiol., 87 (2002), pp. 538–547.

[2] T. Bem and J. Rinzel, *Short duty cycle destabilizes a half-center oscillator, but gap junctions restabilize the anti-phase pattern*, J. Neurophysiol., 91 (2004), pp. 693–703.

[3] C. Bou-Flores and A. J. Berger, *Gap junctions and inhibitory synapses modulate inspiratory motoneuron synchronization*, Neurophysiol., 85 (2001), pp. 1543–1551.

[4] C. C. Chow and N. Kopell, *Dynamics of spiking neurons with electrical coupling*, Neural Comp., 1 (1994), pp. 313–321.

[5] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, Software Environ. Tools 14, SIAM, Philadelphia, 2002.

[6] B. Ermentrout, J. W. Wang, J. Flores, and A. Gelperin, *Model for transition from waves to synchrony in the olfactory lobe of Limax*, J. Comput. Neurosci., 17 (2004), pp. 365–383.

[7] G. B. Ermentrout, *Stable periodic solutions to discrete and continuum arrays of weakly coupled nonlinear oscillators*, SIAM J. Appl. Math., 52 (1992), pp. 1665–1687.

[8] G. B. Ermentrout and D. Kleinfeld, *Travelling electrical waves in cortex: Insight from phase dynamics and speculation on a computational role*, Neuron, 29 (2001), pp. 33–44.

[9] G. B. Ermentrout and N. Kopell, *Multiple pulse interactions and averaging in systems of coupled neural oscillators*, J. Math. Biol., 29 (1991), pp. 195–217.

[10] G. B. Ermentrout and N. Kopell, *Inhibition-produced patterning in chains of coupled nonlinear oscillators*, SIAM J. Appl. Math., 54 (1994), pp. 478–507.

[11] M. Galarreta and S. Hestrin, *A network of fast-spiking cells in the neocortex connected by electrical synapses*, Nature, 402 (1999), pp. 72–75.

[12] N. Kopell, *Toward a theory of modeling central pattern generators*, in Neural Control of Rhythmic Movements in Vertebrates, A. Cohen, ed., John Wiley, New York, 1988, pp. 396–413.

[13] N. Kopell and B. Ermentrout, *Chemical and electrical synapses perform complementary roles in the synchronization of interneuronal networks*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 15482–15487.

[14] N. Kopell and G. B. Ermentrout, *Symmetry and phaselocking in chains of weakly coupled oscillators*, Comm. Pure Appl. Math., 39 (1986), pp. 623–660.

[15] Y. Kuramoto, *Chemical Oscillations, Waves and Turbulence*, Springer, New York, 1984.

[16] T. J. Lewis and J. Rinzel, *Dynamics of spiking neurons connected by both inhibitory and electrical coupling*, J. Comput. Neurosci., 14 (2003), pp. 283–309.

[17] B. Pfeuty, G. Mato, D. Golomb, and D. Hansel, *The combined effects of inhibitory and electrical synapses in synchrony*, Neural Comp., 17 (2005), pp. 633–670.

[18] R. D. Traub, *Model of synchronized population bursts in electrically coupled interneurons containing active dendrites*, J. Comput. Neurosci., 2 (1995), pp. 283–289.

[19] C. van Vreeswijk, L. F. Abbott, and G. B. Ermentrout, *When inhibition not excitation synchronizes neural firing*, J. Comput. Neurosci., 1 (1994), pp. 313–321.

[20] M. A. Whittington, R. D. Traub, N. Kopell, B. Ermentrout, and E. H. Buhl, *Inhibition based rhythms: Experimental and mathematical observations on network dynamics*, Int. J. Psychophysiol., 38 (2000), pp. 315–336.

[21] A. T. Winfree, *Biological rhythms and the behavior of populations of coupled oscillators*, J. Theor. Biol., 16 (1967), pp. 15–42.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 530–542

THE EFFECT OF NOISE ON β-CELL BURST PERIOD*

MORTEN GRAM PEDERSEN† AND MADS PETER SØRENSEN†

Abstract. Bursting electrical behavior is commonly observed in a variety of nerve and endocrine cells, including that in electrically coupled β-cells located in intact pancreatic islets. However, individual β-cells usually display either spiking or very fast bursting behavior, and the difference between isolated and coupled cells has been suggested to be due to stochastic fluctuations of the plasma membrane ion channels, which are supposed to have a stronger effect on single cells than on cells situated in clusters (the channel sharing hypothesis). This effect of noise has previously been studied using numerical simulations. We show here how the application of two recent methods allows an analytic treatment of the stochastic effects on the location of the saddle-node and homoclinic bifurcations, which determine the burst period. Thus, the stochastic system can be analyzed similarly to the deterministic system, but with a quantitative description of the effect of noise. This approach supports previous investigations of the channel sharing hypothesis.

Key words. bursting oscillations, stochastic Melnikov method, stochastic bifurcations

AMS subject classifications. 37H, 34F05, 60H10, 60H30, 92C

DOI. 10.1137/060655663

1. Introduction.

The pancreatic β-cells are crucial for maintaining blood sugar levels in a narrow range. When subjected to glucose the β-cells produce and secrete insulin, and the amount of secreted insulin correlates with intracellular calcium levels [10].

In situ the β-cells are electrically coupled in the islets of Langerhans, where they show bursting electrical activity with burst periods of tens of seconds. Bursting consists of the membrane potential alternating between a silent hyperpolarized phase and an active phase of spiking rising from a depolarized plateau. During the active phase, calcium enters the cells, raises the intracellular Ca^{2+} concentration, and triggers insulin secretion. The plateau fraction, i.e., the ratio of the active phase duration to the burst period, is decisive for intracellular Ca^{2+} concentrations and for the amount of secreted insulin [2].

However, early recordings of single isolated pancreatic β-cells showed that the membrane potential exhibits noisy spiking activity [19], and although it was later found that only approximately one third of isolated cells spike, while half of the single cells are fast bursters with burst period less than 5 seconds [11], there is a fundamental difference in the behavior of single and electrically coupled cells. Importantly, this difference is reflected in intracellular calcium levels [24].

It was suggested early that stochastic fluctuations of ion channels in the plasma membrane were responsible for disrupting the bursting behavior and transforming the isolated cells to spikers, but that the effective sharing of the channels by electrically coupled cells averages the noise and lets the bursting phenomena appear [3]. This was analyzed by Chay and Kang [4] and Sherman, Rinzel, and Keizer [21] using mathematical modeling. The burst period and plateau fraction in the deterministic version

*Received by the editors March 30, 2006; accepted for publication (in revised form) November 28, 2006; published electronically February 9, 2007. This work was supported by the European Union through the Network of Excellence BioSim, contract LSHB-CT-2004-005137.

http://www.siam.org/journals/siap/67-2/65566.html

†Department of Mathematics, Technical University of Denmark, Matematiktorvet Building 303, 2800 Kgs. Lyngby, Denmark (m.g.pedersen@mat.dtu.dk, m.p.soerensen@mat.dtu.dk).

of the Sherman–Rinzel–Keizer model was later analyzed by bifurcation analysis and Melnikov's method [16].

De Vries and Sherman [6] studied the electrical behavior of coupled pancreatic β-cells with focus on the *beneficial* influence of noise. It had previously been shown that weak coupling between identical spiking cells can induce bursting [20], and it is now known that heterogeneous but spiking cells start to burst when coupled with physiologically realistic coupling strengths [7]. The main result presented in [6] is that noise dramatically increases the interval of coupling strengths for which bursting is seen for identical cells, and this observation was supported by analyzing a bifurcation diagram. It was later shown that the beneficial influence is more likely through heterogeneity masquerading as noise, and that the explanation of the enhancement of emergent bursting must be modified accordingly [14].

The investigations of the effect of noise on β-cells have so far been done partly by numerically solving the stochastic differential equations (SDEs) describing the system and partly by analyzing *deterministic* bifurcation diagrams [1, 4, 6, 14, 21]. The transition from the SDEs to the bifurcation analysis was rather weakly motivated from a theoretical point of view.

We look for a more natural deterministic description of the stochastic system, with the aim of characterizing how noise shortens or interrupts bursting. This is based on the ideas from Pernarowski, Miura, and Kevorkian [16] using a stochastic version of a polynomial minimal model [15].

For the transition from the silent to the active phase, we consider the distribution of the solution over time; i.e., we follow the probability that the system is in a certain area of state space over time. The time evolution of the distribution is described by the Fokker–Planck equation (FPE), which is a partial differential equation. Since bifurcation analysis is better performed on a system of ordinary differential equations (ODEs), and the FPE is computationally expensive to solve, we assume that the distribution solving the FPE, and hence describing the system, is Gaussian at any point in time. Doing this, we obtain a set of ODEs describing how the distribution evolves in time. This approach is based on work on models of noisy spiking neurons [18, 22], and the ODEs describe the evolution of the mean and lower order moments of the assumed Gaussian distribution. A similar approach [12, 13] assumed a Gaussian-like distribution around the deterministic solution, and was used to describe a neural burster [13]. For the transition out of the active phase we use a stochastic Melnikov method [9], thus allowing us to use the ideas from [16] in a stochastic setting.

We find that noise makes both the active and the silent phases terminate earlier than for the deterministic model, but that it has a stronger effect on the exit from the active phase than from the silent phase. Thus, we explain why simulations show that noise shortens both phases and consequently the burst period, in this way transforming normal bursters into fast bursters. This supports the idea that stochastic fluctuations in membrane ion channels can disrupt normal bursting and that channel sharing can restore it [3, 4, 21].

2. The β-cell model with noise. Pernarowski [15] introduced a minimal, deterministic, polynomial model capable of modeling both the spiking and the bursting phenomena seen in β-cells. The fact that the involved functions are polynomials will be of importance when describing the moments of the distribution [22]. The model is

$$\frac{du}{dt} = f(u) - w - z, \tag{2.1a}$$

$$\frac{d\,w}{dt} = g(u) - w + \sigma\Gamma_t, \tag{2.1b}$$

$$\frac{d\,z}{dt} = \epsilon(h(u) - z), \tag{2.1c}$$

where we have added the white noise term Γ_t to include noise, the strength of which is given by σ. f and g are third order polynomials, while h is a first order polynomial. u mimics the membrane potential of the cell, while w is a fast gating variable. We assume that the ion-channel controlled by w is fluctuating stochastically, and hence we add the noise term to this equation. z is, on the other hand, a slow gating variable due to the small number ϵ. Thus, we have a fast subsystem (u, w) responsible for the spikes during an active phase of bursting, and a slow z controlling the transition between the silent and active phases.

Following [15], we differentiate (2.1a) with respect to t and then transform system (2.1) to

$$\frac{d^2u}{dt^2} + F(u)\frac{d\,u}{dt} + G(u) + z = -\epsilon(h(u) - z) - \sigma\Gamma_t, \tag{2.2a}$$

$$\frac{d\,z}{dt} = \epsilon(h(u) - z), \tag{2.2b}$$

or, equivalently,

$$\frac{d\,u}{dt} = y, \tag{2.3a}$$

$$\frac{d\,y}{dt} = -F(u)y - G(u) - z - \epsilon(h(u) - z) - \sigma\Gamma_t, \tag{2.3b}$$

$$\frac{d\,z}{dt} = \epsilon(h(u) - z), \tag{2.3c}$$

where

$$F(u) = a\big((u - \hat{u})^2 - \eta^2\big), \tag{2.4}$$

$$G(u) = u^3 - 3(u + 1), \tag{2.5}$$

$$h(u) = \beta(u - u_\beta). \tag{2.6}$$

With appropriate parameters, the system shows a bursting pattern, but increasing the strength of the noise shortens the bursts; see Figure 1, left panels.

This simulation as well as all other simulations and bifurcation diagrams were done using XPPAUT [8]. The stochastic equations were solved by the backward Euler method with time step $dt = 0.005$. For each time step XPPAUT draws a random number from an appropriately scaled normal distribution to simulate the Wiener process. Control simulations showed that the use of smaller time steps did not change the results.

The deterministic system ($\sigma = 0$) can be analyzed from a bifurcation diagram of the fast subsystem with z as the bifurcation parameter [15, 17]. This is done by setting $\epsilon = 0$. The fixed points of the fast system fall on the Z-shaped curve $z = -G(u)$; see Figure 2. The fast system is stable for low z values, but upon increasing z, this stability is lost in a Hopf-bifurcation (HB in Figure 2). The fixed points on the middle branch of the Z-shaped curve are saddle-points, while they are stable on the lower branch. The middle branch meets the upper and lower branch in saddle-node bifurcations (SN in Figure 2). The Hopf-bifurcation gives rise to stable periodic solutions around the

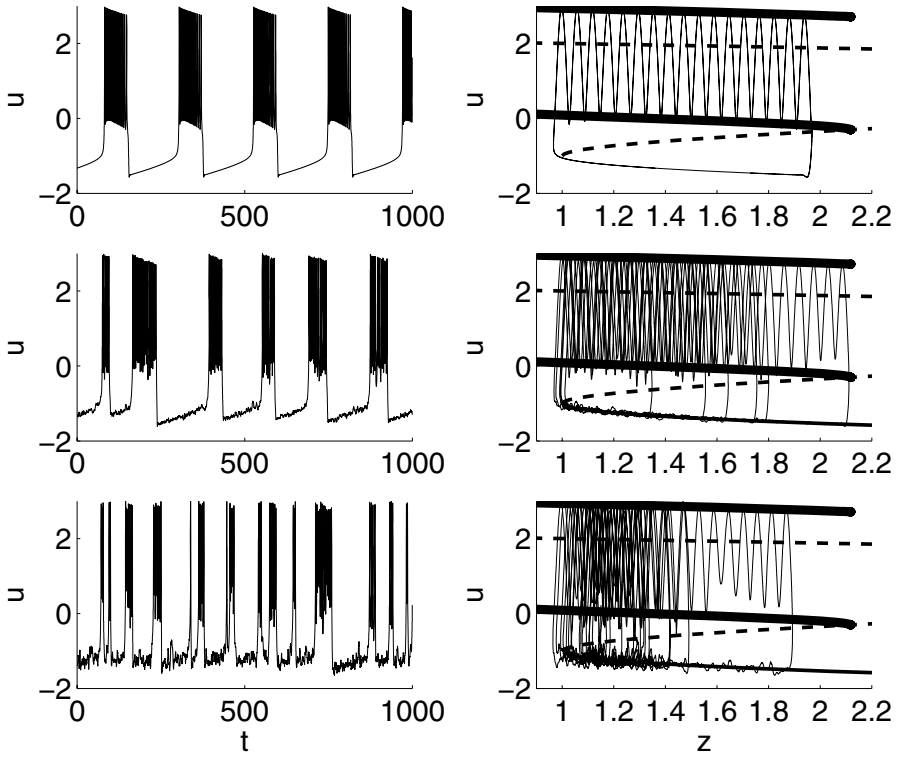


FIG. 1. *Numerical simulations of bursting with different noise strengths. The left panels show time series of the membrane potential u, while the right panels show the corresponding projection on the z-u plane and the deterministic bifurcation diagram from Figure* 2. *The upper panels show the deterministic case* $\sigma = 0$, *in the center panels* $\sigma = 0.1$, *and in the lower panels* $\sigma = 0.3$. *Other parameters are, here and throughout the manuscript,* $a = 0.25$, $\hat{u} = 1.6$, $\beta = 4$, $u_\beta = -0.954$, $\epsilon = 0.0025$, *and* $\eta = 0.7$.

unstable fixed points on the upper branch, but these periodic solutions disappear in a homoclinic bifurcation (HC in Figure 2) for sufficiently large z. The mechanism underlying bursting is based on the bistability between the stable fixed points on the lower branch and the stable periodic solutions for a range of z-values. When we reintroduce the slow variation of z for $0 < \epsilon \ll 1$, we can explain bursting. When the solution of the system is near the lower branch, trajectories move slowly to the left since u is low, and thus $\frac{dz}{dt} < 0$ here. This continues until the stable branch disappears in the left saddle-node bifurcation. The solution now leaves the lower branch (silent phase) and goes to the stable periodic solutions (active phase), where u is high and $\frac{dz}{dt} > 0$. Hence, the trajectory now moves to the right until it meets the homoclinic bifurcation and the stable periodic solutions disappear. The solution then leaves the active phase and settles on the lower branch, and the scenario is repeated.

This explanation gives a hint of how noise shortens the bursts. The random perturbations to the system can make trajectories leave the silent as well as the active phase prematurely when the system is randomly kicked across the corresponding thresholds. When the noise intensity increases, this will happen more often since the stochastic fluctuations are larger. In Figure 1 (right panels), we see that in general the noisy system leaves the active phase prematurely, while the early exit from the silent phase is less pronounced. We now aim at understanding this observation better.

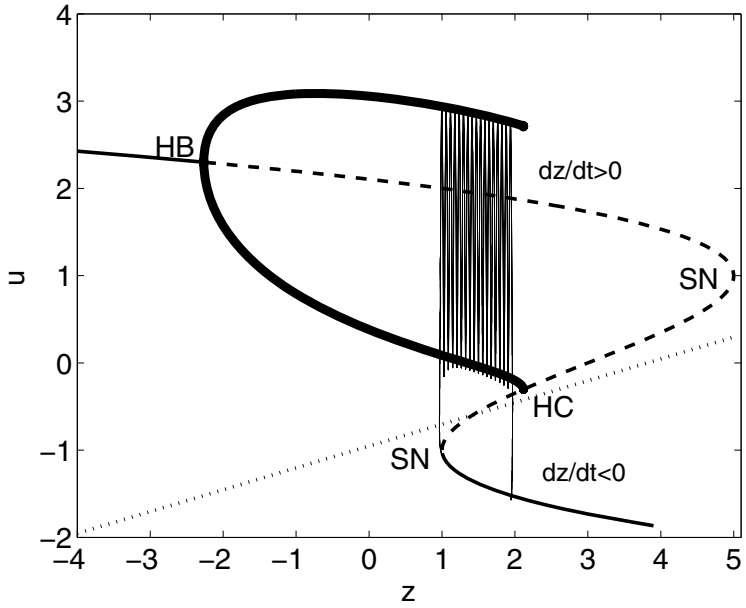


FIG. 2. *Bifurcation diagram of the fast subsystem with z as the bifurcation parameter. Thin solid curves indicate stable fixed points, thin broken curves correspond to unstable fixed points, and the thick solid curve shows the extrema of periodic solutions. The dotted curve shows the z-nullcline, $\frac{dz}{dt} = 0$. A simulation of the deterministic system is projected onto the z-u plane for comparison. See the text for more details.*

3. Location of the left saddle-node bifurcation. The exit from the silent phase happens near the left saddle-node bifurcation; see Figure 2. We expect that for increasing noise, the bifurcation will effectively happen for larger z-values, since the noise will tend to push the system across the threshold and into the active phase prematurely.

To analyze this, we look at the distribution of the system under all possible realizations of the noise. Since the fluctuations of u and y are rather small during the silent phase, we expect that for fixed z ($\epsilon = 0$) the distribution of (u, y) will be approximately Gaussian. This allows us to use the so-called G-method [22], which is a development of the method from [18].

The idea is that a Gaussian distribution is described completely by its mean and covariance matrix. Hence we follow, for fixed z, the means $\bar{u} = \langle u \rangle$ and $\bar{y} = \langle y \rangle$, the variances $S_u = Var(u)$, and $S_v = Var(y)$, and the covariance $C = Cov(u, y)$.

Following [22], we average (2.3a), (2.3b), and the time derivative of the (co)variances $(u - \bar{u})^2$, $(y - \bar{y})^2$, and $(u - \bar{u})(y - \bar{y})$ using Itô's formula and the fact that the odd moments vanish for a Gaussian distribution. To illustrate the procedure, we derive the equation for C in greater detail, as follows:

$$\begin{aligned}
\frac{dC}{dt} &= \frac{d}{dt}\Big\langle (u-\bar{u})(y-\bar{y})\Big\rangle = \Big\langle \frac{d}{dt}[(u-\bar{u})(y-\bar{y})]\Big\rangle \\
&= \Big\langle (u-\bar{u})\frac{d\,(y-\bar{y})}{dt}\Big\rangle + \Big\langle (y-\bar{y})\frac{d\,(u-\bar{u})}{dt}\Big\rangle \\
&= \Big\langle (u-\bar{u})(-F(u)y - G(u) - z - \sigma\Gamma_t)\Big\rangle + \Big\langle (y-\bar{y})^2\Big\rangle \\
&= -\Big\langle (u-\bar{u})F(u)y\Big\rangle - \Big\langle (u-\bar{u})G(u)\Big\rangle + S_y.
\end{aligned} \tag{3.1}$$

The first term is found from the Taylor polynomial of F around $\bar{u}$,

$$\begin{aligned}
&\Big\langle (u-\bar{u})F(u)y \Big\rangle \\
&= \left\langle (u-\bar{u})\big((y-\bar{y})+\bar{y}\big)\left(F(\bar{u})+F'(\bar{u})(u-\bar{u})+\frac{1}{2}F''(\bar{u})(u-\bar{u})^2\right)\right\rangle \\
&= F(\bar{u})C + F'(\bar{u})\bar{y}S_u + a\Big\langle (u-\bar{u})^3(y-\bar{y})\Big\rangle,
\end{aligned} \tag{3.2}$$

where we have again used that the odd moments vanish. Finally, the last term of (3.2) is equal to $3aS_uC$ by the Gaussian joint variable theorem. The second term of (3.1) is treated similarly.

In summary, we obtain the equations

$$\frac{d\,\bar{u}}{dt} = \bar{y}, \tag{3.3a}$$

$$\frac{d\,\bar{y}}{dt} = -F(\bar{u})\bar{y} - G(\bar{u}) - z - \big(F''(\bar{u})\bar{y} + G''(\bar{u})\big)\frac{S_u}{2} - F'(\bar{u})C, \tag{3.3b}$$

$$\frac{d\,S_u}{dt} = 2\ C, \tag{3.3c}$$

$$\frac{d\,S_y}{dt} = 2\ \big[-F(\bar{u})S_y - \big(F'(\bar{u})\bar{y} + G'(\bar{u})\big)\,C\,\big] + \sigma^2 - 6aS_u^2, \tag{3.3d}$$

$$\frac{d\,C}{dt} = S_y - \big(F'(\bar{u})\bar{y} + G'(\bar{u})\big)S_u - F(\bar{u})C - 3aS_uC. \tag{3.3e}$$

These are exact equations for the means and (co)variances due to F and G being polynomials [22]. Since the system (3.3) is deterministic, we can perform bifurcation analysis on these equations using z as the bifurcation parameter. Starting from the silent phase $\bar{u} \approx -1$, $\bar{y} = S_u = S_y = C = 0$, we find a branch of stable fixed points similar to the lower branch of Figure 2 (not shown). This branch ends in a saddle-node bifurcation, as for the deterministic case. However, the rest of the bifurcation structure breaks down, and the system (3.3) has, e.g., fixed points with negative S_y values, which are of course impossible solutions, since S_y is a variance. We believe that this breakdown is because the assumption of a Gaussian distribution holds only in the silent phase, and hence the system (2.3) is no longer described by system (3.3) after leaving the lower branch. Nevertheless, the saddle-node where the silent phase branch of (3.3) ends can be followed in a two-parameter bifurcation diagram with σ as the other bifurcation parameter; see Figure 3. For increasing noise intensity σ the saddle-node moves to the right, indicating that the noisy system leaves the silent phase earlier for greater noise strength. This corresponds well to direct simulations of the z-value for which the noisy system (2.3) leaves the silent phase (Figure 3).

4. Location of the homoclinic bifurcation. To follow the exit from the active phase for different noise intensities, we apply a stochastic Melnikov method. The deterministic Melnikov technique was first applied to β-cell models by Pernarowski, Miura, and Kevorkian [16] and for the model we use here in [15].

The Melnikov function is used to determine the distance between the stable and unstable manifolds of a saddle-point for systems, which are small perturbations of a Hamiltonian system with a homoclinic saddle-point. In the deterministic case of the β-cell model, the active phase ends in a homoclinic bifurcation, which happens exactly when the stable and unstable manifolds of the saddle-point coincide, i.e., when the Melnikov function is zero.

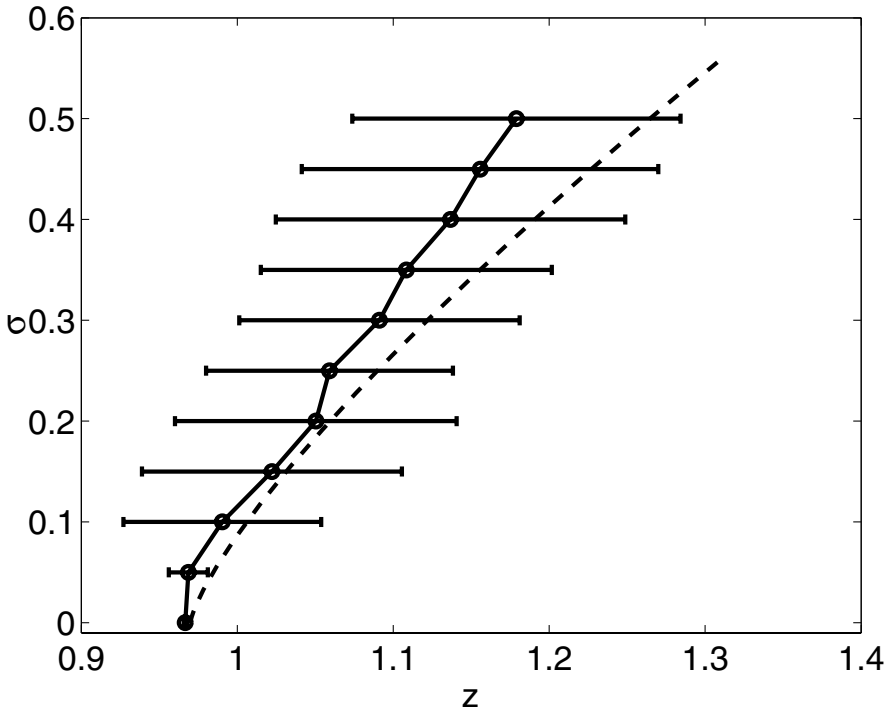


FIG. 3. *Two-parameter bifurcation diagram showing the location of the saddle-node bifurcation where the silent phase branch for system* (3.3) *ends (broken curve). This bifurcation corresponds to the left saddle-node bifurcation in Figure* 2. *Direct simulations of the noisy system* (2.2) *show that the prediction from* (3.3) *is faithful, since the z-values for which the system* (2.2) *leaves the silent phase (measured as u passing through the Poincaré section* $u = -0.55$ *from below) agree well. The bars are mean values of* $z \pm$ *one standard deviation for a simulation until* $t = 10000$. *We have shifted the broken curve* 0.03 *to the left, since the deterministic version overestimates the z-value by this amount.*

We write (2.3) with $\epsilon = 0$ as

$$\frac{d\,u}{dt} = y, \tag{4.1}$$

$$\frac{d\,y}{dt} = -G(u) - z \; + \; [-F(u)y - \sigma\Gamma_t], \tag{4.2}$$

from which it is seen that the term in the square brackets is a perturbation of the Hamiltonian system $\frac{d^2u}{dt^2} + G(u) + z = 0$, which has a saddle-point $(a_s(z), 0)$ with a homoclinic orbit (u_s, y_s) [5, 15]. The Hamiltonian is $H(p,u) = \frac{1}{2}p^2 + V(u; a_s(z))$ with potential

$$V(u, a_s) = \frac{1}{4}(u - a_s)^2[u^2 + 2a_s u + 3a_s^3 - 6]. \tag{4.3}$$

The homoclinic orbit can then be written as

$$(u_s, y_s) = (u_s, \pm\sqrt{-2V(u_s, a_s(z))}\;). \tag{4.4}$$

For the deterministic case, $\sigma = 0$, the square bracket in (4.2) reduces to $-F(u)y$. Pernarowski [15] showed that this term is indeed small for all z values between the

left saddle-node bifurcation and the homoclinic bifurcation, and Melnikov's method is therefore applicable. The Melnikov function is in the deterministic case [5, 15]

$$M_{det} = -a\left[e_2(a_s(z))(\hat{u}^2 - \eta^2) + e_1(a_s(z))\hat{u} + e_0(a_s(z))\right], \tag{4.5}$$

where

$$e_0(a_s) = -\frac{12}{5}\sqrt{3}(a_s^4 - 2a_s^2 - 4)\sqrt{1 - a_s^2} + 6\sqrt{2}a_s(a_s^2 - 3)\Delta(a_s), \tag{4.6}$$

$$e_1(a_s) = 6\sqrt{3}a_s(3 - a_s^2)\sqrt{1 - a_s^2} + 3\sqrt{2}(a_s^2 - 3)(a_s^2 + 1)\Delta(a_s), \tag{4.7}$$

$$e_2(a_s) = 4\sqrt{3}\sqrt{1 - a_s^2} + 2\sqrt{2}a_s(a_s^2 - 3)\Delta(a_s), \tag{4.8}$$

$$\Delta(a_s) = \cos^{-1}(2a_s/\sqrt{6 - 2a_s^2}). \tag{4.9}$$

We have changed the sign of M_{det} compared to [15] such that $M_{det} < 0$ when the stable manifold is outside the unstable manifold of the saddle-point [23]. In this case the fast subsystem has a limit cycle for a given fixed z; see Figure 2. When $M_{det} = 0$ the unstable and stable manifolds coincide and form a homoclinic orbit. This happens at the z-value when the homoclinic bifurcation occurs and the active phase terminates.

The Melnikov function is related to the phase space flux, which is a measure of the transport across the pseudoseparatix approximating the separatix of the Hamiltonian system [23]. For the β-cell model, we are interested in the transport from the inside to the outside of the pseudoseparatix, since this will terminate the active phase. The flux is given by the area of the *turnstile lobe* [23], and to first order it is found as

$$\phi_{det} \approx \int_{t_1}^{t_2} M_{det}^{+} dt = (t_2 - t_1)M_{det}^{+}, \tag{4.10}$$

since M_{det} does not depend on t. Here and in the following, $M^+ = \max\{0, M\}$ is the positive part of M, and t_1 and t_2 are the time points that define the lobe. Note that as long as $M_{det} < 0$, i.e., the unstable manifold lies inside the stable manifold, $\phi_{det} = 0$, indicating that there is no transport (flux) from inside to outside the separatix; i.e., the system is trapped. For the β-cell model the trajectories will follow the limit cycle characterizing the active phase.

Another related variable is the average phase space flux. To first order it is approximated by the flux factor given by [9]:

$$\Phi_{det} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} M_{det}^{+} dt = M_{det}^{+}. \tag{4.11}$$

We now move to the stochastic case with $1 \gg \sigma > 0$ in (4.2). Frey and Simiu [9] approximated the noise process by a harmonic sum with random parameters, so-called Shinozuka noise. Since each realization of the noise—each path—is a harmonic sum, they argued that Melnikov theory can be applied to such scenarios as the one treated here when σ is sufficiently small, even for perturbations containing white noise processes as in (4.2). We note that the deterministic term $-F(u)y$ in the square bracket in (4.2) is still small for all values of z between the left saddle-node bifurcation and the homoclinic bifurcation (not shown) as for the deterministic case.

This approach uses, instead of the deterministic Melnikov function, a stochastic Melnikov process [9], which is given by

$$M_{stoch}(t) = M_{det} + \sigma\Xi_t, \tag{4.12}$$

where Ξ_t is a stochastic process with Gaussian distribution in the case of white noise perturbations Γ_t, which is the case considered here for the β-cell model. Ξ_t has mean zero and variance

$$\sigma_\Xi^2 = \int_0^\infty |H(k)|^2 dk, \tag{4.13}$$

where $H(k) = \int_{\mathbb{R}} h(t)e^{-ikt}dt$ is the Fourier transform of $h(t) = y_s(-t)$. For the Hamiltonian system (4.2), y_s is odd, and hence also h and H are odd. By Parseval's equation and (4.4) we then get

$$\begin{aligned}\sigma_\Xi^2 = \frac{1}{2}\int_{\mathbb{R}} |H(k)|^2 dk &= \frac{1}{2}\, 2\pi \int_{\mathbb{R}} |h(t)|^2 dt \\ &= 2\pi \int_0^\infty |y_s(t)|^2 dt = 2\pi \int_{a_s}^{b_s} \sqrt{-2V(u, a_s(z))}du,\end{aligned} \tag{4.14}$$

where $b_s = -a_s + \sqrt{6 - 2a_s^2}$ is the largest zero of V, corresponding to the point $(b_s, 0)$ on the separatix (u_s, y_s) furthest from the saddle-point [5]. Using standard tables, we get the following expression from (4.3):

$$\sigma_\Xi^2 = \sqrt{2}\pi \left(2\sqrt{6 - 6a_s^2} + a_s(a_s^2 - 3) \left(\pi - 2\sin^{-1} \frac{2a_s}{\sqrt{6 - 2a_s^2}} \right) \right). \tag{4.15}$$

The saddle-point $a_s = a_s(z)$ can be determined analytically. Thus, we have a complete description of the Melnikov process (4.12).

The unstable and stable manifolds of the saddle-point intersect when $M_{stoch}(t_0)$ changes sign and becomes positive, and then the system can escape from the inside of the pseudoseparatix. Hence, it is reasonable to assume that the probability of terminating the active phase at t_0 is proportional to the probability $Pr(M_{stoch}(t_0) > 0)$.

Having characterized the Melnikov process M_{stoch}, we now take a closer look at $Pr(M_{stoch}(t_0) > 0)$. We define $X = \frac{M_{stoch}(t_0) - M_{det}}{\sigma\sigma_\Xi}$, so that the probability of ending the active phase at t_0 is proportional to $Pr(X > \frac{-M_{det}}{\sigma\sigma_\Xi})$ due to the above assumption. Note that $X \sim N(0, 1)$ is a standard Gaussian variable.

It seems plausible that we need at least a certain probability $Pr(X > \frac{-M_{det}}{\sigma\sigma_\Xi}) = \alpha$ in order to effectively end the active phase during a spike period, and there seems to be no a priori reason why this probability should depend on σ. Note that for increasing σ and $M_{det} < 0$, the probability $Pr(X > \frac{-M_{det}}{\sigma\sigma_\Xi})$ increases (for fixed z, and hence M_{det} and σ_Ξ), such that there is a higher probability of ending the active phase prematurely for higher noise strengths, as expected (see Figure 1). In the deterministic limit $\sigma \to 0$, this probability is either 0 (for $M_{det} < 0$) or 1 (for $M_{det} > 0$) in accordance with the above observations for the deterministic scenario.

Continuing this idea, we look for the z value for which the active phase terminates. This is a stochastic event, but on average we expect it to be closely related to the probability discussed above. Since the relation $Pr(X > \frac{-M_{det}}{\sigma\sigma_\Xi}) = \alpha$ determines a fixed $\frac{-M_{det}}{\sigma\sigma_\Xi} = \mu$, we get that for larger σ a larger value of $-M_{det}/\sigma_\Xi$ will be needed for the system to effectively leave the active phase. $-M_{det}/\sigma_\Xi$ is a decreasing function of z (Figure 4), so solving $-M_{det}/\sigma_\Xi = \sigma\mu$ yields a lower z when σ is large than for small σ. This corresponds to the fact that the escape from the active phase will happen earlier for higher noise intensities.

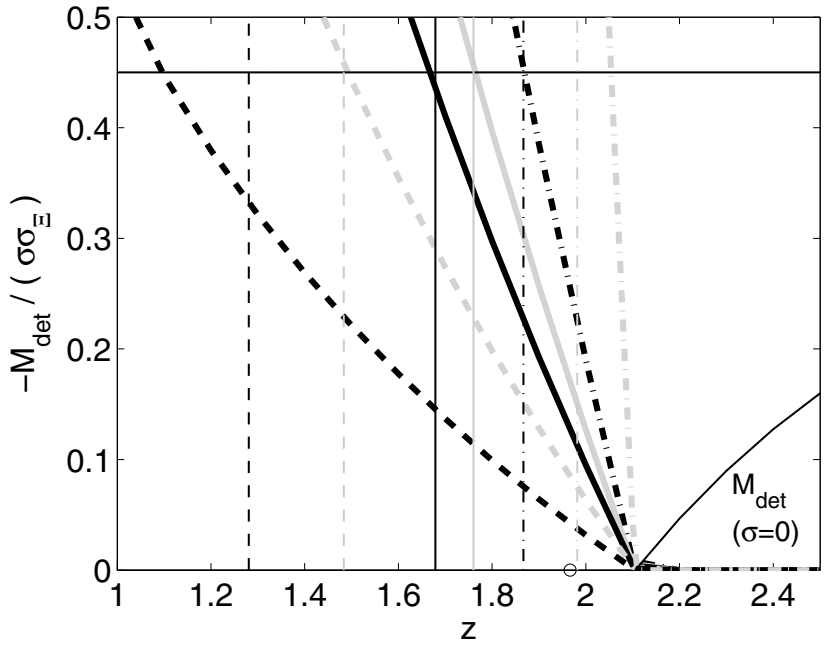


FIG. 4. *Stochastic escape from the active phase is explained by the condition* $Pr(X > \frac{-M_{det}}{\sigma\sigma_\Xi}) \approx 0.32$ *for* $X \sim N(0,1)$. *This corresponds to* $\frac{-M_{det}}{\sigma\sigma_\Xi} \approx 0.45$ *(horizontal black line). The decreasing curves are* $\frac{-M_{det}}{\sigma\sigma_\Xi}$ *calculated for different z-values using* (4.5) *and* (4.15) *for different values of* σ*:* 0.3 *(dashed black curve),* 0.15 *(dashed grey curve),* 0.1 *(solid black curve),* 0.075 *(solid grey curve),* 0.05 *(dash-dotted black curve),* 0.01 *(dash-dotted grey curve). Each of the thin vertical lines indicates for a value of* σ *(same* σ *values and corresponding line types and colors as above) the mean value of a series of z values for which the system left the active phase, defined as passing from above to below* $u = -0.8$*, in a simulation of system* (2.3) *until* $t = 40000$. *For comparison, the full increasing curve is the deterministic* M_{det}. *Note that* M_{det} *does not pass through zero at the z value (*$\circ$*) for which the deterministic system leaves the active phase. This mismatch between the homoclinic bifurcation and the simulated escape from the active phase is also seen in Figures* 1 *and* 2.

These considerations are supported by numerical simulations, which also confirm that the end of the active phase on average happens for a fixed value of $\frac{-M_{det}}{\sigma\sigma_\Xi}$, which is found empirically to be ≈ 0.45 (Figure 4), corresponding to $Pr(X > \frac{-M_{det}}{\sigma\sigma_\Xi}) \approx Pr(X > 0.45) \approx 0.32$. However, for very low or high σ this is not true. For high σ, the reason is that the system enters the active phase at $z > 1$, e.g., $z \approx 1.1$ for $\sigma = 0.3$ (Figure 3). Hence, there is a lower limit on the z value for which the escape can occur, and thus, the average value will be higher than predicted by the considerations above. For low σ we are near the case where $M_{det} = 0$, and hence the considerations above might break down in this deterministic limit, especially considering the probability considerations. For example, the period of the limit cycle in the fast subsystem, corresponding to the spike period of the full system, increases when z approaches the value where the homoclinic bifurcation occurs. This implies that even if the probability $Pr(M_{stoch}(t_0) > 0) = Pr(X > \frac{-M_{det}}{\sigma\sigma_\Xi})$ is smaller at each time point t_0, the probability of leaving the active phase during a spike, $\int_0^T Pr(M_{stoch}(t_0) > 0)dt_0 = T \cdot Pr(X > \frac{-M_{det}}{\sigma\sigma_\Xi})$, can still be large. Thus, the larger spike period compensates for

the lower instantaneous probability, such that the active phase terminates earlier than predicted from $Pr(M_{stoch}(t_0) > 0)$ alone. Moreover, the Melnikov approach predicts a value of z, for which the system leaves the active phase, that is too large even for the deterministic case. This imprecision could be more important for low noise strengths.

The instantaneous flux

$$\phi_{stoch} \approx M^+_{stoch} = (M_{det} + \sigma\Xi_t)^+ \tag{4.16}$$

is at every time t a truncated normal distribution. It has mean equal to the (nonrandom) flux factor Φ_{stoch} given by [9]

$$\Phi_{stoch} = M_{det} + \sigma\sigma_\Xi f(-M_{det}/(\sigma\sigma_\Xi)) - M_{det}F(-M_{det}/(\sigma\sigma_\Xi)), \tag{4.17}$$

where $f = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$ is the standard Gaussian density and $F(z) = \int_\infty^z f(x)dx$ is the corresponding distribution function. We remark that $\Phi_{stoch} \to M^+_{det} = \Phi_{det}$ for $\sigma \to 0$.

However, it is the flux during a finite interval that is relevant for the transition out of the active phase. This finite time flux varies randomly, and hence the mean flux Φ_{stoch} needs to be held against the variance of ϕ_{stoch} in order to determine when the system escapes from the active phase, in the same spirit as for M_{stoch} above. The variance of ϕ_{stoch} is given by

$$\begin{aligned} Var\left(\phi_{stoch}\right) = (\sigma\sigma_\Xi)^2 \Bigg[& \left(1 + \left(\frac{M_{det}}{\sigma\sigma_\Xi}\right)^2\right)\left(1 - F\left(-\frac{M_{det}}{\sigma\sigma_\Xi}\right)\right) + \frac{M_{det}}{\sigma\sigma_\Xi} f\left(-\frac{M_{det}}{\sigma\sigma_\Xi}\right) \\ & - \left(f\left(-\frac{M_{det}}{\sigma\sigma_\Xi}\right) + \frac{M_{det}}{\sigma\sigma_\Xi}\left(1 - F\left(-\frac{M_{det}}{\sigma\sigma_\Xi}\right)\right)\right)^2 \Bigg], \end{aligned} \tag{4.18}$$

and it is readily seen that $Var(\phi_{stoch}) \to 0$ for $\sigma \to 0$.

From simulations, it again appears that the end of the active phase happens for a roughly constant value of $\Phi_{stoch}/\sqrt{Var(\phi_{stoch})} \approx 0.182$ (Figure 5). Thus, as seen above for the Melnikov process, the related approach using phase space flux predicts that the exit from the active phase occurs for a fixed value of the standardized variable $\frac{\phi_{stoch} - \Phi_{stoch}}{\sqrt{Var(\phi_{stoch})}}$.

5. Discussion. We have shown that the escape from the silent as well as from the active phase of the noisy β-cell model can be studied analytically. For the silent phase we used a collective coordinate approach by assuming a Gaussian distribution and the G-method [18, 22]. We could then follow the saddle-node bifurcation at which the silent phase terminates as the noise strength σ varies (Figure 3). For the active phase we used a stochastic Melnikov approach [9], which is new in the context of noisy bursting. We gave an explanation of why a fixed value of $\frac{-M_{det}}{\sigma\sigma_\Xi}$ would predict the z value, for which the system would leave the active phase for different values of σ. The value of $Pr(X > \frac{-M_{det}}{\sigma\sigma_\Xi}) \approx 0.32$ is not obvious, and it should be interesting to see whether it holds for other stochastic systems that are nearly Hamiltonian.

Noise has a bigger influence on the exit from the active phase than on the escape from the silent phase. However, the plateau fraction is roughly unchanged, since a faster escape from the active phase corresponds to the system entering the silent phase later, and vice versa. This is in agreement with the fact that although

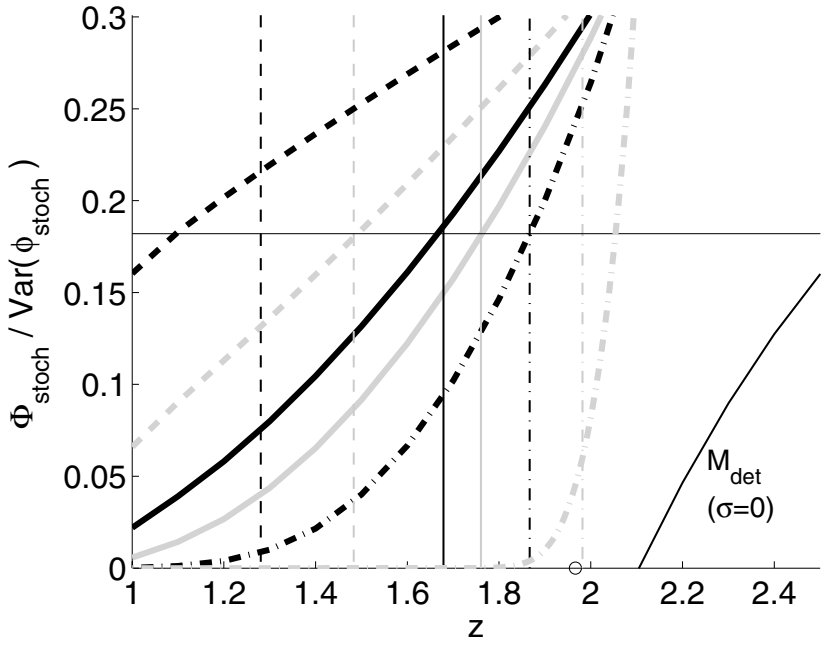


FIG. 5. *Stochastic escape from the active phase is explained by* $\Phi_{stoch}/Var(\phi_{stoch}) \approx 0.182$ *(horizontal black line). The thick curves show* $\Phi_{stoch}/Var(\phi_{stoch})$*, with the same legends and range of* σ *values as in Figure* 4*. They were calculated from* (4.15)*,* (4.17)*, and* (4.18)*. The full curve is* M_{det}*, and vertical lines are means of simulated* z *values as in Figure* 4*.*

single cells have shorter burst periods, the plateau fraction is similar to that of intact islets [11].

We remark that the treatment of the stochastic Melnikov method in [9] was proved rigorously only for the case of uniformly bounded noise, since the paths are then uniformly bounded and current Melnikov theory is applicable. For unbounded processes such as white noise, the results were found and argued to be valid by limit arguments. Biologically, paths are bounded. For example, if w in (2.1) is a gating variable, we have $0 \leq w \leq 1$, which implies that the noise in (2.1) must be bounded. The white noise process was chosen here for mathematical convenience, but in the analysis of biologically more realistic stochastic models the Melnikov approach should be even more reliable.

Assuming that single cells and small clusters of β-cells have shorter burst periods due to noise, the channel sharing hypothesis can explain why. As seen in both simulations as well as the treatment presented here, the stronger the noise intensity, the lower the burst period. If we assume that the cells are coupled with infinite coupling strength (the supercell hypothesis) [4, 21], then in larger groups of cells, the noisy channels will be shared among several cells, and the individual cell would feel smaller fluctuations than if it were isolated, leading to longer burst periods. Extending the methods presented here to groups of coupled cells with finite coupling strength would be interesting in an attempt to gain deeper insight into previously published results [1, 6, 14].

REFERENCES

[1] J. Aguirre, E. Mosekilde, and M. A. Sanjuan, *Analysis of the noise-induced bursting-spiking transition in a pancreatic beta-cell model*, Phys. Rev. E (3), 69 (2004), 041910.

[2] F. M. Ashcroft and P. Rorsman, *Electrophysiology of the pancreatic β-cell*, Prog. Biophys. Mol. Biol., 54 (1989), pp. 87–143.

[3] I. Atwater, L. Rosario, and E. Rojas, *Properties of the Ca-activated K+ channel in pancreatic beta-cells*, Cell Calcium, 4 (1983), pp. 451–461.

[4] T. R. Chay and H. S. Kang, *Role of single-channel stochastic noise on bursting clusters of pancreatic β-cells*, Biophys. J., 54 (1988), pp. 427–435.

[5] G. De Vries, *Multiple bifurcations in a polynomial model of bursting oscillations*, J. Nonlinear Sci., 8 (1998), pp. 281–316.

[6] G. De Vries and A. Sherman, *Channel sharing in pancreatic β-cells revisited: Enhancement of emergent bursting by noise*, J. Theoret. Biol., 207 (2000), pp. 513–530.

[7] G. De Vries and A. Sherman, *From spikers to bursters via coupling: Help from heterogeneity*, Bull. Math. Biol., 63 (2001), pp. 371–391.

[8] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, Software Environ. Tools 14, SIAM, Philadelphia, 2002.

[9] M. Frey and E. Simiu, *Noise-induced chaos and phase space flux*, Phys. D, 63 (1993), pp. 321–340.

[10] J. C. Jonas, P. Gilon, and J. C. Henquin, *Temporal and quantitative correlation between insulin secretion and stably elevated or oscillatory cytoplasmic Ca^{2+} in mouse pancreatic β-cells*, Diabetes, 47 (1998), pp. 1266–1273.

[11] T. A. Kinard, G. De Vries, A. Sherman, and L. S. Satin, *Modulation of the bursting properties of single mouse pancreatic beta-cells by artificial conductances*, Biophys. J., 76 (1999), pp. 1423–1435.

[12] R. Kuske, *Probability densities for noisy delay bifurcations*, J. Statist. Phys., 96 (1999), pp. 797–816.

[13] R. Kuske and S. M. Baer, *Asymptotic analysis of noise sensitivity in a neuronal burster*, Bull. Math. Biol., 64 (2002), pp. 447–481.

[14] M. G. Pedersen, *A comment on noise enhanced bursting in pancreatic beta-cells*, J. Theoret. Biol., 235 (2005), pp. 1–3.

[15] M. Pernarowski, *Fast subsystem bifurcations in a slowly varying Liénard system exhibiting bursting*, SIAM J. Appl. Math., 54 (1994), pp. 814–832.

[16] M. Pernarowski, R. M. Miura, and J. Kevorkian, *Perturbation techniques for models of bursting electrical activity in pancreatic β-cells*, SIAM J. Appl. Math., 52 (1992), pp. 1627–1650.

[17] J. Rinzel, *Bursting oscillations in an excitable membrane model*, in Ordinary and Partial Differential Equations, B. D. Sleeman and R. J. Jarvis, eds., Springer-Verlag, New York, 1985, pp. 304–316.

[18] R. Rodriguez and H. C. Tuckwell, *Statistical properties of stochastic nonlinear dynamical models of single spiking neurons and neural networks*, Phys. Rev. E (3), 54 (1996), pp. 5585–5590.

[19] P. Rorsman and G. Trube, *Calcium and delayed potassium currents in mouse pancreatic β-cells under voltage clamp conditions*, J. Physiol. (London), 374 (1986), pp. 531–550.

[20] A. Sherman and J. Rinzel, *Model for synchronization of pancreatic beta-cells by gap junction coupling*, Biophys. J., 59 (1991), pp. 547–559.

[21] A. Sherman, J. Rinzel, and J. Keizer, *Emergence of organized bursting in clusters of pancreatic β-cells by channel sharing*, Biophys. J., 54 (1988), pp. 411–425.

[22] S. Tanabe and K. Pakdaman, *Dynamics of moments of FitzHugh-Nagumo neuronal models and stochastic bifurcations*, Phys. Rev. E (3), 63 (2001), 031911.

[23] S. Wiggins, *Chaotic Transport in Dynamical Systems*, Springer-Verlag, New York, 1992.

[24] M. Zhang, P. Goforth, R. Bertram, A. Sherman, and L. Satin, *The Ca^{2+} dynamics of isolated mouse β-cells and islets: Implications for mathematical models*, Biophys. J., 84 (2003), pp. 2852–2870.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 543–560

AN OPTIMIZATION APPROACH TO MODELING SEA ICE DYNAMICS. PART 1: LAGRANGIAN FRAMEWORK*

HELGA S. HUNTLEY†, ESTEBAN G. TABAK‡, AND EDWARD H. SUH‡

Abstract. A new model for the dynamics of sea ice is proposed. The pressure field, instead of being derived from a local rheology as in most existing models, is computed from a global optimization problem. Here the pressure is seen as emerging not from an equation of state but as a Lagrange multiplier that enforces the ice's resistance to compression while allowing divergence. The resulting variational problem is solved by minimizing the pressure globally throughout the domain, constrained by the equations of momentum and mass conservation, as well as the limits on ice concentration (which has to stay between 0 and 1). This formulation has an attractive mathematical elegance while being physically motivated. Moreover, it leads to an analytic formulation that is also easily implemented in a numerical code, which exhibits marked stability and is suited to capturing discontinuities. In order to test the theory, the equations for a one-dimensional model are cast in terms of Lagrangian mass coordinates. The solution to the minimization problem is compared to an exact analytic solution derived using jump conditions in a simple test case. Another case is examined, which is somewhat more complicated but still allows our physical intuition to verify the qualitative results of the model. Good agreement is found. A final validation is performed by a comparison with a particle-based model, which tracks individual ice floes and their inelastic interaction in a one-dimensional domain.

Key words. ice dynamics, rheology, Lagrangian fluid dynamics

AMS subject classifications. 76M30, 86A05

DOI. 10.1137/040621156

1. Introduction.

Unlike the dynamics of the ocean waters, sea ice dynamics is relatively little understood. Part of the problem is that ice is a somewhat peculiar substance; it is neither hard as steel, nor elastic like rubber, nor completely fluid like a liquid. Colliding ice floes do not behave like the familiar billiard balls with perfectly round shapes and elastic collisions. An additional challenge for the modeler interested in the large scales of the whole Arctic, or at least an entire strait, is the limit of resolution. It is impossible to follow individual ice floes. Instead the behavior of a field of floes on the order of several square kilometers must be summarized by tracking an average velocity (mass-weighted), a thickness distribution, and the concentration of ice in the grid box. On this scale, one can no longer rely on first physical principles for the solid bodies making up the ice. So what are the laws of physics describing the motion of a half empty box of solid substance that strongly resists compression up to its breaking point but is easily pulled apart due to a multitude of fractures?

Scientists over the years have suggested various analogies, two of which have shown enough promise to have stood the test of time: ice as a fluid, and ice as a granular material. Most models today following the first of these approaches in some way relate back to the sea ice rheology proposed by Hibler in [3], based on the postulate

*Received by the editors December 20, 2004; accepted for publication (in revised form) August 30, 2006; published electronically February 15, 2007. This article derives from work done as part of the Ph.D. thesis of Helga Schaffrin Huntley, while partially supported by the NSF VIGRE program.
http://www.siam.org/journals/siap/67-2/62115.html

†Department of Applied Mathematics, University of Washington, Box 352420, Seattle, WA 98195-2420 (helga@amath.washington.edu).

‡Department of Mathematics, Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., New York, NY 10012 (tabak@cims.nyu.edu, ehs@nyu.edu). The work of the second author was partially supported by grants from the NSF Division of Applied Mathematics.

that ice behaves like a nonlinear viscous-plastic compressible fluid. Improvements have been made on the original model, including an extension with an elastic component to the constitutive law, but the essential character has been maintained (cf., for example, [5]). Parts of the "state-of-the-art" formulation of the rheology are based on the observations that ice resists compression up to a breaking point (plastic); the viscous character is added to avoid multivalued functions, while the elastic terms are mostly introduced for greater numerical stability. While the numerical results are generally good, to some extent this is due to parameter tuning to available (though often scarce) data.

Models based on a cavitating fluid rheology also fall into this first category. They do not naturally incorporate shear strength (although this has been partially addressed), but are computationally somewhat simpler and less expensive, making use of an iterative correction scheme. These models also have had some success in reproducing realistic ice transport (cf., for example, [2]).

In the second category, a granular rheology has been developed by Tremblay and Mysak [11], as well as others (e.g., the CRREL[1] uses a high-resolution granular sea ice model, based on [4]). These are better equipped to handle such tasks as tracking leads in an ice field, although a much higher resolution is required for such problems.

Model intercomparison studies have found that the choice of rheology has a significant impact on the model output; see, e.g., [6] and [1]. A somewhat more comprehensive study, SIMIP, the Sea Ice Model Intercomparison Project, was carried out in the late 1990s. It compared viscous-plastic, cavitating fluid, compressible Newtonian fluid, and free-drift with velocity corrections rheologies. Overall, it was found that the viscous-plastic rheology produced the best results, while the free-drift simulation showed large errors in ice drift, thicknesses, and export through the Fram Strait, the compressible Newtonian model yielded excessive ice thickness build-up in the central Arctic, and the cavitating fluid rheology resulted in errors in ice drift and the thickness pattern. However, in some respects none of the models gave entirely satisfactory results. Thus, for example, in an analysis of the summer sea ice extent anomalies, none of the model results lies consistently within the error band of the observations (from satellite data). The same is true for anomalies of annual mean ice thickness in the Beaufort Sea (where the observational data was collected with upward looking sonars). These results were reported in [7]. The present work on a new approach to treating the internal stress term for sea ice dynamics was motivated in part by the realization that existing models not only disagree significantly with each other but also have remaining difficulties reproducing some observed phenomena, such as the formation of ice arches in the Canadian Arctic Archipelago.

We have chosen to follow the first approach, an analogy to fluids, but are employing a novel global formulation of the rheology as the solution to an optimization problem. We attempt to build a theoretical framework, with supporting numerical experiments, to reproduce and explain the relevant observations of sea ice dynamics. We start with the analogy that ice behaves like a fluid with some special properties. In particular, sea ice exhibits semi-incompressibility: It allows divergence without much resistance, due to the many cracks and leads within the ice pack, but strongly resists convergence at high concentration. The goal is to find an expression for the pressure enforcing this semi-incompressibility, which is mathematically elegant and computationally efficient, allowing for clean analytic solutions in simple cases and numerical simulations of more complicated ones. Reliance on parameterization is to

[1]Cold Regions Research and Engineering Laboratory of the US Army Corps of Engineers.

be minimized. To this end, we start with the fluid equations and investigate how they need to be modified to retain validity in the context of sea ice. The central hypothesis is that the pressure acting within the ice field is the minimum necessary to prevent the unrealistic situation of multiple floes occupying the same space. In other words, the pressure term enforces the condition that the area fraction covered by ice (the concentration) may not exceed 1. Our formulation permits the calculation of the pressure using linear programming, benefiting from existing schemes, without tuning to observations.

The goal of the work reported in this article is to demonstrate the feasibility of having the pressure solve a variational problem. To this end, we concentrate on the simplest possible scenario, one of unforced one-dimensional flows, with uniform ice thickness and infinite yield pressure, so that no crushing occurs. A companion paper (Part 2) discusses the effects of ice yielding. Current work, which will be reported in later publications, involves extending the flow to two horizontal dimensions and allowing for nontrivial ice thickness distributions. A more detailed discussion of many of the results in this paper can be found in [9]. Here we will consider a simplified Lagrangian system, which allows both analytic solutions to basic test cases and a straightforward numerical implementation to be tested against a particle-based model resolving individual ice floes. As will be shown, the minimal pressure hypothesis leads to very encouraging results in the studied test cases, justifying further investigation of this particular rheology.

2. Coarse grained ice dynamics as fluid dynamics. We shall use the following variables:

$$\begin{aligned} c &= \text{concentration of ice (fraction of sea surface area covered by ice)},\\ h &= \text{average thickness of the ice},\\ \mathbf{u} &= (u,v) = \text{horizontal velocities},\\ S &= \text{sources} - \text{sinks of ice (melting and freezing, precipitation)},\\ F^x, F^y &= \text{sum of zonal, meridional forces}\\ &\qquad \text{(Coriolis, wind, currents, sea surface tilt, pressure)}. \end{aligned}$$

Note that in the following we are taking the density of ice ρ, which is nearly constant, to be identically 1. This convention simplifies the notation and has no influence on the qualitative results we are interested in. Equivalently, one can describe this as absorbing the density into the thickness parameter, so that h is measured in $\mathrm{kg/m^2}$.

Most of today's sea ice dynamics models employ a thickness distribution function to allow for ice of various thicknesses in any one grid cell (cf. [10]). For the simple model we are building here to verify the minimal pressure hypothesis, we will not include this level of complexity at this time. Similarly, we are not concerned with tracking a velocity distribution within a grid cell, but assume the velocities u and v to be mass-averaged velocities for the entire box.

In Eulerian coordinates, the mass conservation and momentum conservation equations can be written as

$$(ch)_t + \nabla\cdot(ch\mathbf{u}) = S, \tag{1}$$

$$(chu)_t + \nabla\cdot(chu\mathbf{u}) = F^x, \tag{2}$$

$$(chv)_t + \nabla\cdot(chv\mathbf{u}) = F^y, \tag{3}$$

where the product ch (really $ch\rho$) plays the role of a density in analogy with the conventional fluid equations.

What differentiates the case of ice from standard fluids is that, in addition to these two conservation laws, we also have the constraint that c can vary (unlike the density of incompressible fluids) but may not exceed 1, which is enforced by a pressure force. In this paper, we want to investigate the nature of this pressure force and a new way to calculate it. To isolate this issue, we will consider the one-dimensional case without sources or sinks ($S = 0$). By defining $F = F^x/(ch)$, subtracting u times (1) from (2), and dividing by ch, the system reduces to

$$(ch)_t + (chu)_x = 0, \tag{4}$$

$$u_t + u\,u_x = F. \tag{5}$$

Notice that, even if F were given (which it is not, since the pressure is one of the variables to be determined), this system has only two equations but three unknowns. Mass conservation alone does not provide for a way to evolve the fractional area c and the mean thickness h separately. Extra physical assumptions are necessary to complete the system's description. In the absence of crushing (or sources or sinks), a natural assumption is that the thickness h is advected with ice floes:

$$h_t + u\,h_x = 0. \tag{6}$$

In words, when a pack of floes is pulled apart, the floes do not become thinner: It is the space between them that increases, thus reducing the fractional area coverage c. (The assumption of advection of h is relaxed in Part 2 of this work.)

3. The Lagrangian formulation. To gain further insight into this system and to make analytic solutions easier to obtain, we introduce the Lagrangian mass coordinates[2]

$$\begin{cases} \xi = \int_0^x ch\,d\hat{x}, \\ \tau = t. \end{cases} \tag{7}$$

The resulting Lagrangian equations in these coordinates for one-dimensional ice motion without sources or sinks are

$$\left(\frac{1}{ch}\right)_\tau = u_\xi, \tag{8}$$

$$u_\tau = F. \tag{9}$$

(See the appendix for the derivation.)

Introducing the variable $k \equiv \frac{1}{ch} - 1$ yields

$$k_\tau = u_\xi, \tag{10}$$

$$u_\tau = F. \tag{11}$$

While these equations have a beguilingly simple form, we have now lost sight of the crucial variable c, which we have to constrain. We would like to translate this

[2]Lagrangian mass coordinates are often used in astronomical papers, as well as in studies of hydrodynamics; cf., for just one example, [12].

constraint into a constraint on k. In the case where $h \equiv 1$, this is easily done: The constraint $c \leq 1$ now becomes $k \geq 0$. An extension to varying h is straightforward (the constraint becomes $k \geq \frac{1-h}{h}$), but it obscures the pressure effect. Thus, for easier visualization, we shall focus on the particular case of constant ice thickness $h \equiv 1$. (Variations in ice thickness are reintroduced in Part 2.) Note that, in addition to excluding melting and freezing processes as well as precipitation, fixing h also eliminates crushing from the problem. While a realistic and versatile ice model will ultimately require us to bring these components back, the nature of the pressure force and its role in ice dynamics can more easily be investigated in isolation. For this purpose we make one more simplification and isolate the pressure term by setting all other forces equal to zero.

What form this pressure term should take is not immediately obvious. Various suggestions have been made over the years, which fall into two main categories. On the one hand, there are the viscous-plastic or viscous-plastic-elastic models, which all hark back to the original formulation in terms of the stress/strain yield curve in [3]. On the other hand are the cavitating fluid models, which calculate the pressure in a series of correction steps (cf. [2]). While these are backed up by more or less realistic model outputs, they are dependent on empirical parameter tuning and exhibit some numerical shortcomings in the first case and follow an iterative relaxation scheme in the latter. The goal of our approach is to minimize the reliance on empirical parameter tuning and iterative schemes to the extent possible. Our physical intuition tells us that the pressure does not act unless it is necessary to prevent ice concentration from exceeding 1, i.e., to prevent two ice floes from occupying the same space. In other words, while c is far from 1, the pressure force $F = 0$, and the flow follows $u_\tau = 0$. Each parcel's velocity does not change with time. By adding a pressure to the system, we would like to deviate as little as possible from this route while satisfying the constraint. This suggests a mathematical formulation as a constrained optimization problem, where the pressure is defined as a Lagrange multiplier.

We do not expect the pressure force to push the ice apart, as any elastic rheology would. Rather, it builds up as the ice converges, solely to prevent multiple floes from occupying the same space. The goal then is to find a pressure that allows us to minimize the change in u over time. Lagrange multipliers were invented for just such constrained optimization problems. Here we have two constraints acting on u, the equation for the evolution of k and the lower limit on k. Since k is defined at each point ξ, this in fact amounts to infinitely many constraints, one for each $k(\xi)$.

A note on notation: In the following (except for section 9), we will be discussing only the Lagrangian formulation of the problem. Thus, we will simplify our notation by replacing the Greek letters for the independent variables by their Roman cousins again, i.e., by writing the Lagrangian mass coordinate ξ as x and the Lagrangian time coordinate τ as t.

4. Pressure as a Lagrange multiplier. The problem is at each time t to

(12)	minimize	$\|u_t\|$,
(13)	given	$k_t = u_x$,
(14)	subject to the constraint	$k \geq 0$.

Discretizing this system in time, using an implicit scheme, we can reformulate the

problem for each n as follows:

$$\text{minimize} \qquad \|u^{n+1} - u^n\|, \tag{15}$$

$$\text{given} \qquad k^{n+1} = k^n + \Delta t\, u_x^{n+1} \quad \forall\, n \geq 0, \tag{16}$$

$$\text{subject to the constraint} \qquad k^n \geq 0 \quad \forall\, n \geq 0. \tag{17}$$

To simplify the notation, let $\tilde{u} = u^{n+1}$, $u = u^n$, and $k = k^n$. Choosing to define the norm used on u_t as the 2-norm, our task is then to

$$\text{minimize} \quad \int (\tilde{u} - u)^2\, dx, \tag{18}$$

$$\text{subject to} \quad k + \Delta t\, \tilde{u}_x \geq 0. \tag{19}$$

The corresponding variational principle states that there exist Lagrange multipliers $\lambda(x) \leq 0$ such that

$$\delta \int \left\{ (\tilde{u} - u)^2 + \lambda(x)(k + \Delta t\, \tilde{u}_x) \right\} dx = 0, \tag{20}$$

from which it follows (as shown in the appendix) that

$$2\,(\tilde{u} - u) = \Delta t\, \lambda_x \quad \Longrightarrow \quad \frac{\tilde{u} - u}{\Delta t} = \frac{1}{2}\lambda_x.$$

The equivalent continuous statement is

$$u_t = \frac{1}{2}\,\lambda_x. \tag{21}$$

We will adopt the convention of writing the pressure as $p = -\lambda/2$. Our final system for the Lagrangian formulation of one-dimensional sea ice dynamics with constant thickness and without external forces is thus

$$k_t = u_x, \tag{22}$$

$$u_t = -p_x, \tag{23}$$

$$k \geq 0, \tag{24}$$

$$p \geq 0. \tag{25}$$

5. The conundrum is unresolved: Introducing the concept of minimal pressure. We have just derived a form for the pressure term as the x-derivative of a Lagrange multiplier. While this gives us the nice result that the pressure looks very much like the pressure for incompressible fluids, we still have little information about what the pressure actually is.

If we were able to resolve single floes, a constitutive law for frozen water would provide us with a way to calculate the pressure within each ice floe (provided that we know the external forces, including the pressure applied by other ice floes). Since we do not have resolution at this scale, however, we have to find a different method.

The pressure, constructed as a Lagrange multiplier, serves the purpose of enforcing the restriction that ice concentration cannot exceed 1 (or, equivalently, k cannot dip below 0). As long as k is far from 0, one would consequently expect p to equal 0. On the other hand, when k reaches 0, p needs to take on values to prevent it from decreasing further. (Note that this means that always one of k and p is zero.) There

is no reason, however, why it should push the ice apart, i.e., why the arising pressure should exceed the minimum necessary to satisfy the constraint. It is then a reasonable suggestion that the pressure should be calculated as the minimum necessary to enforce the constraint.

One should note at this point that in the more complex cases where h is allowed to vary, and in particular where crushing of the ice is permitted, this p needs to be limited by a maximal p. This is standard in other ice models (for two different formulations, see [3] and [8]). This extension for the minimization formulation for the pressure as we propose here will be discussed in Part 2.

Beyond the heuristic argument presented above, the formulation for the pressure as minimal suggested here is also mathematically appealing based on optimization theory. The first observation above (that $\mathbf{k} \cdot \mathbf{p} = 0$, where $\mathbf{k}$ and $\mathbf{p}$ denote the vectors of the respective discretized quantities) is nothing but the complementary slackness requirement of the Karush–Kuhn–Tucker conditions for constrained optimization problems.[3] Note also that because of (23), minimizing $\|u_t\|_2$ is equivalent to minimizing $\|p_x\|_2$. Since $p \geq 0$ and $p = 0$ whenever $k > 0$, minimizing $\|p_x\|_2$ is in turn equivalent to minimizing $\|p\|_2$ or $\|p\|_1$, as will be shown below in section 7. It is thus natural to choose a formulation that aids in the calculations, and we will use the 1-norm of p to take advantage of available robust linear optimization techniques.

6. An analytic solution to a well understood problem using jump conditions. To build a better understanding of the form that the pressure takes, let us start by considering a problem whose solution we know from physical considerations. We take the half-infinite (one-dimensional) domain $x \leq 0$, with a wall at $x = 0$. Ice is initially distributed according to the given function $k(0, x) = k_o(x)$ and moves according to $u(0, x) = u_o(x)$. We will assume that initially there are no patches of consolidated ice; i.e., $k_o(x) \neq 0 \ \forall x$. If we assume that pressure arises inside the ice only when necessary, it follows that $p(0, x) = 0 \ \forall x$.

If anywhere in the domain $u_x < 0$, then the ice will accumulate somewhere until $k = 0$ (or, equivalently, $c = 1$). The wall at $x = 0$ requires that $u(t, 0) = 0 \ \forall t$. Thus, a positive initial velocity anywhere will ensure ice accumulation.

The first problem we will consider has the first ice build-up occurring at the wall. (This can, for example, be achieved by setting $u_o(x) \equiv u_o = \text{constant} > 0$.) At the edge of the consolidated ice, a discontinuity in both concentration and velocity arises. We denote the location of this shock, marking the interface between the region with $k = 0$ ($c = 1$) and that with $k > 0$ ($c < 1$), by $x = x_c$ and the time when this shock first forms by $t = t_c$; i.e.,

$$t_c = \min \{t : k(t, x) = 0 \text{ for some } x < 0\}, \tag{26}$$

$$x_c(t) = \min \{\tilde{x} : k(t, x) = 0 \ \forall x \geq \tilde{x}\}. \tag{27}$$

For $t < t_c$, $p(t, x) = 0 \ \forall x$. The equations we solve for this portion are the very

[3] Physically, pressure is generally defined only up to a constant, since only its gradient enters into the dynamics. Similarly, here it is easy to see that adding an arbitrary (positive) constant to a p which solves the system (22)–(25) yields another solution. Strictly speaking, the KKT conditions do not necessitate $p \geq 0$ and $\mathbf{k} \cdot \mathbf{p} = 0$; they do provide for the *existence* of a p satisfying these conditions and solving the system (22)–(25). Hence we will choose the arbitrary constant such that they hold and $\min p = 0$.

simple set

$$k_t = u_x, \tag{28}$$

$$u_t = 0, \tag{29}$$

whose solution is

$$k(t,x) = k_o(x) + tu_{ox}(x), \tag{30}$$

$$u(t,x) = u_o(x), \tag{31}$$

$$p(t,x) = 0. \tag{32}$$

Assuming no consolidation except at the wall, this solution holds on $\{(t,x) : t < t_c \text{ or } x < x_c\}$. Once a region of consolidated ice starts to form along the wall, we know that for $x > x_c$,

$$k(t,x) = 0, \tag{33}$$

$$u(t,x) = 0. \tag{34}$$

Thus, for $t > t_c$ and $x > x_c$,

$$u = 0 \implies u_t = 0 \implies p_x = 0 \implies p(t,x) = p(t). \tag{35}$$

In other words, p is constant throughout the region of consolidation at a given point in time.

The only unknowns are p and x_c. While the solutions to the left and to the right of x_c are smooth, there is a discontinuity, both in k and in u, at x_c itself. Using the corresponding jump conditions, we can calculate p and x_c.

Equations (22) and (23), respectively, imply that

$$[\![k]\!]\dot{x_c} = -[\![u]\!], \tag{36}$$

$$[\![u]\!]\dot{x_c} = [\![p]\!], \tag{37}$$

where $[\![\cdot]\!]$ denotes the jump across the shock and $\dot{x_c}$ is the shock speed. It follows from (36) that

$$\dot{x_c} = -\frac{[\![u]\!]}{[\![k]\!]} = -\frac{u_o(x_c)}{k_o(x_c) + tu_{ox}(x_c)}. \tag{38}$$

Together with the initial condition provided by evaluating (27) at t_c, this completely determines x_c. Combining (36) and (37), we find

$$[\![p]\!] = -\frac{[\![u]\!]^2}{[\![k]\!]} \implies p(t,x) = \frac{[u_o(x_c)]^2}{k_o(x_c) + tu_{ox}(x_c)} \qquad \text{for } x > x_c. \tag{39}$$

One should note that this procedure of analyzing jump conditions to determine the pressure can easily be extended to examples with the consolidated ice not against a wall or with multiple consolidated regions. However, such an approach quickly becomes unmanageable, as the number of coupled nonlinear equations increases with the number of consolidated regions (two for each discontinuity, i.e., four for any region not against a wall). Besides, the locations of the interfaces between consolidated and nonconsolidated areas would need to be tracked, a laborious enterprise. Instead we want to use the minimal pressure hypothesis, which simplifies the calculations by making them global and, in the language of numerical conservation laws, is *capturing* rather than *tracking* the boundaries of the consolidated regions.

7. Pressure minimization and norm comparison. The toy problem we consider for a first validation of the minimal pressure hypothesis follows the example described in the previous section: Ice is moving towards a coast, where it consolidates and pressure builds up. We discretize (22) and (23) using a staggered grid and a backward Euler scheme as follows:

$$k_j^{n+1} = k_j^n + \frac{\Delta t}{\Delta x}\left(u_{j+\frac{1}{2}}^{n+1} - u_{j-\frac{1}{2}}^{n+1}\right), \tag{40}$$

$$u_{j+\frac{1}{2}}^{n+1} = u_{j+\frac{1}{2}}^{n} - \frac{\Delta t}{\Delta x}\left(p_{j+1}^{n+1} - p_j^{n+1}\right). \tag{41}$$

The constraints are

$$k_j^{n+1} \geq 0 \quad \forall\, j, n, \tag{42}$$

$$p_j^{n+1} \geq 0 \quad \forall\, j, n. \tag{43}$$

Note that it is necessary to use an implicit scheme for the evolution equation of k, in order to be able to satisfy the constraint. As p does not have an evolution equation, it is immaterial whether we use an explicit scheme or an implicit scheme for the u-equation. We have chosen to write it implicitly for consistency.

We place the wall at $j = 5$. As boundary conditions, we take that $p_{-1}^{n+1} = p_0^{n+1} = 0$ (until the shock reaches this boundary) and $p_5^{n+1} = p_4^{n+1}\ \forall\, n$. As initial conditions, we assume that the shock is located at $j = 2.75$. To the left of the shock, $p = 0$, $k = 1/2$, and $u = 1$. To the right of the shock, $k = 0$ and $u = 0$. Δt is set to 0.5, and Δx is set to 1. (This allows us to capture the shock, which travels with speed 2. Of course, this resolution is very coarse. However, it serves to illustrate the point in this toy example.)

After the first time step,

$$u_{j+\frac{1}{2}}^{1} = \begin{cases} 1 - \frac{1}{2}\left(p_{j+1}^{n+1} - p_j^{n+1}\right) & \text{if } j < 3, \\ -\frac{1}{2}\left(p_{j+1}^{n+1} - p_j^{n+1}\right) & \text{if } j \geq 3. \end{cases} \tag{44}$$

So the constraints (42) take the following form:

$$k_0^1 = \frac{1}{2} - \frac{1}{4}[p_1^1 - 2p_0^1 + p_{-1}^1] \geq 0, \tag{45}$$

$$k_1^1 = \frac{1}{2} - \frac{1}{4}[p_2^1 - 2p_1^1 + p_0^1] \geq 0, \tag{46}$$

$$k_2^1 = \frac{1}{2} - \frac{1}{4}[p_3^1 - 2p_2^1 + p_1^1] \geq 0, \tag{47}$$

$$k_3^1 = 0 - \frac{1}{2} - \frac{1}{4}[p_4^1 - 2p_3^1 + p_2^1] \geq 0, \tag{48}$$

$$k_4^1 = -\frac{1}{4}[p_5^1 - 2p_4^1 + p_3^1] \geq 0. \tag{49}$$

Now we have a system of five inequalities constraining p_1, p_2, p_3, and p_4. (Recall that $p_{-1} = p_0 = 0$ and $p_5 = p_4$.) Minimizing the 2-norm of p_x while satisfying these

constraints gives for $T = \Delta t$ that

$$\begin{aligned} p_0^1 &= 0, & k_0^1 &= \frac{1}{2}, & u_{.5}^1 &= 1, \\ p_1^1 &= 0, & k_1^1 &= \frac{1}{2}, & u_{1.5}^1 &= 1, \\ p_2^1 &= 0, & k_2^1 &= 0, & u_{2.5}^1 &= 0, \\ p_3^1 &= 2, & k_3^1 &= 0, & u_{3.5}^1 &= 0, \\ p_4^1 &= 2, & k_4^1 &= 0, & u_{4.5}^1 &= 0. \end{aligned}$$

From the jump conditions (cf. (39)) it follows that to the right of the shock

$$p = 2. \tag{50}$$

The shock speed can be calculated from (36) or (37) to be

$$\dot{x_c} = -2. \tag{51}$$

Thus the solution obtained by minimizing $\|p_x\|_2$ is exactly what the analytic solution predicts: After time $\Delta t = 0.5$, the shock traveled one step to the left, with u and k retaining the same values on either side of the shock, while $p = 2$ to the right of the shock. Similarly satisfying results were obtained for longer runs and for different initial conditions. In cases where the discretization does not allow for capturing the location of the shock exactly, a certain amount of numerical smoothing around the shock occurs, with p, k, and u taking on intermediate values.

The three norms we want to compare are

1. the 2-norm of p_x, i.e., $\sum (p_i - p_{i-1})^2$,
2. the 2-norm of p, i.e., $\sum p_i^2$,
3. the 1-norm of p, i.e., $\sum |p_i|$.

There are, of course, many other possible norms. These are some of the most natural choices and can be optimized with existing tools.

Solving the same problem above, but minimizing the 1- or 2-norm of p, gives exactly the same answer. This suggests that the norm chosen (at least from among these three) does not impact the result. In fact, this can be shown as follows.

First, consider $\|p_x\|_2$, with the additional condition that $\min p = 0$ (see footnote 3). Recall that the KKT conditions then require $p = 0$ when $k > 0$. Minimizing one of the norms of p itself also calls for p to be 0 whenever possible, in particular whenever $k > 0$.

Within regions of consolidated ice, where $k = 0$, we know that $k_t \geq 0$, and hence, by (22), $u_x \geq 0$.

(i) If $u_x > 0$, then $k_t > 0$ and $p_x = 0$ is a solution. (Physically, this is the situation where the consolidated ice is being pulled apart without the pressure acting.)

(ii) If $u_x = 0$, then $u_{xt} \geq 0 \Longrightarrow p_{xx} \leq 0$ (by (23)).

The values of p at the endpoints of the consolidated region are determined by the jump conditions, which have to be satisfied because of the constraining equations. Minimizing $\|p\|$, either as a 2-norm or as a 1-norm, with $p_{xx} \leq 0$ requires p to be linear within the consolidated region. The same is true for $\|p_x\|_2$. (Note, however, that minimizing the 1-norm of p_x leads to nonunique solutions, where the one found using the other norms is but one possibility.)

Thus, for any of the three norms:

(i) $p = 0$ outside consolidated regions.

(ii) p at the boundaries of consolidated regions is prescribed by the jump conditions imposed by the constraining equations.

(iii) p is linear within consolidated regions.

It follows that the solutions are exactly the same.

As mentioned before, we have decided for our numerical work to rely on the 1-norm of p in order to facilitate the numerical optimization.

8. The numerical model and another test case. Our one-dimensional Lagrangian ice dynamics model for ice with uniform and constant thickness and without forcing is based on the discretization (40)–(43). For the optimization step, we use either Matlab's built-in function `linprog` or a self-coded simplex method.

The problem described above (ice flowing towards a wall) can be successfully modeled this way (also with much better resolution and different initial conditions). We have also modeled, as another example, the case of a periodic domain, with initially uniformly distributed ice. The initial velocity function is sinusoidal with average velocity 0. Again, our physical intuition tells us what the solution should be: We expect the ice to consolidate in the middle of the domain. This is indeed what happens; see Figure 1. Given the initial conditions, we can also analytically calculate the first consolidation time, which in this case is $1/4\pi \approx 0.0796$. Again, the numerical results agree, predicting the first time for $c = 1$ to lie in the interval $[0.07875, 0.08]$ when a temporal step size of 0.00125 was chosen. The comparison of norms was again carried out in this model, which confirmed the previous conclusion that they yield the same answer, whereby minimizing the 1-norm of p is calculationally the least expensive. Various other initial conditions for u, including nonsymmetric cases, were studied, with the numerical results agreeing with the analytic solution (as far as this was easily obtained) and/or physical intuition.

9. Another validation: Comparison with a particle-based method. Further evidence in favor of the minimal pressure hypothesis comes from a comparison with a particle-based model. Here the one-dimensional sea ice motion is simulated as the interaction of individual "particles" (i.e., floes). Each of these is traced through space and time. Collisions are required to be mass- and momentum-conserving and inelastic (again following the observation that the pressure arising from collisions and consolidation does not lead to divergence). This is a rather costly method of modeling sea ice dynamics. However, for the purposes of validation, it is useful. After making the appropriate transformations from the respective coordinate system of each model to Eulerian coordinates, the concentration distributions resulting from the particle-based model and the continuum model with the minimal pressure hypothesis were compared, exhibiting remarkable agreement.

For the comparison, all external forces as well as crushing were again ignored. The initial concentration was taken to be constant c_o throughout the (periodic) domain. This is implemented in the collision model by an even spacing of identical particles. The number of particles corresponds to the resolution; the initial concentration determines the width of each particle. Later, the concentration is calculated at a particle point as the average area coverage between neighboring particles, i.e., at the point k

$$c(k) = \frac{2w}{x(k+1) - x(k-1)}, \tag{52}$$

where w is the width of a parcel, $c(k)$ is the concentration, and $x(k)$ is the location of the kth parcel. The velocities are tracked for each particle. They remain constant

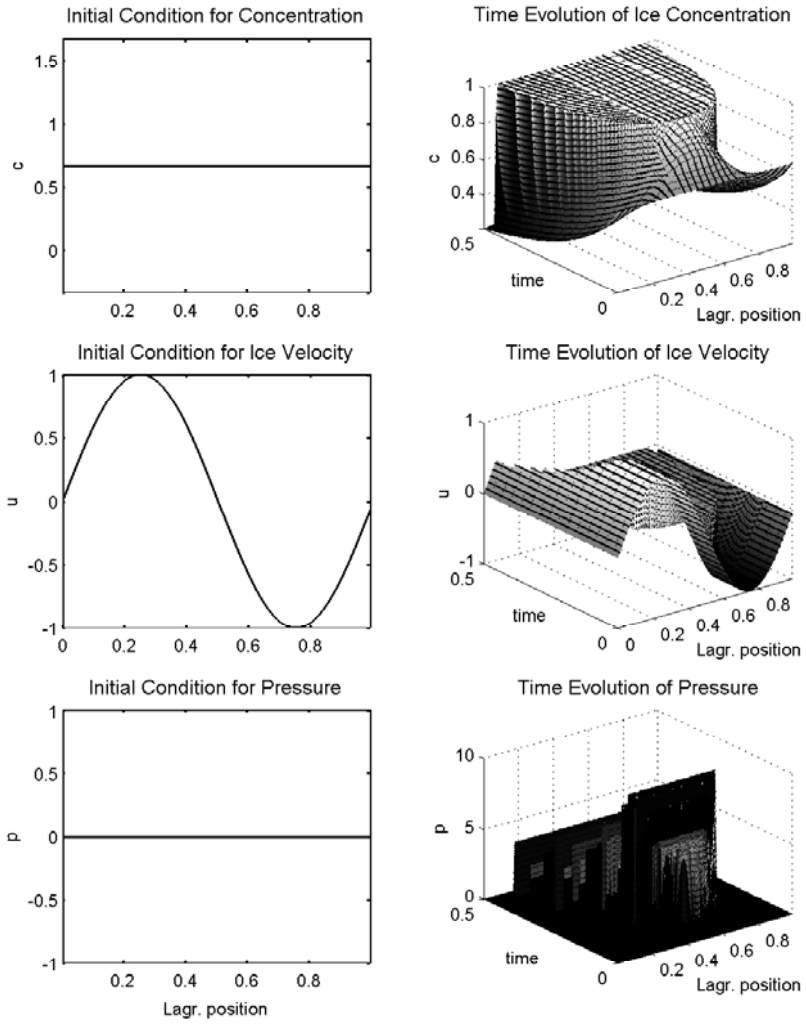


FIG. 1. *The numerical results for ice consolidating in the center of the domain. Initial conditions are given in the plots on the left. The evolution is exhibited in the plots on the right. Note that the x-coordinate here is the Lagrangian position.*

unless a collision occurs, in which case momentum conservation dictates the new velocity of the consolidated region: For particles j_1 through j_2 involved in the collision, each will have new velocity

$$u_{new} = \frac{\sum_{i=j_1}^{j_2} m_i\, u_i}{\sum_{i=j_1}^{j_2} m_i}, \tag{53}$$

where u_i denotes the velocity of the ith particle and m_i its mass. Here, as in the continuum model, constant thickness is assumed. Hence the concentration $c(i)$ is proportional to the mass m_i and can be substituted for it in the formula (53).

The initial velocity for both model runs was taken to be sinusoidal again, given

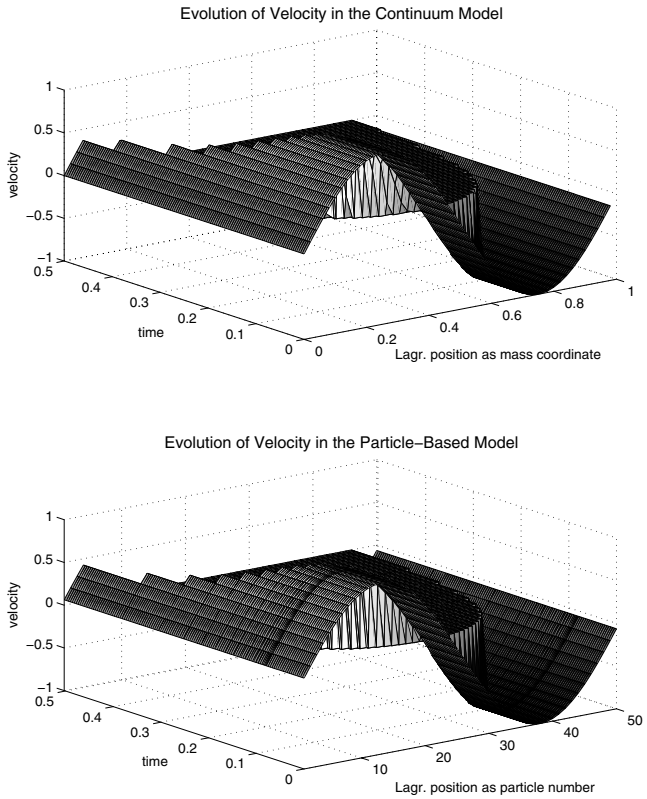


FIG. 2. *The velocity evolutions derived by the continuum model using the minimal pressure hypothesis (top frame) and the particle-based model (bottom frame). Note that the x-coordinates here are the Lagrangian x-coordinates of each model.*

in Eulerian coordinates as $u_i^E(x) = \sin(2\pi c_o x)$. (Note that this is equivalent to the case presented in the previous section. In the Lagrangian mass coordinates used there, the same velocity is written as $u_i^L(\xi) = \sin(2\,\pi\,\xi)$, where we are reverting to the Greek notation for clarification.) Two comparisons were carried out, one of the resulting velocity fields, the other of the resulting concentration fields. The Lagrangian coordinates of the continuum model with the minimal pressure hypothesis can be converted to Eulerian coordinates by numerically integrating the velocity u in time; the Eulerian positions of the particles in the collision model are recorded at each step.

Figure 2 shows the evolution of the velocities from the two models over 125 time steps. They are clearly very similar. A more detailed comparison is given in Figure 3. Here the Lagrangian coordinates are converted to Eulerian ones, and two different times are chosen. Note that in both plots, the velocities agree almost exactly. The one point of the particle model overshooting the continuum model (both in the positive and negative directions) in each plot is simply not resolved by the continuum model.

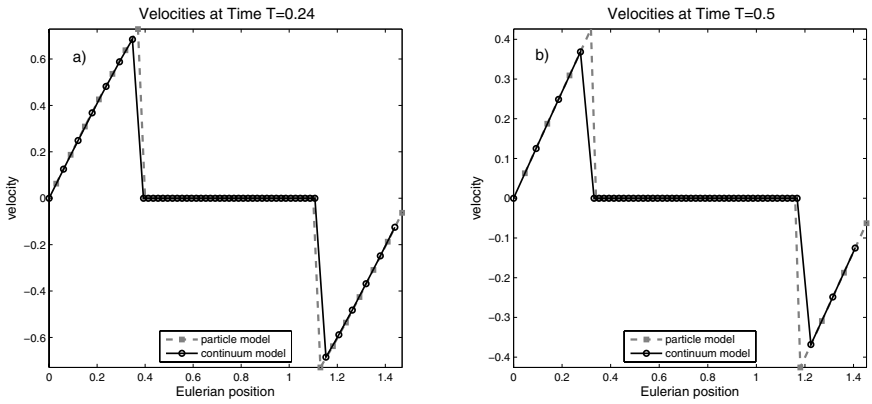


FIG. 3. *The velocities derived with the continuum (solid line, circles) and the particle (dashed line, squares) models are compared at* (a) *an intermediate time and* (b) *the end of the run. The x-coordinates here are Eulerian x-coordinates.*

Figure 4 shows the evolution of the concentrations from the two models over the 125 time steps. Again, a close similarity is apparent. To compare the concentration fields more exactly, we have chosen two points in time, one at the first collision (marked in the continuum model as the first time with nonzero pressure), the other at the end of the run. The results are shown in Figure 5. Up to the first collision, the concentration fields are identical between the two models. Recall that this time can be found analytically to be $\frac{1}{4\pi} \approx 0.0796$. With the resolution used here ($\Delta t = 0.004$, $\Delta\xi = 0.02$), the continuum model places it in the interval $[0.076, 0.080]$, while the particle model predicts that the first collision will occur at time 0.0796. In other words, both models are in good agreement with the exact value. At time $T = 0.5$, small differences in the concentration fields are discernible. However, once again much of this difference can be ascribed to the fact that the two models do not resolve the same points.

One can conclude then that the model suggested here, using a minimization technique to calculate internal pressure, which is used in turn to update the velocity field, produces results entirely consistent with the physical description of sea ice motion as the interaction of ice floes through inelastic collisions.

10. Conclusions. We propose here a new closure for the equations of motion for sea ice modeled as a fluid. In addition to the momentum and the mass conservation equations, an expression for the internal forcing due to pressure has to be found. Based on the observation that collisions of ice floes at natural speeds tend to be inelastic, it was hypothesized that the internal pressure should stay at the minimum required to enforce a concentration of at most 1. This formulation is also consistent with treating the pressure as a Lagrange multiplier in the optimization problem minimizing the deviation from the unforced path under the constraints of the mass conservation equation and the limits on the concentration ($0 \leq c \leq 1$).

To investigate the validity of this closure, a one-dimensional Lagrangian model was set up. Two classes of test cases were studied, ice moving towards a wall with initially constant velocity and ice moving in a periodic domain with initially sinusoidal velocity. Thickness was held constant for easier identification of the effects of the minimal pressure hypothesis, an assumption that will be relaxed in Part 2. For the

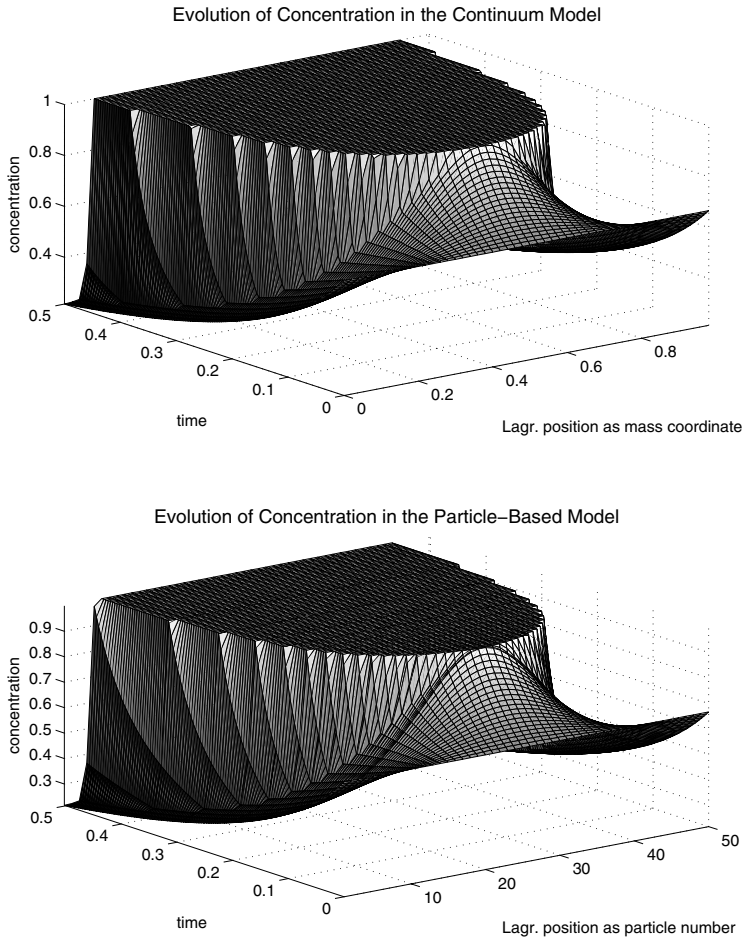


FIG. 4. *The concentration evolutions derived by the continuum model using the minimal pressure hypothesis (top frame) and the particle-based model (bottom frame). Note that the x-coordinates here are the Lagrangian x-coordinates of each model.*

first case, an exact analytic solution was derived—which was reproduced to high accuracy by the model (except for smoothing of the shock when the resolution was too coarse). The second case produced results consistent with physical intuition. A further validation was given by comparison with a particle-resolving model using inelastic collisions. Within the limits of their respective resolutions, the output of these two models was in exact agreement.

It is thus possible to conclude that the minimal pressure hypothesis leads to correct solutions, at least in the cases studied here. This is promising, because the work presented here forms the basis for more complex versions of the dynamics model using the minimal pressure hypothesis. The next step—a translation to Eulerian coordinates, allowing variable ice thickness, and implementing a finite ice strength—is presented in Part 2. Clearly, further study is required to determine the usefulness of this approach in large-scale simulations. Thus, for modeling real situations, an

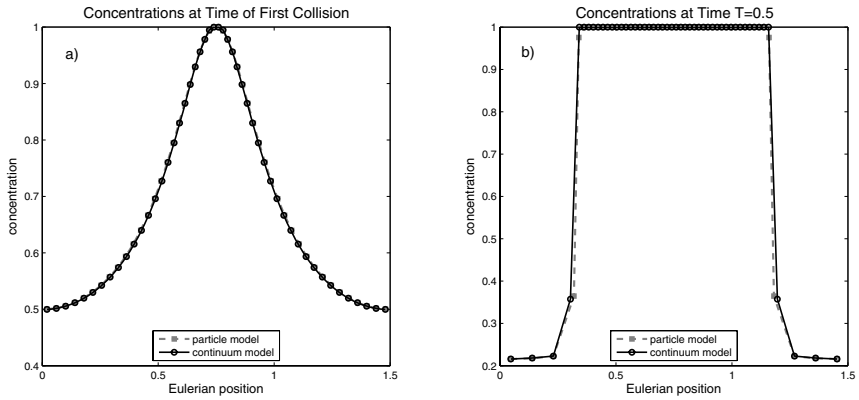


FIG. 5. *The concentrations derived with the continuum (solid line, circles) and the particle (dashed line, squares) models are compared at* (a) *the time of first collision and* (b) *the end of the run. The x-coordinates here are Eulerian x-coordinates.*

extension to two dimensions is necessary. Restoring the other forcing terms will be a minor task. The good agreement of the model results so far with analytic and particle-resolving model solutions justifies such further investigations, currently underway.

Appendix. Derivations.

A.1. Lagrangian equations. Recall that (7) defined the Lagrangian coordinates as

$$\begin{cases} \xi = \int_0^x ch\, d\hat{x}, \\ \tau = t. \end{cases}$$

The partial derivatives relating them to the Eulerian coordinates (t, x) are as follows:

$$\frac{\partial \xi}{\partial x} = ch, \tag{54}$$

$$\frac{\partial \xi}{\partial t} = -chu, \tag{55}$$

$$\frac{\partial \tau}{\partial x} = 0, \tag{56}$$

$$\frac{\partial \tau}{\partial t} = 1, \tag{57}$$

where (4) was used for the partial derivative $\partial\xi/\partial t$. Similarly,

$$\frac{\partial x}{\partial \xi} = \frac{1}{ch}, \tag{58}$$

$$\frac{\partial x}{\partial \tau} = u, \tag{59}$$

$$\frac{\partial t}{\partial \xi} = 0, \tag{60}$$

$$\frac{\partial t}{\partial \tau} = 1, \tag{61}$$

Now consider $u = u(t(\tau,\xi), x(\tau,\xi))$:

$$\begin{aligned}\frac{\partial u}{\partial \tau} &= u_t t_\tau + u_x x_\tau \\ &= u_t + u_x u \\ &= F \qquad\qquad \text{from (5).}\end{aligned}$$

This is (9).

Following the same pattern, we find

$$\begin{aligned}\frac{\partial u}{\partial \xi} &= u_t t_\xi + u_x x_\xi \\ &= \frac{u_x}{ch} \\ &= -\frac{1}{(ch)^2}\left[-(ch)u_x\right] \\ &= -\frac{1}{(ch)^2}\left[(ch)_t + (ch)_x u\right] \qquad \text{from (4)} \\ &= -\frac{1}{(ch)^2}\left[(ch)_t t_\tau + (ch)_x x_\tau\right] \\ &= \left(\frac{1}{ch}\right)_\tau .\end{aligned}$$

This yields (8).

A.2. Solution to the variational problem. The variational problem posed in section 4 is

$$\delta \int \left\{(\tilde{u}-u)^2 + \lambda(x)(k + \Delta t\, \tilde{u}_x)\right\}\, dx = 0.$$

Equivalently,

$$\begin{aligned}\forall\, f \in C_o^\infty, \quad 0 &= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \int \left[(\tilde{u}+\epsilon f - u)^2 + \lambda\left(k + \Delta t\, \frac{\partial}{\partial x}(\tilde{u}+\epsilon f)\right)\right]\, dx \\ &= \int \left[2\tilde{u}f - 2uf + \Delta t\, \lambda f_x\right]\, dx \\ &= \int f\left[2\left(\tilde{u}-u\right) - \Delta t\, \lambda_x\right]\, dx,\end{aligned}$$

which implies that

$$2\left(\tilde{u}-u\right) = \Delta t\, \lambda_x.$$

Acknowledgments. We would like to thank David Holland for many fruitful conversations on the subject of sea ice dynamics, and Oliver Bühler for stimulating suggestions. We are also grateful to the anonymous reviewer for the valuable comments.

REFERENCES

[1] T. E. Arbetter, J. A. Curry, and J. A. Maslanik, *Effects of rheology and ice thickness distribution in a dynamic-thermodynamic sea ice model*, J. Phys. Oceanogr., 29 (1999), pp. 2656–2670.
[2] G. M. Flato and W. D. Hibler, III, *Modeling pack ice as a cavitating fluid*, J. Phys. Oceanogr., 22 (1992), pp. 626–651.
[3] W. D. Hibler, III, *A dynamic thermodynamic sea ice model*, J. Phys. Oceanogr., 9 (1979), pp. 817–846.
[4] M. A. Hopkins, *On the mesoscale interaction of lead ice and floes*, J. Geophys. Res., 101 (1996), pp. 18315–18326.
[5] E. C. Hunke and J. K. Dukowicz, *An elastic-viscous-plastic model for sea ice dynamics*, J. Phys. Oceanogr., 27 (1997), pp. 1849–1867.
[6] C. F. Ip, W. D. Hibler, III, and G. M. Flato, *On the effect of rheology on seasonal sea-ice simulations*, Ann. Glaciol., 15 (1991), pp. 17–25.
[7] M. Kreyscher, M. Harder, P. Lemke, and G. M. Flato, *Results of the sea ice model intercomparison project: Evaluation of sea ice rheology schemes for use in climate simulations*, J. Geophys. Res., 105 (2000), pp. 11299–11320.
[8] J. E. Overland and C. H. Pease, *Modeling ice dynamics of coastal seas*, J. Geophys. Res., 93 (1988), pp. 15619–15637.
[9] H. Schaffrin, *An Optimization Approach to Sea Ice Dynamics*, Ph.D. thesis, Department of Mathematics, Courant Institute of Mathematical Sciences, New York University, New York, NY, 2005.
[10] A. S. Thorndike, D. A. Rothrock, G. A. Maykut, and R. Colony, *The thickness distribution of sea ice*, J. Geophys. Res., 80 (1975), pp. 4501–4513.
[11] L.-B. Tremblay and L. A. Mysak, *Modeling sea ice as a granular material, including the dilatancy effect*, J. Phys. Oceanogr., 27 (1997), pp. 2342–2360.
[12] B. Zhang, *On a local existence theorem for a simplified one-dimensional hydrodynamic model for semiconductor devices*, SIAM J. Math. Anal., 25 (1994), pp. 941–947.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 561–581

AN OPTIMIZATION APPROACH TO MODELING SEA ICE DYNAMICS, PART 2: FINITE ICE STRENGTH EFFECTS*

HELGA S. HUNTLEY† AND ESTEBAN G. TABAK‡

Abstract. The effects of a finite ice strength on a new model for sea ice dynamics, deriving the internal pressure field from a global optimization problem, rather than a local rheology, are examined. Building on the promising results from the one-dimensional Lagrangian model described previously, here we add one of the key properties of sea ice. In order to investigate the behavior of the model under ice yielding, the equations are cast in an Eulerian framework, now allowing for variable thickness. The model is first tested under conditions of infinite ice strength, to ensure that the numerics behave as desired. A finite ice strength is incorporated into the model as a second optimization step, minimizing the change in ice thickness necessary to satisfy the upper bound on the pressure, whereby ice strength is taken to be a linear function of thickness, following typical parameterizations in the literature. The theory is implemented numerically, and several test cases are discussed, which show good agreement with physically based expectations.

Key words. ice dynamics, rheology, fluid dynamics, ice yielding

AMS subject classifications. 76M30, 86A05

DOI. 10.1137/060668651

1. Introduction. Much progress has been made in the field of sea ice dynamics modeling over the last half a century, as models have evolved from using free drift to incorporating complex rheologies, derived from various physical considerations. Nonetheless, some salient features of the polar ice covers (as, for example, sea ice arches in the straits of the Canadian Arctic Archipelago) are still not being reproduced satisfactorily. With the ultimate goal of remedying this shortfall, we are developing a novel method for modeling the dynamics.

In Part 1 of this study [6], we introduced a new way to calculate the internal stress arising in converging sea ice; here we discuss how a finite ice strength can be incorporated into the model. The theory was developed in a Lagrangian frame of reference. This simplified the equations to a degree that it was possible to verify the numerical model results by comparing them to an analytic solution to a well-understood toy problem. Beginning with the analogy that coarse grained sea ice can be described as a semi-incompressible fluid (i.e., a fluid that is always allowed to diverge, but can converge only if the ice strength is insufficient to stop the motion), it was argued that the problem of finding the internal stress can be phrased as an optimization problem, where the pressure plays the role of a Lagrange multiplier.

In order to be able to carry out a relatively straightforward qualitative analysis of the results, we make (in both Parts 1 and 2) several simplifications to the full ice dynamics problem. Thus, thermodynamic effects are ignored. We also do not incorporate a thickness distribution. Thorndike et al. [11] and subsequent studies argued

*Received by the editors December 20, 2004; accepted for publication (in revised form) August 30, 2006; published electronically February 15, 2007. This article derives from work done as part of the Ph.D. thesis of Helga Schaffrin Huntley, while partially supported by the NSF VIGRE program.
http://www.siam.org/journals/siap/67-2/66865.html

†Department of Applied Mathematics, University of Washington, Box 352420, Seattle, WA 98195-2420 (helga@amath.washington.edu).

‡Department of Mathematics, Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., New York, NY 10012 (tabak@cims.nyu.edu). This author's work was partially supported by grants from the NSF Division of Applied Mathematics.

well for the importance of and provided a method for tracking such a distribution. It is particularly relevant for thermodynamics, as the freezing and melting properties of thin ice and thick ice differ significantly. However, since ice strength is also a function of ice thickness, resolving subgridscale variations is desirable even in a purely dynamic model. Realizing the resulting limitations, we ignore these effects. Similarly, we do not employ a velocity distribution, but consider all velocities mass-averaged over a grid cell. As the work reported here is intended as a feasibility study, we also set all external forces equal to 0. (They are easily added back into the dynamics, as will be described below, but complicate the analysis of the results.) Finally, we restrict ourselves to one dimension, where only isotropic stress exists. How to handle shear stresses will be addressed in subsequent work. The assumption from Part 1 that is dropped here is that of constant ice thickness throughout the domain, in order to allow crushing of the ice.

The appealing simplicity of the equations derived in the Lagrangian framework in Part 1 resulted in part from the assumption of a constant ice thickness. Permitting the thideness to vary over the domain and especially for individual floes (or their grid-averaged equivalent) due to yielding negates the advantages of the formulation we used previously. Thus, following a quick review of the Lagrangian model in section 2, we will here return to an Eulerian point of view. The new theory is presented, and the corresponding numerical model, with variable ice thickness but still infinite ice strength, is compared to the Lagrangian one of Part 1 in sections 3 and 4. In section 5, then, ice strength is limited, and ice is allowed to yield. For the parameterization of ice strength, we rely on suggestions from the literature. (Note that, as we are not carrying out quantitative studies or direct checks against data, the *form* of the parameterization is more important than the exact empirical—or tuned—values used.) It turns out that employing the limiting ice strength as a truncation value for the internal pressure directly leads to undesirable effects. Hence, we instead reformulate the problem in terms of a double optimization, where the ice strength becomes a new constraint on the pressure. Some numerical results are discussed in section 6, and conclusions presented in section 7. The appendix provides some of the mathematical details of the model derivation; for greater detail on some of the other results, the reader is referred to [10].

2. Lagrangian dynamics. We will use the same notation as in Part 1 [6]. In particular, the variables are defined as follows:

c = concentration of ice (fraction of sea surface area covered by ice),
h = thickness of the ice, averaged over a grid box,
u = horizontal velocity,
F = sum of all forces under consideration,
p = internal stress.

Recall that we are not modeling any thermodynamic effects, so that all sources and sinks of ice (melting, freezing, and precipitation) are set to 0. Similarly, no forces other than the internal stress are considered in this paper, so that F denotes the force due to the internal stress. The density of ice, which is nearly constant, is taken to be identically 1. Alternatively, one can consider it as absorbed in h, which then represents the product of ice thickness and density or, in effect, an areal density. The authors find it most useful, however, to continue thinking of h as thickness.

The Eulerian spatial coordinate will be denoted by x, while the Lagrangian mass coordinate is given by

$$\xi = \int_0^x ch\, d\hat{x}. \tag{1}$$

For the Lagrangian formulation, we also define the new variable

$$k = \frac{1}{ch} - 1,$$

which simplifies the form of the governing equations.

The mass and momentum conservation equations in one dimension are given, in Eulerian coordinates, by

$$(ch)_t + (chu)_x = 0, \tag{2}$$

$$(chu)_t + (chu^2)_x = F. \tag{3}$$

We will return to these in the next section.

Translating the system into the Lagrangian coordinate defined above in (1) and substituting the new variable k yields

$$k_t = u_\xi, \tag{4}$$

$$u_t = \tilde{F}, \tag{5}$$

where $\tilde{F} = F/(ch)$ is still unknown.

We argued in Part 1 [6] (where ice strength was taken to be infinite) that the internal stress arises solely due to the semi-incompressibility of the ice. It serves the purpose of preventing further convergence when $c = 1$, or, in other words, of enforcing the constraint that $c \leq 1$, equivalently that $k \geq 0$ (for h constant and taken to be 1). Consequently, it was suggested that the problem of finding F amounts to a constrained optimization problem, which can be solved using Lagrange multipliers. The pressure arose naturally as such a multiplier. It was shown that for the case without ice yielding, minimizing $\|p\|$ is equivalent to minimizing $\|p_\xi\|$. So the system was ultimately phrased as follows:

$$\text{minimize } \|p\|, \tag{6}$$

$$\text{subject to the constraints } k_t = u_\xi, \tag{7}$$

$$u_t = -p_\xi, \tag{8}$$

$$k \geq 0, \tag{9}$$

$$p \geq 0. \tag{10}$$

As the numerical model arising from this theory behaved well under the verification tests (comparison to an analytic solution, comparison with a particle-resolving model, and qualitative assessment of other model runs), we concluded that this approach to sea ice dynamics is promising.

The analysis carried out in Part 1 [6], however, relied on the unrealistic assumptions of constant ice thickness and infinite ice strength. This paper is intended to carry the model one step further by eliminating those simplifications.

3. Translation to Eulerian coordinates. Allowing variable ice thickness in the Lagrangian model outlined in the previous section leads to some difficulties. For one, the constraint on k becomes the rather cumbersome $k \geq \frac{1-h}{h}$, versus the simple $k \geq 0$. In fact, this constraint is not even well defined if we let $h = 0$. Of course, in this case, c is also somewhat arbitrarily defined, but it is preferable to work with a variable that is not occasionally constrained to be greater than or equal to infinity. It should also be noted that the Lagrangian mass coordinate distorts the resolution in favor of thick ice. In the discretized optimization, then, minimizing the pressure in thicker parts weighs heavier than doing so for thinner parts, which distorts the pressure field as well. These observations argue strongly for proceeding in Eulerian coordinates. In addition, it is unclear whether the mass coordinates can be extended into a second dimension while retaining any of the desired simplifications. Thus, we return to Eulerian coordinates for studying the effects of a finite ice strength.

Translating the system (6)–(10) into Eulerian coordinates results in

$$\text{minimize } \|p\|, \tag{11}$$

$$\text{subject to the constraints } (ch)_t + (chu)_x = 0, \tag{12}$$

$$(chu)_t + (chu^2)_x = -p_x, \tag{13}$$

$$0 \leq c \leq 1, \tag{14}$$

$$0 \leq p. \tag{15}$$

As a side note, if one had decided not to ignore external forces, these could easily be added to the left-hand side of (13). Everything that follows could be carried out as described here (except, of course, that a discretization would have to be found for the additional terms).

The reader will notice that even once p is determined by the optimization, there are but two equations for the three unknowns c, h, and u. Let us continue to assume infinite ice strength for now, before adding this additional complexity in the next section. In this case, since ice does not crush, it is reasonable to take ice thickness to be conserved following ice floes, or, in other words, h to be advected:

$$h_t + u\,h_x = 0. \tag{16}$$

Using the mass conservation equation (12), we can rewrite this as

$$c_t + (cu)_x = 0. \tag{17}$$

This equation is then added as an additional constraint to the optimization problem (11)–(15).

For the numerical implementation, we chose to rewrite the momentum conservation equation as a velocity evolution equation:

$$u_t + \left(\frac{u^2}{2}\right)_x = -\frac{p_x}{ch}. \tag{18}$$

Retaining momentum (chu) as a fundamental variable to be updated each time step is attractive, since it can then be exactly conserved. However, a preliminary numerical implementation following this approach proved to be far more prone to numerical instabilities than one updating velocity explicitly. (It may be worth mentioning that velocity is updated directly in many of today's ice dynamics models; see, e.g., [4] and [5].)

Note that this derivation is valid only if $c \neq 0$. However, wherever $c = 0$, the velocity is intrinsically not well defined as a physical quantity. Defining the ratio on the right-hand side of (18) to be 0 in this case and evolving u accordingly provides for a solution for u to fill in the gaps. This particular choice has the advantage that regions without ice do not have to be treated separately in the numerical implementation. At first glance, one might expect that defining the ratio on the right-hand side to be 0 when $c = 0$ might lead to strange discontinuities in the forcing, since $\lim_{c\to 0} 1/c = \infty$. Yet there is no inconsistency, since the pressure is also zero whenever c is far from 1. The situation with $h = 0$ can be handled similarly. In this case, both u and c are arbitrary, so that c can be taken far from 1.

While advancing momentum as a fundamental variable leads to undesirable effects, we do retain mass (ch) as a fundamental variable, so that it can be conserved exactly. Thickness h never appears explicitly in the equations. It is tracked as a derived variable.

A staggered grid is used, defining u at half steps from c, ch, and p. This follows the setup for the Lagrangian model. Unlike in that case, however, it is here unfortunately not possible to avoid all need for interpolation of any variables. The placement of the variables was chosen to allow for a semi-implicit discretization for (17) (necessary to be able to satisfy the constraint on c) and to keep mass and concentration in the same locations for deriving thickness.

The discretizations of (17), (12), and (13) take the form

$$c_{j+\frac{1}{2}}^{n+1} = c_{j+\frac{1}{2}}^{n} - \frac{\Delta t}{\Delta x}\left(c_{j+1}^{n}\, u_{j+1}^{n+1} - c_{j}^{n}\, u_{j}^{n+1}\right), \tag{19}$$

$$(ch)_{j+\frac{1}{2}}^{n+1} = (ch)_{j+\frac{1}{2}}^{n} - \frac{\Delta t}{\Delta x}\left((ch)_{j+1}^{n}\, u_{j+1}^{n+1} - (ch)_{j}^{n}\, u_{j}^{n+1}\right), \tag{20}$$

$$u_{j}^{n+1} = u_{j}^{n} - \frac{1}{2}\frac{\Delta t}{\Delta x}\left[\left(u_{j+\frac{1}{2}}^{n}\right)^2 - \left(u_{j-\frac{1}{2}}^{n}\right)^2\right] - \frac{\Delta t}{\Delta x}\frac{1}{(ch)_{j}^{n}}\left[p_{j+\frac{1}{2}}^{n+1} - p_{j-\frac{1}{2}}^{n+1}\right]. \tag{21}$$

Observe that (20) serves only to update the variable (ch); it does, in fact, not represent an additional constraint in the optimization. Thus, the linear constraint for the minimization of $\|p\| = \sum p_i$ is given by (19), where (21) is used to substitute for u^{n+1}. The optimization provides updated values of p and c, which are then used to find updated values of u and subsequently of (ch). The variables are interpolated where necessary by a Godunov-type scheme. (See the appendix for more details.) The results presented here were obtained from an implementation on Matlab, using the built-in "`linprog`" function for the minimization, which uses a linear interior point solver. (Alternatively, a simplex method can be prescribed. The resulting differences in the output are negligible.) Also, periodic boundary conditions were imposed.

4. Results in Eulerian coordinates without yielding. A series of tests was carried out on this model to check its behavior under various initial conditions. Here we will present only two of the results. First, we will compare the behavior of this Eulerian model to that of the Lagrangian one from Part 1 [6]. The equations and hence the numerics are more complicated here, requiring the choice of an interpolation scheme in addition to a discretization. Moreover, it can be shown that, with the choices made here, in general,

$$\sum_j (ch)_j^{n+1} u_j^{n+1} - \sum_j (ch)_j^n u_j^n \neq 0.$$

In other words, momentum is not exactly conserved. It is, thus, desirable to check that the deviations are acceptably small.

For the comparison, we initialize both models with a sinusoidal initial velocity, constant ice thickness ($h_o = 1$), and uniform ice concentration. Since spatial and temporal resolutions are fixed throughout the runs but space is measured in different coordinates, it is not possible to retain equivalent spatial step sizes in the two models. In particular, the specifications are as follows:

$$\begin{aligned} c_o &= 0.5, & k_o &= 1, \\ u_o &= \sin(2\pi x), & u_o &= \sin(4\pi\xi), \\ \Delta x &= 0.0125, & \Delta\xi &= 0.0125, \\ \frac{\Delta t}{\Delta x} &= 0.1, & \frac{\Delta t}{\Delta\xi} &= 0.1. \end{aligned}$$

(Recall that k was defined as $k = (1 - ch)/ch$ and $\xi = \int_0^x ch\, d\tilde{x}$.)

For the plots, we have chosen four times:

$t = 0.0175$ fairly early on, after 15 time steps,

$t = 0.0500$ shortly before consolidation begins, after 41 time steps,

$t = 0.0900$ shortly after consolidation begins, after 73 time steps,

$t = 0.1500$ about 2/3 through this run, after 121 time steps.

At each of these times, we have converted the Lagrangian spatial coordinate to the Eulerian one for a direct comparison. Figures 1 and 2 display the concentrations and velocities from the two models corresponding to these times. Figure 3 shows the momenta.

The two models show good agreement in both concentration and velocity; the lines coincide almost exactly. The differences are mostly due to resolution, as different points are resolved in each model. The same is true for the momentum plots. Here the Eulerian model produces somewhat greater values for both the positive and the negative velocities near the consolidated region, which may be related to the interpolation. (The momentum is the product of two quantities that are tracked in different spatial locations. Thus, even for the Lagrangian results, interpolation is necessary to calculate it.) Recall that the Lagrangian model was designed to conserve momentum exactly. The discretization chosen for the Eulerian model, on the other hand, does not do the same. (Of course, in the limit as the resolution becomes finer, this error vanishes.) The fact that the momentum profiles from the Eulerian model agree so well is evidence that momentum is close to being conserved here as well. Further support comes from the fact that the constant velocity at which the consolidated ice travels is 0 (up to four decimal places), just as predicted by the theory based on momentum conservation. (This velocity test, by the way, holds up in other examples as well, although the accuracy degrades somewhat to only 10^{-2} when the average velocity is not 0. Some of this loss in accuracy can be ascribed to numerical diffusion.)

Before the ice begins to consolidate, the numerical solutions can also be compared to exact analytic solutions. Both the Lagrangian and the Eulerian model produce very good approximations (not shown). For the latter, the type of interpolation chosen can make a noticeable difference. (Again, see the appendix for a description of the scheme we use.) The first accumulation time (when c first reaches 1 and p becomes nonzero) is theoretically predicted in this example to be $1/4\pi \approx 0.079577$. The model here predicts it to be in the interval $[0.07875, 0.08000]$; i.e., it captures it very well.

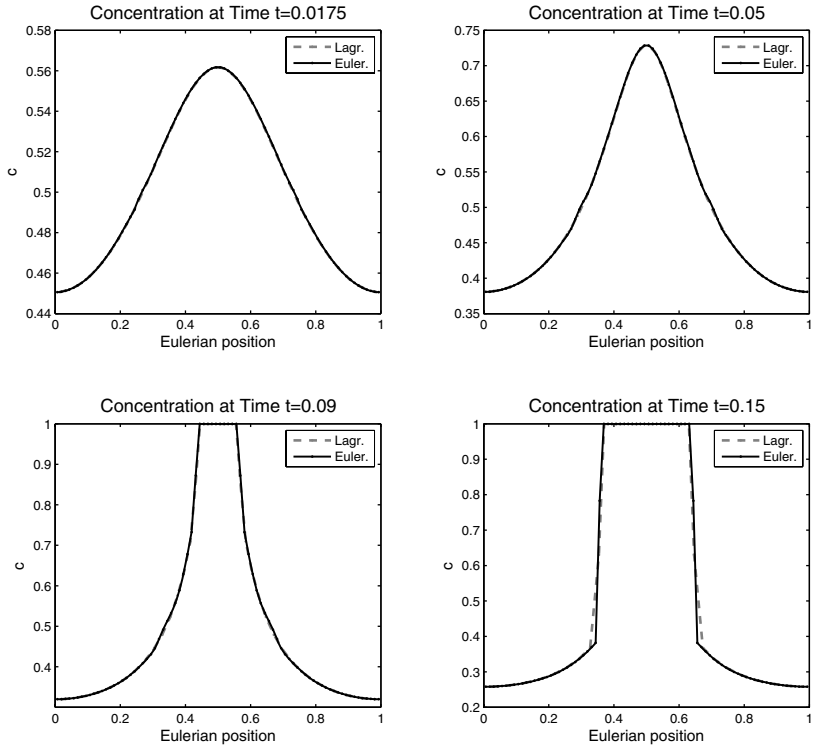


FIG. 1. *Comparing ice concentrations at various times from the Lagrangian and the Eulerian models. The Eulerian output is interpolated with a solid black line, the Lagrangian with a dashed grey line.*

To test how the model handles variable thickness (which, after all, was the primary purpose for it), we ran an example with a parabolic initial thickness distribution, while using again the sinusoidal initial velocity and uniform initial concentration. Figure 4 beautifully exhibits the behavior. As the ice is pushed together in the center of the domain, the thickness profile steepens. Once concentration reaches 1, h no longer changes. This agrees well with expectations from physical intuition.

5. Incorporating finite ice strength. Now that we have a model allowing for variable ice thickness, we can introduce the process of ice yielding.[1] It is well known that sea ice has a finite strength, which depends on ice thickness. It may also depend on other properties, such as the age of the ice or the salinity of the water from which it was formed. These variables are not being tracked in this model. The literature, moreover, seems to agree generally that thickness is the most important factor.

[1] We will use the terms "yielding," "crushing," and "ridging" interchangeably, although the actual processes of crushing and ridging may be quite different. Ridging typically occurs when ice floes slide under each other, while crushing implies that the ice actually breaks. The net result of both processes is thicker ice, as a consequence of yielding. It is generally accepted that resistance to crushing is significantly higher than that to ridging. Hence, most of the yielding accounted for here will technically be ridging rather than crushing.

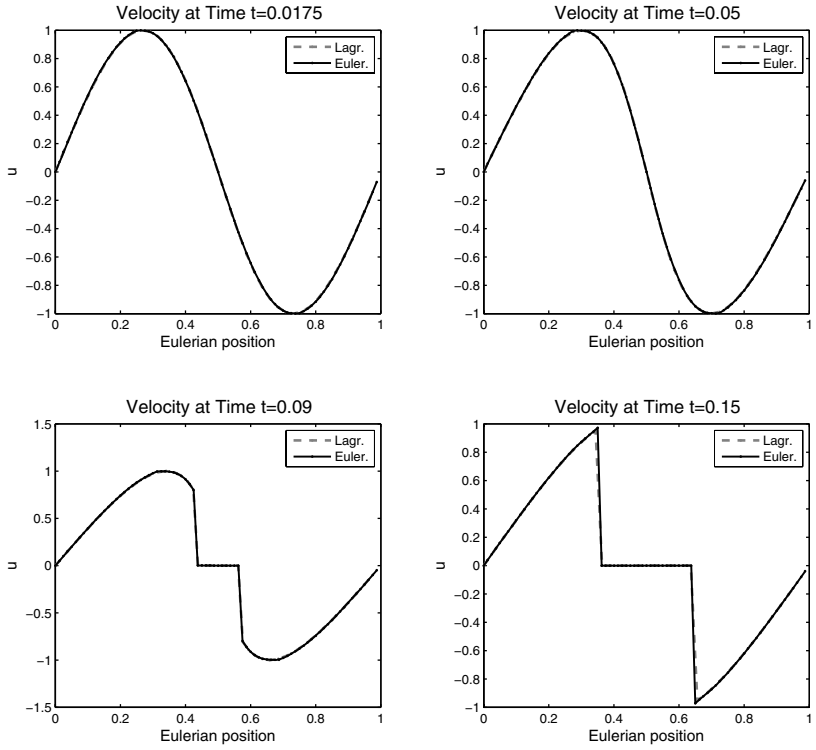


FIG. 2. *Comparing ice velocities at various times from the Lagrangian and the Eulerian models. The Eulerian output is interpolated with a solid black line, the Lagrangian with a dashed grey line.*

Ice dynamics models currently in use do not typically include dependence on anything but ice thickness and concentration.

The importance of a failure criterion for sea ice dynamics has been recognized at least since the mid-1970s (see [1]). It has been incorporated into plastic (e.g., [8]), viscous-plastic (e.g., [4]), elastic-plastic (e.g., [2] and [9]) and elastic-viscous-plastic (e.g., [5]) rheologies, but it has also appeared in other descriptions of the constituency law for ice, such as the cavitating fluid (e.g., [3]) and granular flow (e.g., [12]) rheologies.

The literature offers essentially two types of ice strength parameterizations. On the one hand is a formulation relating ice strength P^* to the potential energy change associated with the changes in ice thickness. This was first suggested by [9], was adopted by [11], and has been used ever since in connection with the thickness distribution theory pioneered by the latter.

For simpler models, using a two-category thickness distribution, distinguishing within a grid box only between ice and open water (the type employed here), ice strength has uniformly been taken as a function of ice thickness and concentration, although the functional dependence is not always the same. The other factors that may influence ice strength (such as age) are generally not taken into account.

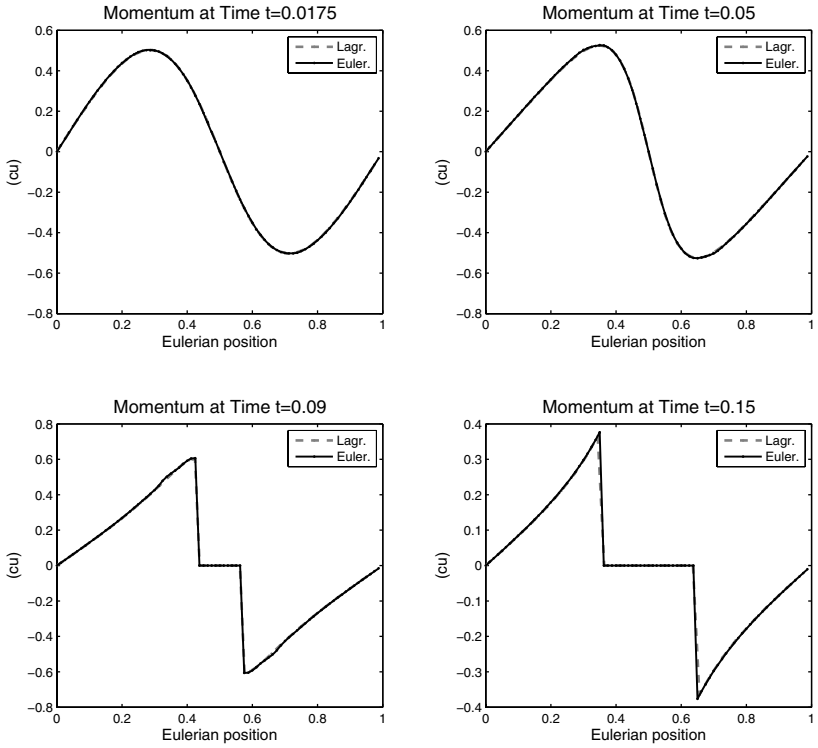


FIG. 3. *Comparing ice momenta at various times from the Lagrangian and the Eulerian models. The Eulerian output is interpolated with a solid black line, the Lagrangian with a dashed grey line.*

Most models follow the formulation presented by [4]:

$$P^* = P_o c h e^{-b(1-c)}, \tag{22}$$

where P^* is the ice strength expressed as a maximal pressure and P_o and b are empirical parameters.[2] This parameterization exhibits several of the desired properties. The thicker the ice, the stronger it is. The greater the concentration, the greater is the total strength. However, it is also apparent that even if c is relatively far from 1 (say 0.6 or so), the ice strength is still significant.

Overland and Pease [8] suggest an alternative. Instead of a linear dependence on h, a quadratic law is hypothesized, which leads to a better approximation in their study of observed ridging:

$$P^* = P_o \rho c h^2 e^{-b(1-c)}. \tag{23}$$

The only other difference is that they include a dependence on ice density, which is typically (as in this work) taken to be a constant and can hence be absorbed into P_o. The functional form (22) has proven to be the more popular of these two.

[2]Hibler's notation in the 1979 paper [4] is actually $P = P^* h \exp[-C(1-A)]$. The formulation here reflects the definition of the variables we are using throughout. Note also that Hibler's h is an *effective* ice thickness, i.e., more akin to ch in our notation.

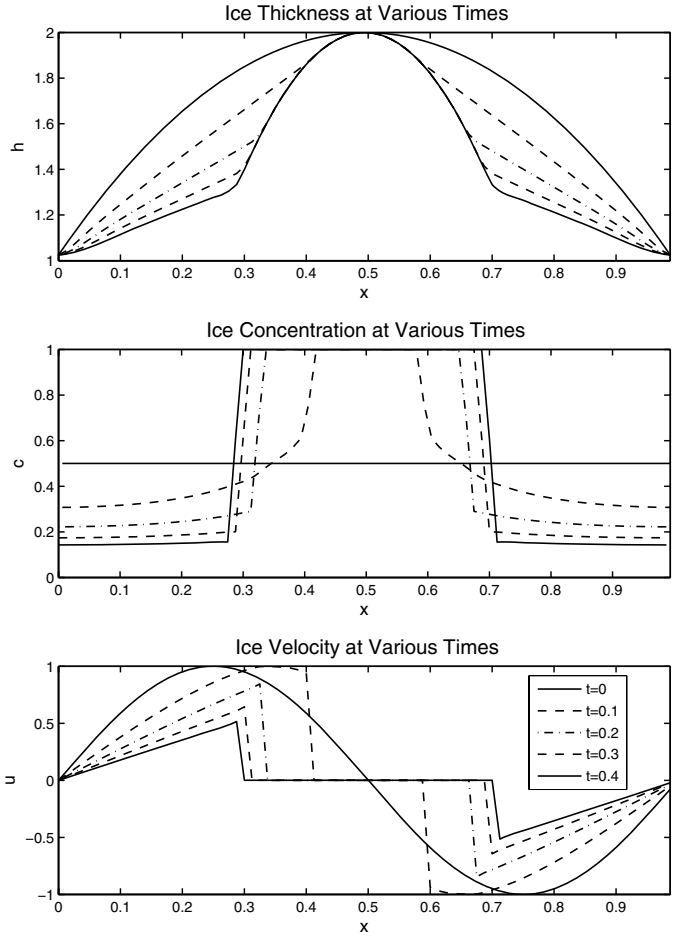


FIG. 4. *The Eulerian model was run with spatial step size set to* 0.0125 *and temporal step size set to* 0.00125. *Initial conditions were* $c_o = 0.5$, $h_o = -4x^2 + 4x + 1$, *and* $u_o = \sin(2\pi x)$. *The evolution of* h, c, *and* u *is illustrated in parallel with snapshots from five different times.*

For our purposes, either description would work; we have chosen a linear dependence of ice strength on ice thickness, mirroring (22). However, we do not include the dependence on ice concentration. One of the underlying assumptions of our model is that $p = 0$ whenever $c \neq 1$. Hence ice strength is irrelevant (or could be taken to be 0: the ice does not resist convergence) in this situation. Setting $c = 1$ in (22), the parameterization becomes simply

$$P^* = P_o h. \tag{24}$$

Flato and Hibler found empirical values for P_o (and the b appearing in the exponential term of (22)) in their 1992 paper [3]. We are not, at present, concerned with

producing *quantitatively* correct results and will hence adjust P_o to assist in investigating the effect of a limiting pressure value, rather than using their recommendation or an otherwise physically justified value.

Initially, the limiting ice strength was treated as a truncation value for the pressure. p was calculated as before as the solution to the constrained minimization problem (11) with constraints (13)–(15) and (17). Then it was truncated by P^*. The resulting values for p were used to calculate the updated values for u and (ch) using (14) and (12). Equation (17) no longer holds where ice begins to yield. Thus, it is valid only in regions without crushing. However, since ice only yields where it is consolidated, we know that everywhere else $c = 1$.

This procedure is not advisable. Since the limiting values of p depend on h, the truncated p can have undesirable shapes. In particular, a convex thickness can lead to a convex pressure, which causes the ice to diverge artificially. The error in this procedure lies in the assumption that ice yielding in one area of a consolidated region does not change the pressure anywhere else. This, however, is not necessarily true.

The algorithm we would like to suggest here hence rephrases the optimization problem, instead of attempting to correct the pressure profile after the minimization. The finite ice strength is, in fact, an additional constraint for the optimization. On the other hand, we lose the constraint (17), since we can no longer assume advection of thickness. This poses problems, since we need to be able to say something about the advanced c, in order to constrain it to stay between 0 and 1.

Observe that as long as the ice is not allowed to yield, $\frac{Dh}{Dt} = h_t + u\,h_x = 0$; ice thickness does not change following particles. When ice does buckle, it does so only to the degree necessary to satisfy the upper bound on the internal pressure. One can argue that h will change as little as possible—suggesting a second constrained optimization.

The new procedure then is as follows:

(1) Minimize the changes in ice thickness globally:

$$\text{find} \quad \min \left\| \frac{Dh}{Dt} \right\|, \tag{25}$$

$$\text{given the constraints} \quad (ch)_t + (chu)_x = 0, \tag{26}$$

$$(chu)_t + (chu^2)_x = -p_x, \tag{27}$$

$$\frac{Dh}{Dt} \geq 0, \tag{28}$$

$$0 \leq c \leq 1, \tag{29}$$

$$0 \leq p \leq P^*. \tag{30}$$

The first inequality constraint (28) arises because crushing can only increase ice thickness. Equation (26) again is used to advance (ch) rather than as a true constraint on the problem. The norm for (25) is chosen to be the 1-norm, primarily for numerical considerations: This choice keeps the problem linear.

This minimization will give values for c and p, which determine u and h, at the new time level that minimize $\left\| \frac{Dh}{Dt} \right\|$. However, the answer is typically not unique. The only unique quantity is the $\min \left\| \frac{Dh}{Dt} \right\|$.

(2) The first step provides the closure for h, so that it is now possible to proceed

with the pressure minimization:

$$\text{find} \quad \min \|p\|, \tag{31}$$

$$\text{given the constraints} \quad (ch)_t + (chu)_x = 0, \tag{32}$$

$$(chu)_t + (chu^2)_x = -p_x, \tag{33}$$

$$\left\| \frac{Dh}{Dt} \right\| = \text{value found in step (1)}, \tag{34}$$

$$\frac{Dh}{Dt} \geq 0, \tag{35}$$

$$0 \leq c \leq 1, \tag{36}$$

$$0 \leq p \leq P^*. \tag{37}$$

Note that the first inequality (35) needs to be retained in addition to the last equality constraint (34), since it is a pointwise, rather than an integral, statement.

This minimization must have a solution for c and p, from which u and h are derived, satisfying all the constraints, since one was found in the first step.

In the case where no ice yielding occurs, this two-step algorithm simplifies to the procedure outlined in section 3 above, as desired: The minimum of $\|Dh/Dt\|$ is zero. Since Dh/Dt is constrained to be nonnegative pointwise, this implies that, in fact, $Dh/Dt = 0$ everywhere. This is the constraint used in section 3.

The numerical implementation follows the pattern above, for the model without ice yielding. The same grid and discretizations are used. Recall that the variable h is not directly updated at each time step. The objective function for the first minimization is hence rewritten in terms of (ch) and c. It follows from mass conservation (see (26)) that

$$c\left[h_t + uh_x\right] = -h\left[c_t + (cu)_x\right]. \tag{38}$$

Since $c \geq 0$ and $h \geq 0$, the inequality constraint $\frac{Dh}{Dt} \geq 0$ is equivalent to $c_t + (cu)_x \leq 0$. Also, ice only crushes where $c = 1$. In other words, if $c \neq 1$, then $\frac{Dh}{Dt} = 0$. Hence minimizing $\left\| \frac{Dh}{Dt} \right\|$ is equivalent to minimizing $\left\| c\frac{Dh}{Dt} \right\|$ or $\left\| -\frac{(ch)}{c}[c_t + (cu)_x] \right\|$.

6. Results with ice yielding. To illustrate the effects of finite ice strength, we present here the model results from two different problems. First, we look at a somewhat degenerate example, where the initial ice thickness is taken to be constant. (Of course, here it does not remain constant, since P_o is chosen so that the ice does yield.) The initial concentration is also set constant at 0.5, while the velocity is initialized with a sine-curve.

Figure 5 shows the evolution of h, c, and u during the run up to time $t = 0.2$. As expected, the ice begins to consolidate in the center of the domain. Shortly after the concentration reaches 1, the pressure begins to exceed the ice strength, and the ice begins to ridge. Initially, this happens right at $x = 0.5$. Later, the ice yields at the edges of the consolidated region, where it is still thinner (not having ridged there yet), leading to a profile with multiple peaks. This can be seen in Figure 6, which shows several snapshots of the ice thickness from this run.

The attentive reader may have noticed a slight inconsistency in Figure 6: At some times (e.g., $t = 0.08750$ or $t = 0.10500$) the ice thickness actually decreases below its initial value of $h = 1$. This should not happen. On the other hand, these deviations

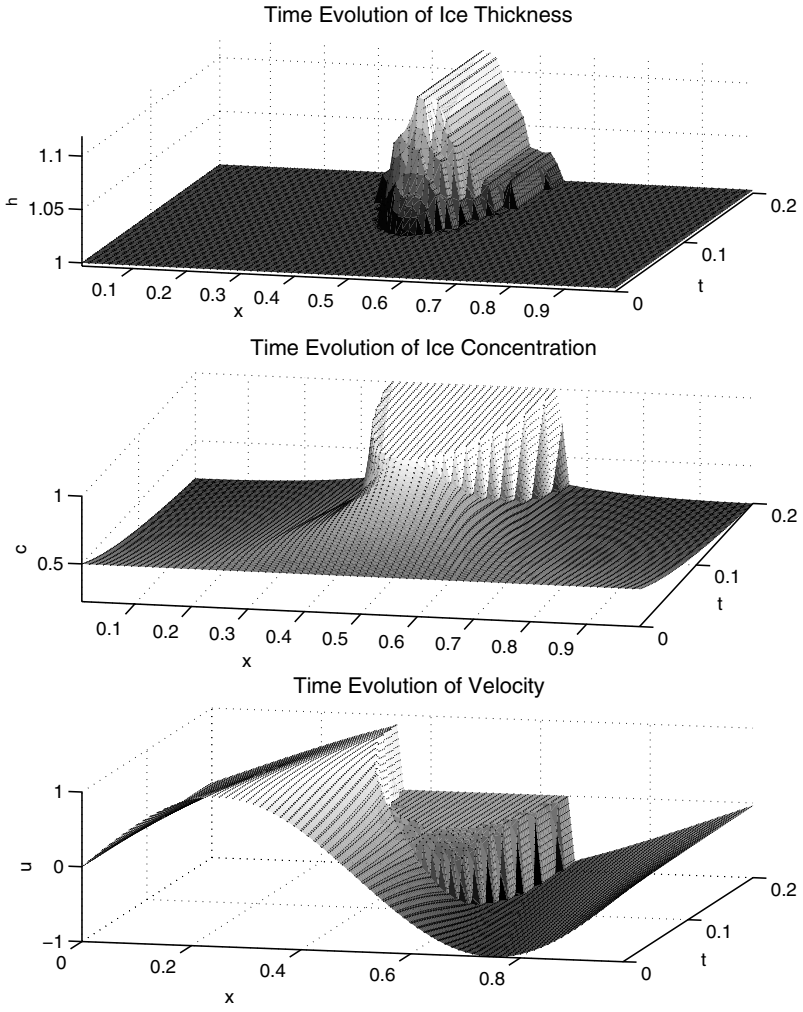


FIG. 5. *The model allowing for crushing was run with spatial step size set to* 0.0125 *and temporal step size set to* 0.00125. *Initial conditions were* $c_o = 0.5$, $h_o = 1$, *and* $u_o = \sin(2\pi x)$. P_o *was set to* 2. *The evolution of* h, c, *and* u *is shown up to time* $t = 0.2$.

are on a small scale, i.e., never greater than about 0.003. Moreover, recall that h is a derived quantity. The locations of the too-small values for h are invariably at the very edge of the consolidated domain, so that it stands to reason that the errors are due to a not quite precise enough capturing of the discontinuity in c. Indeed a higher spatial resolution does improve the situation. Thus, a halving of Δx leads to a reduction of the dips in h below 1 by a whole order of magnitude.

As a second example, we will examine a case where the ice thickness is not uniform, so that it is easier to predict where the ice should yield first. The initial conditions for this run are $c_o = 0.5$, $h_o = \cos(2\pi x) + 1.1$, and $u_o = \sin(2\pi x) + 1.5$. As expected,

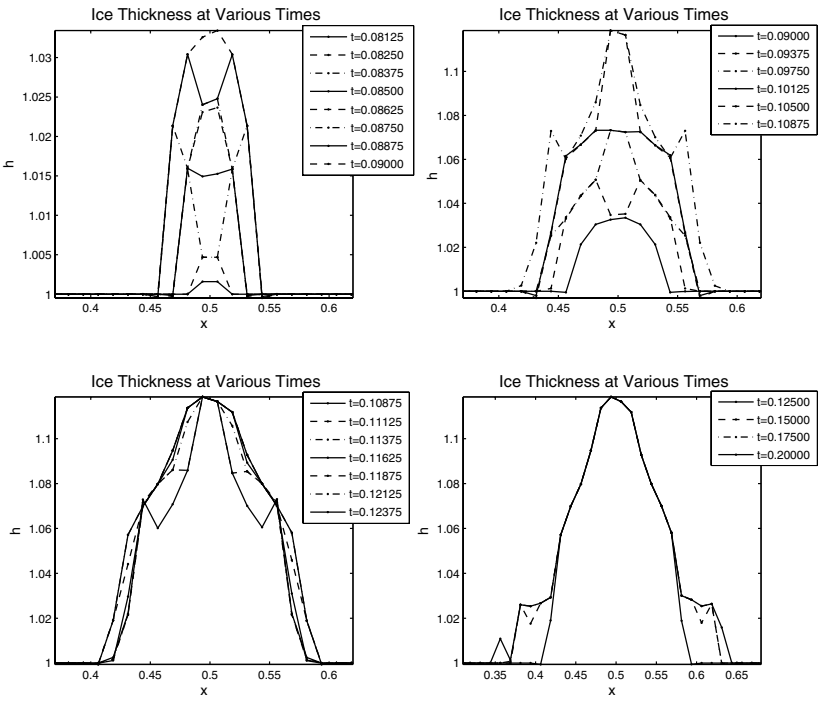


FIG. 6. *Snapshots from the thickness evolution of the run from Figure* 5. *The ice yields initially at the center of the domain, but later at other points within the consolidated region, leading temporarily to profiles with multiple peaks.*

the ice yields first where it is thinnest within the consolidated region. Figures 7 and 8 show the evolution of thickness, concentration, and velocity for this particular experiment. Note that the second figure is translated to the left, so that the center of the plot is at $x = 1$ rather than at $x = 0.5$. (Recall that periodic boundary conditions are being used.)

For a more detailed look, Figure 9 shows h, c, and u in one frame at two distinct times. In the plot on the left, for time $t = 0.2$, one can clearly see that the ice has begun to ridge in the consolidated region. It is also apparent that the velocity for this area is no longer constant; the yielding implies further convergence of the ice despite c equaling 1. In the plot on the right, for time $t = 0.4$, the ice is no longer yielding, and the consolidated ice travels at a uniform velocity of approximately 1.48 (close to the expected 1.5). By this time, the thickness profile within the region of $c = 1$ is again a more or less smooth valley; the thinner parts from the earlier picture have ridged as well.

7. Conclusions. We are investigating the feasibility of a novel formulation of the sea ice dynamics. In Part 1 [6], the method for calculating the internal pressure term as the solution to an optimization problem was derived, and results of a Lagrangian model with infinite ice strength were discussed. In this part, we set out to show how a finite ice strength can be handled by a model of this type. The implementation turns out to be easier in an Eulerian framework. Thus, we translated the formulation,

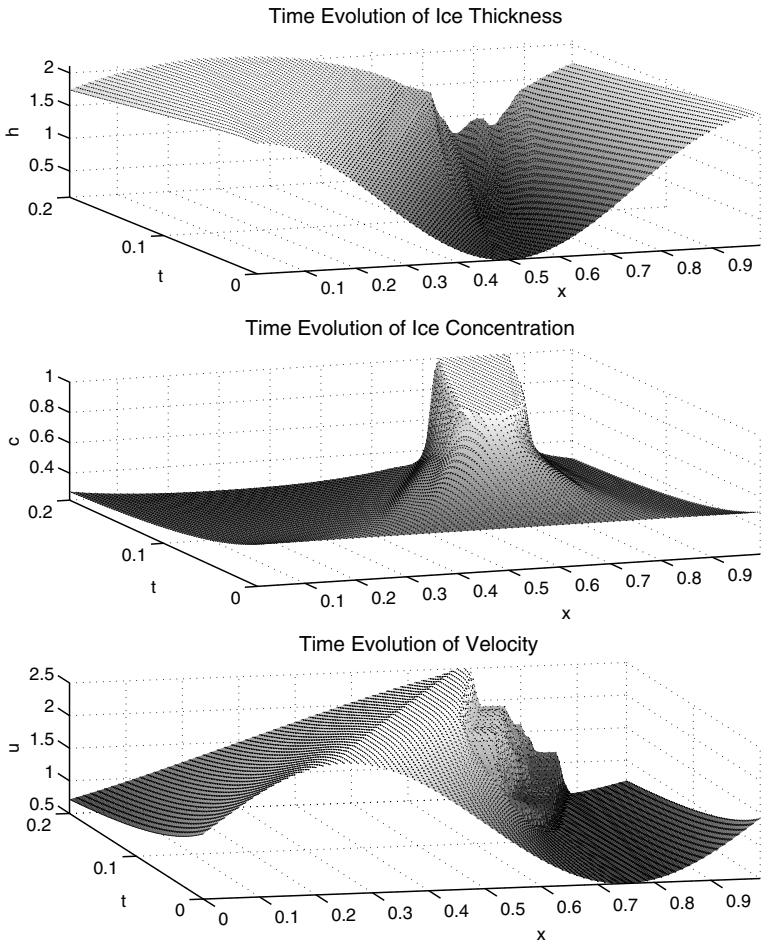


FIG. 7. *The model allowing for crushing was run with spatial step size set to* 0.0125 *and temporal step size set to* 0.00125. *Initial conditions were* $c_o = 0.5$, $h_o = \cos(2\pi x) + 1.1$, *and* $u_o = \sin(2\pi x) + 1.5$. P_o *was set to* 1. *The evolution up to time* $t = 0.2$ *is shown. The ice begins to yield at the weakest point.*

allowing now for a variable ice thickness, although initially still working with an infinite ice strength. This new model was compared to the Lagrangian one, to ensure that the choices made for the numerics, such as which variables to update explicitly at each time step and how to discretize the equations, did not lead to unreasonable results. In particular, we wanted to make sure that momentum, while not exactly conserved by the discretized equations, does not vary significantly over the course of a run. The Eulerian model output was also tested against an exact solution before consolidation, with very good agreement.

To incorporate a finite ice strength, we chose to adopt a parameterization of the ice strength as a multiple of the ice thickness, based on parameterizations common in the ice dynamics literature. Limiting the pressure the ice can withstand then requires

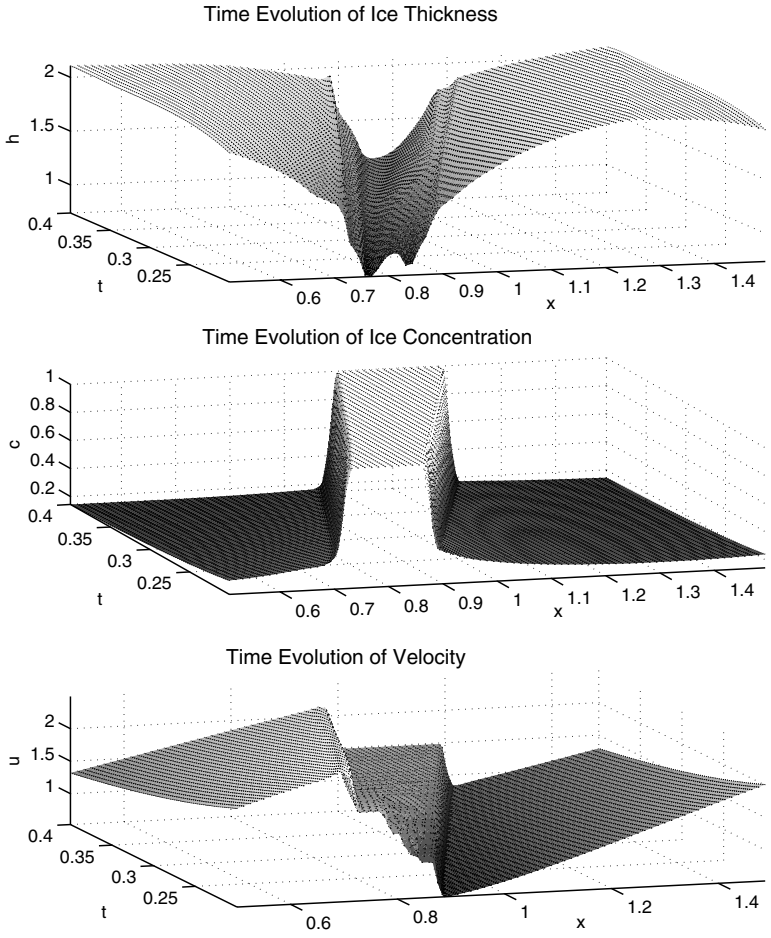


FIG. 8. *The continuation of the plot in Figure 7; here the evolution from time $t = 0.2$ until $t = 0.4$ is shown. Thinner ice yields first, leading to a smooth convex profile for h within the consolidated region at the end of the run.*

a second optimization problem, which determines how much the ice has to yield in order to satisfy this new constraint. While the solution to this minimization tends not to be unique, the minimum found is; this minimum then becomes an additional constraint on the second optimization, minimizing the pressure.

Two example runs were shown from this model, one beginning with a constant ice thickness, the second beginning with a varied one. In both cases, the behavior of the numerical results was qualitatively correct. Ice yields in the consolidated regions, whereby thinner ice tends to yield before thicker ice. Some convergence of the ice occurs during the ridging, but when the ice is thick enough to withstand the pressure exerted by the surrounding ice, all convergence is stopped—as desired.

It can be concluded, thus, that this procedure for determining internal stress for

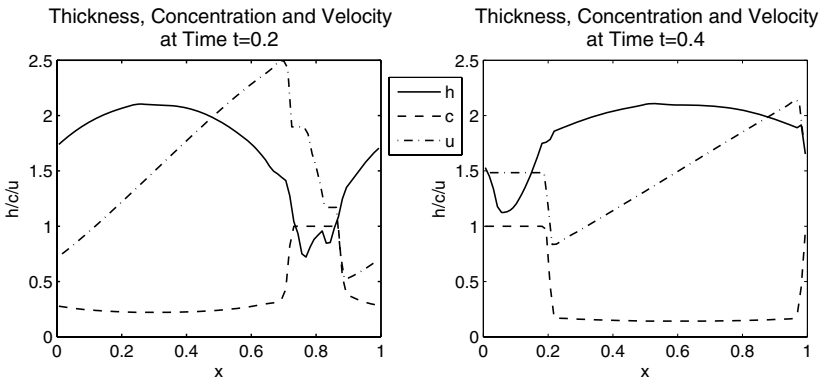


FIG. 9. *Detail for the second experiment, as in Figures* 7 *and* 8. *The left panel shows thickness (solid line), concentration (dashed line), and velocity (dot-dashed line) at time* $t = 0.2$. *The right panel shows the same measures at time* $t = 0.4$.

sea ice is indeed able to handle finite ice strength. Further work on a two-dimensional implementation is underway, where the new challenge of shear stresses and a limited shear strength arises. The findings presented here indicate that the method being investigated produces realistic results in one-dimensional cases, which is prerequisite to creating a working two-dimensional model.

Appendix. Interpolation scheme. As mentioned above, unlike the Lagrangian model, the Eulerian one requires that the variables be interpolated. They are defined on a staggered grid, but for the flux-form of the discretized equations we need to know their values at points in between. The strategy adopted here is a Godunov-type upwinding. See [7] for an extensive discussion of Godunov schemes.

For the concentration c and mass (ch), the upwinding works as follows: u is defined at the interfaces to which each of these is to be interpolated. The sign of the velocity determines the side from which the values are taken. In our implementation, we also perform a linear extrapolation. Thus, since c is defined at half steps of the grid,

$$c_j = \begin{cases} \frac{3}{2}c_{j-\frac{1}{2}} - \frac{1}{2}c_{j-\frac{3}{2}} & \text{if } u_j \geq 0, \\ \frac{3}{2}c_{j+\frac{1}{2}} - \frac{1}{2}c_{j+\frac{3}{2}} & \text{if } u_j < 0. \end{cases} \tag{39}$$

This method generally produces a better estimate of the value at the interface than a constant approximation (i.e., letting c_j equal either $c_{j-\frac{1}{2}}$ or $c_{j+\frac{1}{2}}$). However, one needs to include a safeguard not to overshoot the desired values. Thus, the extrapolations are capped by the minimum and the maximum of the adjoining points. If, for example, the extrapolation predicts a value for c_j larger than $\max\{c_{j-\frac{1}{2}}, c_{j+\frac{1}{2}}\}$, then it is reset to the maximum (similarly for the minimum). The variable (ch) is interpolated according to the same rules.

To find the value of $u_{j+\frac{1}{2}}$ is a somewhat more complex problem. One wants to solve the local Riemann problem for the system of equations

$$c_t + (ch)_x = 0, \tag{40}$$

$$(ch)_t + (chu)_x = 0, \tag{41}$$

$$(chu)_t + (chu^2)_x = -p_x. \tag{42}$$

(We assume no crushing here in between two time steps.) It turns out, however,

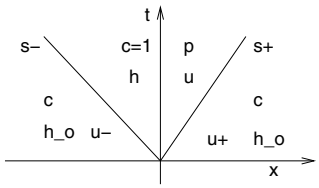


FIG. 10. *The two-shock setup with* $c^- = c^+$ *and* $h^- = h^+ = h_o$.

that this system of equations has only one set of characteristics and only one Riemann invariant. This is insufficient in general. In the case of divergence or contact discontinuities, there is no convergence, and hence we can assume that $p = 0$ is a solution. This observation allows us to simplify the momentum equation (42), using mass conservation (41), to Burgers' equation

$$u_t + \left(\frac{u^2}{2}\right)_x = 0. \tag{43}$$

Classical theory provides a solution to the local Riemann problem here. Thus, if we denote by u^- and u^+ the extrapolated values to the left and to the right, respectively, of the interface, then

$$u_{j+\frac{1}{2}} = \begin{cases} u^- & \text{if } u^+ \geq u^- \geq 0, \\ u^+ & \text{if } u^- \leq u^+ \leq 0, \\ 0 & \text{if } u^- < 0 < u^+. \end{cases} \tag{44}$$

The case of convergence and shocks is a little trickier. In fact, it can be shown that now p cannot be 0. Physical considerations suggest that when ice converges, not one but two discontinuities in the velocity form, one on each side of the newly consolidated ice. The general case for a nonstaggered grid, where concentration, thickness, and velocity all differ from one side of the interface to the other, requires an additional assumption, such as pressure minimization or thickness advection. (See [10] for a detailed discussion of this case.) Luckily, on the staggered grid we have chosen, this is not necessary. The two-shock setup of the local Riemann problem is illustrated in Figure 10. Ice thickness and concentration on each side of the interface (or the shocks, once these develop) are denoted by h_o and c, respectively. The velocities to the left and to the right (the extrapolated values) are denoted by u^- and u^+, as before. In between the shocks, the ice is consolidated; hence $c = 1$, the thickness is h, pressure p, and velocity u.

In the following derivation, we will assume that $0 < c < 1$; the other cases will be treated separately below.[3] For each shock, there are three jump conditions derived from (40)–(42):

$$s = \frac{[\![cu]\!]}{[\![c]\!]} = \frac{[\![chu]\!]}{[\![ch]\!]} = \frac{[\![chu^2 + p]\!]}{[\![chu]\!]}, \tag{45}$$

where s is the shock speed. Using s^- for the shock speed on the left and s^+ for that

[3]This restriction is desirable to ensure that the denominators in the jump conditions are nonzero.

on the right, these six equations become

$$s^- = \frac{cu^- - u}{c-1}, \tag{46}$$

$$\frac{cu^- - u}{c-1} = \frac{ch_o u^- - hu}{ch_o - h}, \tag{47}$$

$$\frac{ch_o u^- - hu}{ch_o - h} = \frac{ch_o (u^-)^2 - hu^2 - p}{ch_o u^- - hu}, \tag{48}$$

$$s^+ = \frac{cu^+ - u}{c-1}, \tag{49}$$

$$\frac{cu^+ - u}{c-1} = \frac{ch_o u^+ - hu}{ch_o - h}, \tag{50}$$

$$\frac{ch_o u^+ - hu}{ch_o - h} = \frac{ch_o (u^+)^2 - hu^2 - p}{ch_o u^+ - hu}. \tag{51}$$

Having six equations to constrain five unknowns (s^-, s^+, u, p, and h) opens the possibility for inconsistency. However, as we will see below, (47) and (50) are, in fact, redundant.

Solving (48) and (51) for p yields

$$p = \frac{ch_o h(u^- - u)^2}{h - ch_o} \tag{52}$$

and

$$p = \frac{ch_o h(u^+ - u)^2}{h - ch_o}. \tag{53}$$

By assumption, $c \neq 0$. One can also assume that $h \neq 0$ and $h_o \neq 0$, since this would mean that there is no ice near this interface and u becomes arbitrarily defined. Hence,

$$(u^- - u)^2 = (u^+ - u)^2. \tag{54}$$

Also, $u^- > u^+$ in the convergence case considered here, so that

$$u^- - u = u - u^+. \tag{55}$$

It follows that

$$u = \frac{u^- + u^+}{2}. \tag{56}$$

Note that one may have predicted this result, namely that the ice masses, once consolidated, move at the average velocity (since ice thickness and concentration are equal on each side).

Using (47), one can solve for h—which, as one might expect, turns out to be equal to h_o. (Since a finite ice strength does not enter this picture, there should not be any yielding or change in thickness.) Equation (50) yields the same answer (showing that one of them is unnecessary).

Substituting expression (56) for u into (46) and (49), we can solve for the two shock speeds:

$$s^- = \frac{u^+ + u^-(1-2c)}{2(1-c)}, \quad s^+ = \frac{u^- + u^+(1-2c)}{2(1-c)}. \tag{57}$$

In order for the picture in Figure 10 to be accurate, we also need that $s^- < s^+$. This is true.

Proof. Since we are only concerned with convergence, $u^- > u^+$. From (56), it can be concluded that $u^- > u > u^+$. Also $0 < c < 1$. It follows that

$$cu^- > cu, \qquad cu > cu^+, \tag{58}$$

$$u - cu^- < u - cu, \qquad u - cu < u - cu^+, \tag{59}$$

$$\frac{u - cu^-}{1 - c} < u, \qquad u < \frac{u - cu^+}{1 - c}, \tag{60}$$

$$s^- < u, \qquad u < s^+, \tag{61}$$

where (46) and (49) were used for the last step. ∎

The interpolation scheme resulting from this analysis is the following:

$$u_{j+\frac{1}{2}} = \begin{cases} u^- & \text{if } u^- \geq u^+, 0 < s^- < s^+, \\ u^+ & \text{if } u^- \geq u^+, s^- < s^+ < 0, \\ u = \frac{u^- + u^+}{2} & \text{if } u^- \geq u^+, s^- \leq 0 \leq s^+. \end{cases} \tag{62}$$

(We have chosen to group the cases $s^- = 0$ and $s^+ = 0$ into the last category, while they could as well fit into the two previous scenarios, respectively.)

Finally, we will say a word about the cases $c = 0$ and $c = 1$. If $c = 0$, there is no ice near the interface, and the velocity of the nonexisting ice is arbitrary. For consistency and smoothness, we will continue to use the same interpolation scheme as above (even though the derivation does not necessarily hold, starting with the observation that the concentration needs to be 1 between the shocks).

If $c = 1$, the ice is consolidated around the interface in question, which means that it should all be traveling at the same speed (in the absence of crushing). In other words, $u^- = u^+$. If this is not the case, due to numerical error, we set

$$u_{j+\frac{1}{2}} = \frac{u^- + u^+}{2}. \tag{63}$$

This completes the description of the interpolation scheme.

Acknowledgment. We would like to thank David Holland for his comments on this work.

REFERENCES

[1] M.C. Coon, *Mechanical behavior of compacted arctic ice floes*, J. Pet. Technol., 26 (1974), pp. 466–470.

[2] M.C. Coon, S.A. Maykut, R.S. Pritchard, D.A. Rothrock, and A.S. Thorndike, *Modeling the pack ice as an elastic-plastic material*, AIDJEX Bull., 24 (1974), pp. 1–105.

[3] G.M. Flato and W.D. Hibler, III, *Modeling pack ice as a cavitating fluid*, J. Phys. Oceanogr., 22 (1992), pp. 626–651.

[4] W.D. Hibler, III, *A dynamic thermodynamic sea ice model*, J. Phys. Oceanogr., 9 (1979), pp. 817–846.

[5] E.C. Hunke and J.K. Dukowicz, *An elastic-viscous-plastic model for sea ice dynamics*, J. Phys. Oceanogr., 27 (1997), pp. 1849–1867.

[6] H.S. Huntley, E.H. Suh, and E.G. Tabak, *An optimization approach to modeling sea ice dynamics. Part* 1: *Lagrangian framework*, SIAM J. Appl. Math., to appear.

[7] R.J. LeVeque, *Numerical Methods for Conservation Laws*, 2nd ed., Birkhäuser Boston, Cambridge, MA, 1992.

[8] J.E. Overland and C.H. Pease, *Modeling ice dynamics of coastal seas*, J. Geophys. Res., 93 (1988), pp. 15619–15637.
[9] R.S. Pritchard, *An elastic-plastic constitutive law for sea ice*, J. Appl. Mech., 42E (1975), pp. 379–384.
[10] H. Schaffrin, *An Optimization Approach to Sea Ice Dynamics*, Ph.D. thesis, Department of Mathematics, Courant Institute of Mathematical Sciences, New York University, New York, NY, 2005.
[11] A.S. Thorndike, D.A. Rothrock, G.A. Maykut, and R. Colony, *The thickness distribution of sea ice*, J. Geophys. Res., 80 (1975), pp. 4501–4513.
[12] L.-B. Tremblay and L.A. Mysak, *Modeling sea ice as a granular material, including the dilatancy effect*, J. Phys. Oceanogr., 27 (1997), pp. 2342–2360.

SIAM J. APPL. MATH.
Vol. 67, No. 2, pp. 582–603

A FAST AND ACCURATE MOMENT METHOD FOR THE FOKKER–PLANCK EQUATION AND APPLICATIONS TO ELECTRON RADIOTHERAPY*

MARTIN FRANK†, HARTMUT HENSEL‡, AND AXEL KLAR‡

Abstract. This paper represents a first step toward a moment method for dose calculations in radiotherapy. Starting from a deterministic transport model for electron radiation and its Fokker–Planck approximation, a new macroscopic model is presented. We investigate several ways to simplify the deterministic model having two goals in mind, lower computation times on the one hand and high accuracy and model inherent incorporation of tissue inhomogeneities on the other hand. While being fast, the second property is lost in the often used pencil-beam models. We discuss the properties of well-known macroscopic models and design a new model, which combines their advantages. Several test cases, including the irradiation of a water phantom, are presented.

Key words. radiotherapy, moment method, minimum entropy

AMS subject classifications. 78M05, 92C50, 93A30

DOI. 10.1137/06065547X

1. Introduction.

Treatment with high energy ionizing radiation is one of the main methods in modern cancer therapy that is in clinical use. Since the early days of radiation treatment, high energy photons have been the most important type of radiation. Other types of radiation include high energy electrons and heavy charged particles like protons and ions. The latter type of radiation is of growing importance but has not reached the widespread use of photons and electrons, yet.

Before the treatment of the patient can be started, the expected dose distribution, i.e., the distribution of absorbed radiative energy in the patient, has to be calculated. During the past decades two main approaches to dose calculation were executed. The most accurate way to calculate the dose is given by Monte Carlo simulations [2]. Based on the well-known interactions of radiation in human tissue, a rigorous model of the energy distribution in the patient's body can be developed. Monte Carlo models allow for an exact computation of the dose distribution. Due to their high computational costs they have not found their way to everyday clinical use yet.

An alternative way to determine the dose was developed during the last 25 years. The so-called pencil-beam models [1] offer a reliable and fast alternative for most types of radiation treatment. These models are based on the Fermi–Eyges theory of radiative transfer [32, 14]. Despite their success in most clinical problems, they fail in complicated settings like air cavities or other inhomogeneities. This failure is caused by the underlying Fermi–Eyges theory because this approximation allows only for cross sections that vary spatially in the one-dimensional direction of the central axis of the beam. This assumption leads to a depth-dependence of the physical parameters

*Received by the editors March 28, 2006; accepted for publication (in revised form) October 18, 2006; published electronically February 15, 2007. This work was partially supported by the German Research Foundation DFG under grant KL 1105/14/2 and by the Rheinland-Pfalz Excellence Cluster "Dependable Adaptive Systems and Mathematical Modeling."

http://www.siam.org/journals/siap/67-2/65547.html

†Fachbereich Mathematik, TU Kaiserslautern, Erwin-Schrödinger-Str., 67663 Kaiserslautern, Germany (frank@mathematik.uni-kl.de).

‡Fraunhofer Institut für Techno- und Wirtschaftsmathematik, Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany (hensel@itwm.fraunhofer.de, klar@itwm.fraunhofer.de).

that is in general not observed in reality. Therefore pencil-beam models can account only for layered heterogeneities, and their effect on the dose has to be approximated by a rescaling of the kernels [18].

A third way to dose calculation has not attracted much attention in the medical physics community. This access is based on deterministic transport equations of radiative transfer. Similar to Monte Carlo simulations a rigorous model of the physical interactions in human tissue is modeled that can in principle be solved exactly. Recent studies for pure electron radiation can be found in [5, 7, 24, 22]. Electron and combined photon and electron radiation were studied in [36, 37] in the context of inverse therapy planning. The focus in these studies was more on rigorous analytical investigations with emphasis on mathematical aspects rather than a physically based modeling of the radiation-tissue interactions. A discrete transport model was proposed in [35, 19, 27, 20]. A consistent model of combined photon and electron radiation was developed [17] that includes the most important physical interactions. It allows for the computation of dose distributions for any geometric setting and is not restricted to layered inhomogeneities as the pencil-beam models are. Similar to Monte Carlo models, all interactions are modeled in detail which requires long computation times that are far beyond those needed for a clinical use.

This paper represents a first step toward simplified models for electron beam calculations in radiotherapy. In electron transport, small angle scattering with small energy loss, thus forward-peaked scattering kernels, is important. In transport calculations, these are either treated separately (e.g., $\delta - N$ method [34]) or the Boltzmann–Fokker–Planck equation can be used [31, 9]. In the latter approach, the scattering kernels are split into forward-peaked or more isotropic kernels, representing soft or hard collisions, respectively. Applying the Fokker–Planck asymptotic to the forward-peaked kernels, the (numerical) singularities are removed. In this work, we restrict ourselves to the Fokker–Plank approximation of the deterministic model although it does not contain all the physics necessary for realistic dose calculation. The reason for this is that the models we present here have all been successfully applied to radiative transfer with rather isotropic scattering, but it turns out that several of these models have problems in the Fokker–Planck limit of strong forward-peaked scattering. Here, we therefore modify the models to account for this difficulty alone. In future work, we will apply these new models to the combined Boltzmann–Fokker–Planck equation.

We investigate several ways to simplify the deterministic Fokker–Planck model having two goals in mind: On one hand the computation times should be lowered by several magnitudes. Our aim is a solver that can compute a typical dose distribution within minutes so that the model gets attractive for clinical use. On the other hand the errors in the dose distribution following the model simplifications should be very small. Especially the model inherent incorporation of tissue inhomogeneities must not be lost during the approximation steps. After all simplifications the model should still be able to fully reflect the heterogeneity of the human body, and the dose distribution should differ within only a few percent from more accurate benchmark simulations.

The rest of this paper is organized as follows. In section 2, we give an overview of the Boltzmann-type model, which was developed in [17] and serves as the starting point of our investigations. We also review the Fokker–Planck and Boltzmann–Fokker–Planck approximations. Well-known macroscopic approximations in the case of the Boltzmann equation are applied to the Fokker–Planck equation in section 3. Advantages and drawbacks of the macroscopic models are discussed in section 4. This discussion leads us to the proposal of a new model which unites some of the

best properties of the known models. We present numerical comparisons between several models in section 5 and include a realistic dose calculation using the full physical model from [17]. We conclude with a discussion of the results and present open problems for future work in section 6.

2. A deterministic model for dose calculation. A ray of high energy electrons that interacts with human tissue is subject to elastic scattering processes (Mott scattering) and inelastic ones (Møller scattering). Møller scattering is the electron-electron scattering process. It increases the number of free electrons when sufficient energy is transferred to the secondary electrons. It is this latter process that leads to energy deposition in the tissue, i.e., to the absorbed dose.

To formulate a transport equation for electrons we study their fluence in phase space. Let $\psi(r,\varepsilon,\Omega)\cos\Theta dA d\Omega d\epsilon_e dt$ be the number of electrons that move in time dt through area dA into the element of solid angle $d\Omega$ around Ω with an energy in the interval $(\epsilon_e, \epsilon_e + d\epsilon_e)$. The angle between direction Ω and the outer normal of dA is denoted by Θ. The direction is denoted by

$$\begin{aligned}\Omega = (\mu,\eta,\xi)^T &= (\mu, \sqrt{1-\mu^2}\cos\varphi, \sqrt{1-\mu^2}\sin\varphi)^T \\ &= (\cos\vartheta, \sin\vartheta\cos\varphi, \sin\vartheta\sin\varphi)^T,\end{aligned} \tag{2.1}$$

with φ and ϑ being the azimuth and polar angle in a cartesian coordinate system, respectively. The kinetic energy ϵ_e of the electrons is the relativistic kinetic energy.

2.1. Boltzmann transport equation. The transport equation formulated in [17] for the electrons is

$$\begin{aligned}\Omega\cdot\nabla\psi(r,\varepsilon,\Omega) = {}& \rho_e(r)\int_\varepsilon^\infty\int_{\mathcal{S}^2_{1/4}} \tilde\sigma_M(\epsilon_e',\epsilon_e,\Omega'\cdot\Omega)\psi(r,\varepsilon,\Omega')d\Omega_e' d\epsilon_e' \\ &+ \rho_e(r)\int_\varepsilon^\infty\int_{\mathcal{S}^2_{2/4}} \tilde\sigma_{M,\delta}(\epsilon_e',\epsilon_e,\Omega'\cdot\Omega)\psi(r,\varepsilon,\Omega')d\Omega_e' d\epsilon_e' \\ &+ \rho_c(r)\int_{\mathcal{S}^2}\sigma_{\mathrm{Mott}}(r,\epsilon_e,\Omega'\cdot\Omega)\psi(r,\varepsilon,\Omega')d\Omega_e' \\ &- \rho_e(r)\sigma_{M,\mathrm{tot}}(\epsilon_e)\psi(r,\varepsilon,\Omega) \\ &- \rho_c(r)\sigma_{\mathrm{Mott,tot}}(r,\epsilon_e)\psi(r,\varepsilon,\Omega),\end{aligned}$$

with $\tilde\sigma_M$ being the differential scattering cross section for Møller scattering of primary electrons; $\tilde\sigma_{M,\delta}$ being the differential Møller cross section for secondary electrons and σ_{Mott} being the differential cross section for Mott scattering; σ_M^{tot} and $\sigma_{\mathrm{Mott}}^{\mathrm{tot}}$ are the total cross sections for Møller and Mott scattering, respectively; and ρ_e and ρ_c are the densities of tissue electrons and tissue atomic cores, respectively. Explicit formulas for the cross sections can be found in Appendix A, see also [17]. In contrast to the model in [17] we set the lower energy bound of the transport equation to zero. The energy integration is performed over (ε,∞) since the electrons lose energy in every scattering event. Also, we consider only electron radiation. (2.2) could also be used to model electrons, which are generated by the interactions of photons with matter, as in [17]. In this case we would have an additional source term on the right-hand side for the generated electrons.

Due to kinematical reasons of the scattering processes, the range of solid angles in Møller (electron-electron) scattering is restricted. The electron, which has the

higher energy after the collision, is called the primary electron; the other electron is the secondary. Here, an incoming electron with energy ε' hits an electron at rest. After the collision, the angle between the directions of the electrons is at most $\pi/2$. Electrons are indistinguishable. For an angle in $[0, \pi/4]$, the electron with energy ε is the primary electron; for an angle in $[\pi/4, \pi/2]$, it is the secondary electron. Therefore we write the Møller term in the Boltzmann equation as

$$\int_{\mathcal{S}^2_{1/4}} f(\varphi,\vartheta)d\Omega := \int_0^{2\pi}\int_0^{\pi/4} f(\varphi,\vartheta)\sin\vartheta d\vartheta d\varphi \quad \text{and}$$
$$\int_{\mathcal{S}^2_{2/4}} f(\varphi,\vartheta)d\Omega := \int_0^{2\pi}\int_{\pi/4}^{\pi/2} f(\varphi,\vartheta)\sin\vartheta d\vartheta d\varphi,$$

where the axes of reference are given by Ω in the scattering integrals in (2.2).

Besides the transport equation one needs an equation for the absorbed dose. It was derived in [17] as an asymptotic limit of a model with a finite lower energy bound $\epsilon_{\mathrm{s}} > 0$. The formula is exact if one chooses the lower energy limit $\epsilon_{\mathrm{s}} = 0$, as we do here.

$$D(r) = \frac{T}{\rho(r)}\int_0^\infty S_{\mathrm{M}}(r,\epsilon'_{\mathrm{e}})\psi^{(0)}(r,\epsilon'_{\mathrm{e}})d\epsilon'_{\mathrm{e}},$$

with

$$\psi^{(0)}(r,\epsilon_{\mathrm{e}}) := \int_{S^2}\psi(r,\epsilon_{\mathrm{e}},\Omega')d\Omega', \tag{2.2}$$

T being the duration of the irradiation of the patient, and ρ being the mass density of the irradiated tissue. If all quantities are calculated in System Internationale (SI) units, (2.1) leads to SI units J/kg or Gray (Gy) for the dose.

S_{M} is the stopping power related to the Møller cross section. It is defined as

$$S_{\mathrm{M}}(r,\epsilon_{\mathrm{e}}) = \rho_{\mathrm{e}}(r)\int_{\epsilon_{\mathrm{B}}}^{(\epsilon_{\mathrm{e}}-\epsilon_{\mathrm{B}})/2}\epsilon'_{\mathrm{e}}\sigma_{\mathrm{M}}(\epsilon_{\mathrm{e}},\epsilon'_{\mathrm{e}})d\epsilon'_{\mathrm{e}}.$$

ϵ_{B} is the binding energy of electrons in tissue atoms or molecules, cf. [17] for details.

The above Boltzmann-type transport equation can be approximated by a Boltzmann–Fokker–Planck equation taking into account the fact that the great majority of the electron collisions are soft collisions. These collisions can be approximated by a Fokker–Planck-type term, see the next section. However, some electrons will also experience hard collisions with large changes in direction and energy losses which have to be described by Boltzmann integral terms. For reasons given above, we will consider only the Fokker–Planck part. Approximate methods for the full Boltzmann–Fokker–Planck equation will be considered in a forthcoming paper.

2.2. Fokker–Planck approximation. Electron transport in tissue has very distinctive properties. The soft collision differential scattering cross sections have a pronounced maximum for small scattering angles and small energy loss. This allows for a simplification of the scattering terms in the Boltzmann equation. The Fokker–Planck equation is the result of an asymptotic analysis for both small energy loss and small deflections. It has been rigorously derived in [30] and has been applied to the

above Boltzmann model in [17]. The Fokker–Planck equation is

$$\Omega\cdot\nabla\psi(r,\varepsilon,\Omega)=(T_{\mathrm{M}}(r,\varepsilon)+T_{\mathrm{Mott}}(r,\varepsilon))\Delta_\Omega\psi(r,\varepsilon,\Omega)+\partial_\varepsilon(S_{\mathrm{M}}(r,\varepsilon)\psi(r,\varepsilon,\Omega)), \tag{2.3}$$

with

$$T_{\mathrm{M}}(r,\varepsilon)=\pi\rho_{\mathrm{e}}(r)\int_{\varepsilon_{\mathrm{B}}}^{(\varepsilon-\varepsilon_{\mathrm{B}})/2}\int_{-1}^{1}(1-\mu)\tilde{\sigma}_{\mathrm{M}}(\varepsilon,\varepsilon',\mu)d\mu d\varepsilon', \tag{2.4}$$

$$T_{\mathrm{Mott}}(r,\varepsilon)=\pi\rho_{\mathrm{e}}(r)\int_{-1}^{1}(1-\mu)\sigma_{\mathrm{Mott}}(\varepsilon,\mu)d\mu, \tag{2.5}$$

and the Laplacian on the unit sphere

$$\Delta_\Omega=\frac{\partial}{\partial\mu}(1-\mu^2)\frac{\partial}{\partial\mu}+\frac{1}{1-\mu^2}\frac{\partial}{\partial\varphi}. \tag{2.6}$$

In the following, we will interpret the energy variable ε as time. We assign boundary values to the Fokker–Planck equation (2.3) and will consider the following initial boundary value problem. We have the spatial domain $r\in\mathcal{Z}$ with boundary $\partial\mathcal{Z}$ and outward normal n, energy $\varepsilon\in[0,\infty)$, direction $\Omega\in\mathcal{S}^2$. We prescribe the ingoing radiation at the spatial boundary,

$$\psi(r,\varepsilon,\Omega)=\psi_b(r,\varepsilon,\Omega)\quad\text{for}\quad r\in\partial\mathcal{Z},n\cdot\Omega<0. \tag{2.7}$$

For the energy, we prescribe the "initial value"

$$\psi(r,\infty,\Omega)=0. \tag{2.8}$$

In the numerical simulations, we use a large cutoff energy.

2.3. Fermi–Eyges theory. To derive the Fermi–Eyges approximation, which is the basis of many schemes used in practice, several additional simplifying assumptions have to be made. As incident radiation, a monoenergetic pencil beam is assumed. We consider an infinite plate $(x,y,z)\in(0,1)\times\mathbb{R}^2$ and prescribe [7]

$$\psi(0,y,z,\varepsilon,\Omega)=\delta(y)\delta(z)\delta(\varepsilon-\varepsilon_0)\frac{\delta(1-\mu)}{2\pi}\quad\text{for}\quad 0<\mu<1, \tag{2.9}$$

$$\psi(1,y,z,\varepsilon,\Omega)=0\quad\text{for}\quad -1<\mu<0. \tag{2.10}$$

Energy transfer and change in direction are treated separately. First, if we neglect angular deflections, we obtain the straight ahead approximation,

$$\Omega\cdot\nabla\psi(r,\varepsilon,\Omega)=\frac{\partial}{\partial\varepsilon}(S_{\mathrm{M}}(r,\varepsilon)\psi(r,\varepsilon,\Omega)). \tag{2.11}$$

The method of characteristics for conservation laws relates energy and penetration depth by the initial value problem

$$\frac{d\varepsilon(x)}{dx}=-S_{\mathrm{M}}(x,\varepsilon(x)),\quad\varepsilon(0)=\varepsilon_0. \tag{2.12}$$

This means that we obtain the energy of the monoenergetic beam as a function of penetration depth. We have assumed that S_{M} depends only on the depth x.

On the other hand, we neglect energy loss and assume small angle scattering. The direction Ω is approximated by

$$\Omega = (\mu, \eta, \xi) \approx (1, \eta, \xi). \tag{2.13}$$

The Fokker–Planck equation becomes the Fermi equation

$$\begin{aligned} &\frac{\partial}{\partial x}\psi(r,\varepsilon,\eta,\xi) + \eta\frac{\partial}{\partial y}\psi(r,\varepsilon,\eta,\xi) + \xi\frac{\partial}{\partial z}\psi(r,\varepsilon,\eta,\xi) \\ &\qquad = \frac{T_{\mathrm{M}}(r,\varepsilon) + T_{\mathrm{Mott}}(r,\varepsilon)}{2}\Delta_{\eta,\xi}\psi(r,\varepsilon,\eta,\xi), \end{aligned} \tag{2.14}$$

where [5]

$$\Delta_{\eta,\xi} = \frac{\partial^2}{\partial\eta^2} + \frac{\partial^2}{\partial\xi^2} = \frac{1}{\mu}\frac{\partial}{\partial\mu}\left(\frac{1-\mu^2}{\mu}\frac{\partial}{\partial\mu}\right) + \frac{1}{1-\mu^2}\frac{\partial^2}{\partial\varphi^2}. \tag{2.15}$$

The Fermi equation permits only directions $\mu > 0$. An asymptotic derivation is given in [6].

The starting point of Fermi–Eyges theory is to combine the Fermi equation and the straight ahead approximation, i.e., $\varepsilon = \varepsilon(r)$ in (2.14). The Fermi equation is defined for $\eta^2 + \xi^2 < 1$. By artificially extending the range of η and ξ to the real line (ψ small for η, ξ large), Fermi ([32], pp. 265–268) was able to give an explicit solution for the pencil-beam problem for constant coefficients. Eyges [14] generalized this solution to space-dependent coefficients. Using the straight ahead approximation, it is thus possible to incorporate energy dependence.

In cylindrical coordinates, $r = (x, y, z) = (x, \rho\cos\phi, \rho\sin\phi)$, the Fermi equation reads

$$\frac{\partial}{\partial x}\psi(x,\rho,\eta) = -\eta\frac{\partial}{\partial\rho}\psi(x,\rho,\eta) + \frac{T_{\mathrm{M}}(x,\varepsilon(x)) + T_{\mathrm{Mott}}(x,\varepsilon(x))}{2}\frac{\partial^2}{\partial\eta^2}\psi(x,\rho,\eta). \tag{2.16}$$

Note that, due to symmetries in the pencil-beam problem, the solution depends neither on φ nor on ξ. Therefore pencil-beam models can account only for layered heterogeneities. Moreover, we assumed that the scattering coefficients T_{M} and T_{Mott} depend only on the penetration depth, i.e., the medium is layered. The explicit solution due to Eyges is

$$\psi_{FE}(x,\rho,\eta) = \frac{1}{2\pi\sqrt{A_0A_2(x) - A_1^2(x)}}\exp\left(-\frac{A_2(x)\eta^2 - 2A_1(x)\eta\rho + A_0\rho^2}{4A_0(A_0A_2(x) - A_1^2(x))}\right), \tag{2.17}$$

where

$$A_k(x) = \int_0^x (x-y)^k\frac{T_{\mathrm{M}}(y,\varepsilon(y)) + T_{\mathrm{Mott}}(y,\varepsilon(y))}{2}dy. \tag{2.18}$$

The so-called pencil-beam models based on the Fermi–Eyges theory offer a reliable and fast alternative for most types of radiation treatment as mentioned in the introduction. However, they fail in complicated settings, because the Fermi–Eyges approximation allows only for cross sections that vary spatially in the one-dimensional direction of the axis of the beam, i.e., for layered heterogeneities. Its generalization to nonlayered media is not obvious. Many heuristics have been introduced to that end [1]. Here, the basic Fermi–Eyges solution will serve as a comparison to the macroscopic methods.

3. Macroscopic models. In this section, we give an overview over several well-known macroscopic models as applied to the Fokker–Planck equation. For the sake of simplicity, we restrict ourselves to one space dimension. The ideas however are general and have been applied to two or three space dimensions.

3.1. Spherical harmonics. The spherical harmonics approach was developed first for radiative transfer [21, 13]. With high orders of the expansion, it has been proposed for use in photon transport medical physics problems, see, e.g., [4]. The idea of the spherical harmonics approach is to express the angular dependence of the distribution function in terms of a truncated Fourier series,

$$\psi_{SH}(r,\varepsilon,\mu) = \sum_{l=0}^{N} \psi^{(l)}(r,\varepsilon)\frac{2l+1}{2}P_l(\mu), \tag{3.1}$$

where P_l are the Legendre polynomials. This means that we assume $\psi^{(l)} = 0$ for $l > N$. The Legendre polynomials form an orthogonal basis of the space of polynomials with respect to the standard scalar product on $[-1,1]$. The Fourier coefficients are the moments

$$\psi^{(l)}(r,\varepsilon) = \int_{-1}^{1} \psi(r,\mu,\varepsilon)P_l(\mu)d\mu. \tag{3.2}$$

We test the Fokker–Planck equation with P_l, integrate with respect to μ over $[-1,1]$, and use a recursion relation to obtain the P_N equations

$$\begin{aligned} -S_{\mathrm{M}}\partial_\varepsilon\psi^{(l)} + \partial_x\left(\frac{l+1}{2l+1}\psi^{(l+1)} + \frac{l}{2l+1}\psi^{(l-1)}\right) \\ = -\frac{T_{\mathrm{M}}+T_{\mathrm{Mott}}}{2}l(l+1)\psi^{(l)} + (\partial_\varepsilon S_{\mathrm{M}})\psi^{(l)} \end{aligned} \tag{3.3}$$

for $l = 0,\dots,N$. The P_3 equations, which we will use later, read

$$-S_{\mathrm{M}}\partial_\varepsilon\psi^{(0)} + \partial_x\psi^{(1)} = (\partial_\varepsilon S_{\mathrm{M}})\psi^{(0)}, \tag{3.4}$$

$$-S_{\mathrm{M}}\partial_\varepsilon\psi^{(1)} + \partial_x\left(\frac{1}{3}\psi^{(0)} + \frac{2}{3}\psi^{(2)}\right) = -(T_{\mathrm{M}}+T_{\mathrm{Mott}})\psi^{(1)} + (\partial_\varepsilon S_{\mathrm{M}})\psi^{(1)}, \tag{3.5}$$

$$-S_{\mathrm{M}}\partial_\varepsilon\psi^{(2)} + \partial_x\left(\frac{2}{5}\psi^{(1)} + \frac{3}{5}\psi^{(3)}\right) = -(T_{\mathrm{M}}+T_{\mathrm{Mott}})\psi^{(2)} + (\partial_\varepsilon S_{\mathrm{M}})\psi^{(2)}, \tag{3.6}$$

$$-S_{\mathrm{M}}\partial_\varepsilon\psi^{(3)} + \partial_x\frac{3}{7}\psi^{(2)} = -(T_{\mathrm{M}}+T_{\mathrm{Mott}})\psi^{(3)} + (\partial_\varepsilon S_{\mathrm{M}})\psi^{(3)}. \tag{3.7}$$

3.2. Diffusion and Fermi age. The diffusion approximation can be derived in several ways. One way is to start from the first spherical harmonics moment equation,

$$-S_{\mathrm{M}}\partial_\varepsilon\psi^{(0)} + \partial_x\psi^{(1)} = (\partial_\varepsilon S_{\mathrm{M}})\psi^{(0)}, \tag{3.8}$$

and to approximate the second equation by Fick's law,

$$\partial_x\frac{1}{3}\psi^{(0)} = -(T_{\mathrm{M}}+T_{\mathrm{Mott}})\psi^{(1)}. \tag{3.9}$$

This gives

$$-\partial_\varepsilon(S_{\mathrm{M}}(\varepsilon)\psi^{(0)}) = \frac{1}{3(T_{\mathrm{M}}+T_{\mathrm{Mott}})}\partial_x^2\psi^{(0)}. \tag{3.10}$$

Alternatively, the diffusion equation can be derived by scaling arguments. One may start with the Fokker–Planck equation (2.3). Using the diffusion scaling (spatial scale of order $1/\delta$) combined with a rescaling of the term $\partial_\varepsilon(S_{\mathrm{M}}(r,\varepsilon))$ of order δ^2, one obtains with the classical arguments, see, e.g., [25], the above diffusion equation. However, the relevant physical parameters, see the last section, do not coincide with such a scaling; i.e., we are not in the range of validity of the diffusion equation. See also the numerical results in the last section. In Fermi age theory, the name under which the diffusion approximation is known in nuclear; we introduce the slowing down density $\tilde{\psi} = S_{\mathrm{M}}\psi^{(0)}$. Furthermore, we introduce the age τ by

$$\frac{d\tau}{d\varepsilon} = -\frac{1}{3(T_{\mathrm{M}} + T_{\mathrm{Mott}})}\frac{1}{S_{\mathrm{M}}}. \tag{3.11}$$

Note that the age has dimension length squared. Its square root is a characteristic length called the fast diffusion length. The diffusion equation attains the simple form

$$\frac{\partial}{\partial\tau}\tilde{\psi} = \Delta\tilde{\psi}. \tag{3.12}$$

3.3. Minimum entropy. The approximations based on the expansion of the distribution function into a polynomial suffer from several drawbacks [11]. Most importantly, the distribution function can become negative, and thus the moments computed from the distribution can become unphysical. One way to overcome this problem is to use an entropy minimization principle to obtain the constitutive equation to close the moment equations. This principle has become the main concept of rational extended thermodynamics [29].

The minimum entropy $M1$ model [11, 3] for electrons [8] can be derived in the following way. We start with the first two equations of the spherical harmonics method above:

$$-S_{\mathrm{M}}\partial_\varepsilon\psi^{(0)} + \partial_x\psi^{(1)} = (\partial_\varepsilon S_{\mathrm{M}})\psi^{(0)}, \tag{3.13}$$

$$-S_{\mathrm{M}}\partial_\varepsilon\psi^{(1)} + \partial_x\left(\frac{1}{3}\psi^{(0)} + \frac{2}{3}\psi^{(2)}\right) = -(T_{\mathrm{M}} + T_{\mathrm{Mott}})\psi^{(1)} + (\partial_\varepsilon S_{\mathrm{M}})\psi^{(1)}. \tag{3.14}$$

This system is underdetermined: two equations for three unknowns. To close the system we determine a distribution function ψ_{ME} that minimizes the entropy of the electrons,

$$H_R^*(\psi) = -\int_{-1}^{1}\psi\log\psi d\mu, \tag{3.15}$$

under the constraint that it reproduces the lower order moments,

$$\int_{-1}^{1}\psi_{ME}d\mu = \psi^{(0)} \quad \text{and} \quad \int_{-1}^{1}\mu\psi_{ME}d\mu = \psi^{(1)}. \tag{3.16}$$

The entropy minimizer can be written as [8]

$$\psi_{ME}(\mu) = \alpha e^{\beta\mu}. \tag{3.17}$$

This is a Maxwell–Boltzmann-type distribution, and α, β are Lagrange multipliers enforcing the constraints.

When dealing with nonlinear closure functions, it is always crucial to ask which moments can be realized by the concrete distribution. In this case, we have as a necessary condition that

$$\psi^{(0)} = \int_{-1}^{1} \psi_{ME} d\mu \geq 0 \quad \text{and} \quad |\psi^{(1)}| = |\int_{-1}^{1} \mu \psi_{ME} d\mu| \leq \psi^{(0)}. \tag{3.18}$$

This is also sufficient; i.e., for each pair of moments $(\psi^{(0)}, \psi^{(1)})$ that satisfies (3.18) one can find Lagrange multipliers α and β such that the moments are realized by the distribution function ψ_{ME}.

It is not possible to express the highest moment $\psi^{(2)}$ explicitly in terms of $\psi^{(0)}$ and $\psi^{(1)}$, but we can write for the flux function

$$\frac{1}{3}\psi^{(0)} + \frac{2}{3}\psi^{(2)} = \chi\left(\frac{\psi^{(1)}}{\psi^{(0)}}\right)\psi^{(0)}. \tag{3.19}$$

The Eddington factor χ is defined for $|\psi^{(1)}/\psi^{(0)}| \leq 1$ and can be computed numerically [8] from the set of equations

$$\frac{\psi^{(1)}}{\psi^{(0)}} = \beta \coth \beta - 1, \tag{3.20}$$

$$\chi\left(\frac{\psi^{(1)}}{\psi^{(0)}}\right) = \frac{\psi^{(2)}}{\psi^{(0)}} = \beta^2 + 2 - 4\beta \coth \beta, \tag{3.21}$$

where β is the Lagrange multiplier. However, also the models based on the minimum entropy closure yield unphysical solutions in certain situations, see, e.g., the numerical results.

3.4. Half-moment approximation. A model which has been successfully applied to anisotropic radiative transfer, removing some drawbacks of the minimum entropy model in the last section, is the half-moment approximation [33, 38]. A typical drawback of the minimum entropy solution is displayed in the simulations in Figure 5.1. The idea is to average not over all directions but over certain subsets, e.g., particles moving left or right. In one dimension, this means to integrate over $[-1, 0]$ and $[0, 1]$. We denote the half-moments by

$$\psi_+^{(l)}(r, \varepsilon) = \int_0^1 \psi(r, \varepsilon, \mu) P_l(\mu) d\mu \quad \text{and} \quad \psi_-^{(l)}(r, \varepsilon) = \int_{-1}^0 \psi(r, \varepsilon, \mu) P_l(\mu) d\mu. \tag{3.22}$$

Applying this approach to the Fokker–Planck equation we obtain

$$-S_{\mathrm{M}} \partial_\varepsilon \psi_+^{(0)} + \partial_x \psi_+^{(1)} = \frac{T_{\mathrm{M}} + T_{\mathrm{Mott}}}{2} \int_0^1 \partial_\mu (1 - \mu^2) \partial_\mu \psi d\mu + (\partial_\varepsilon S) \psi_+^{(0)}. \tag{3.23}$$

If we use integration by parts, the integral on the right-hand side becomes

$$\int_0^1 \partial_\mu (1 - \mu^2) \partial_\mu \psi d\mu = -\partial_\mu \psi(0). \tag{3.24}$$

We note that, in contrast to the spherical harmonics approach, on the right-hand side a microscopic term, i.e., the distribution itself instead of its moments, appears. In

a similar way, we can derive equations for higher moments, as well as the negative half-space:

$$-S_{\mathrm{M}}\partial_\varepsilon\psi_+^{(0)} + \partial_x\psi_+^{(1)} = -\frac{T_{\mathrm{M}}+T_{\mathrm{Mott}}}{2}\partial_\mu\psi(0) + (\partial_\varepsilon S_{\mathrm{M}})\psi_+^{(0)}, \tag{3.25}$$

$$-S_{\mathrm{M}}\partial_\varepsilon\psi_+^{(1)} + \partial_x\left(\psi_+^{(2)} + \frac{1}{3}\psi_+^{(0)}\right) = \frac{T_{\mathrm{M}}+T_{\mathrm{Mott}}}{2}(\psi(0) - 2\psi_+^{(1)}) + (\partial_\varepsilon S_{\mathrm{M}})\psi_+^{(1)}, \tag{3.26}$$

$$-S_{\mathrm{M}}\partial_\varepsilon\psi_-^{(0)} + \partial_x\psi_-^{(1)} = \frac{T_{\mathrm{M}}+T_{\mathrm{Mott}}}{2}\partial_\mu\psi(0) + (\partial_\varepsilon S_{\mathrm{M}})\psi_-^{(0)}, \tag{3.27}$$

$$-S_{\mathrm{M}}\partial_\varepsilon\psi_-^{(1)} + \partial_x\left(\psi_-^{(2)} + \frac{1}{3}\psi_+^{(0)}\right) = \frac{T_{\mathrm{M}}+T_{\mathrm{Mott}}}{2}(-\psi(0) - 2\psi_-^{(1)}) + (\partial_\varepsilon S_{\mathrm{M}})\psi_-^{(1)}. \tag{3.28}$$

In principle, we can model the boundary terms with the underlying distribution, which is used to model the higher order moments. Again, we can choose either a polynomial or minimum entropy closure.

For the half-moment P_1 closure, sometimes also called the double P_1 closure, we take

$$\psi_{HP1}(\mu) = \begin{cases} \alpha_- + \beta_-\mu & \text{for} \quad \mu \in [-1,0], \\ \alpha_+ + \beta_+\mu & \text{for} \quad \mu \in [0,1]. \end{cases} \tag{3.29}$$

If we model the microscopic terms using this function by defining $\psi(0)$ and $\partial_\mu\psi(0)$ as the limit from the right and left, respective of the positive or negative half-space, we obtain the closed half-moment P_1 system:

$$-S_{\mathrm{M}}\partial_\varepsilon\psi_+^{(0)} + \partial_x\psi_+^{(1)} = 3(T_{\mathrm{M}}+T_{\mathrm{Mott}})(\psi_+^{(0)} - 2\psi_+^{(1)}) + (\partial_\varepsilon S_{\mathrm{M}})\psi_+^{(0)}, \tag{3.30}$$

$$-S_{\mathrm{M}}\partial_\varepsilon\psi_+^{(1)} + \partial_x\left(\chi_+\left(\frac{\psi_+^{(1)}}{\psi_+^{(0)}}\right)\psi^{(0)}\right) = 2(T_{\mathrm{M}}+T_{\mathrm{Mott}})(\psi_+^{(0)} - 2\psi_+^{(1)}) + (\partial_\varepsilon S_{\mathrm{M}})\psi_+^{(1)}, \tag{3.31}$$

$$-S_{\mathrm{M}}\partial_\varepsilon\psi_-^{(0)} + \partial_x\psi_-^{(1)} = 3(T_{\mathrm{M}}+T_{\mathrm{Mott}})(\psi_-^{(0)} + 2\psi_-^{(1)}) + (\partial_\varepsilon S_{\mathrm{M}})\psi_-^{(0)}, \tag{3.32}$$

$$-S_{\mathrm{M}}\partial_\varepsilon\psi_-^{(1)} + \partial_x\left(\chi_-\left(\frac{\psi_-^{(1)}}{\psi_-^{(0)}}\right)\psi^{(0)}\right) = 2(T_{\mathrm{M}}+T_{\mathrm{Mott}})(\psi_-^{(0)} + 2\psi_-^{(1)}) + (\partial_\varepsilon S_{\mathrm{M}})\psi_-^{(1)}. \tag{3.33}$$

The half-Eddington factors for the P_1 closure are

$$\chi_\pm\left(\frac{\psi_\pm^{(1)}}{\psi_\pm^{(0)}}\right) = -\frac{1}{6} \pm \frac{\psi_\pm^{(1)}}{\psi_\pm^{(0)}}. \tag{3.34}$$

The half-moment minimum entropy closure is

$$\psi_{HME}(\mu) = \begin{cases} \alpha_- e^{\beta_-\mu} & \text{for} \quad \mu \in [-1,0], \\ \alpha_+ e^{+\beta_+\mu} & \text{for} \quad \mu \in [0,1]. \end{cases} \tag{3.35}$$

All physically reasonable half-moments can be realized; i.e., for $\psi_+^{(0)} \geq 0$ and $0 \leq \psi_+^{(1)} \leq \psi_+^{(0)}$ there exist Lagrange multipliers α_+ and β_+ such that the moments are realized by ψ_{HME} (respectively for "$-$"). The half-Eddington factors and the additional terms must be computed numerically. They can be obtained from the

system

$$\frac{\psi_+^{(1)}}{\psi^{(0)}} = \frac{\beta_+\beta_- e^{\beta_+} - \beta_-(e^{\beta_+}-1)}{\beta_+\beta_-(e^{\beta_+}-1) + \beta_+^2(1-e^{-\beta_-})}, \tag{3.36}$$

$$\frac{\psi_-^{(1)}}{\psi^{(0)}} = \frac{\beta_+\beta_- e^{-\beta_-} - \beta_+(1-e^{-\beta_-})}{\beta_-^2(e^{\beta_+}-1) + \beta_+\beta_-(1-e^{-\beta_-})}, \tag{3.37}$$

$$\chi_+\left(\frac{\psi_+^{(1)}}{\psi_+^{(0)}}\right) = \frac{\beta_- e^{\beta_+}}{\beta_-(e^{\beta_+}-1) + \beta_+(1-e^{-\beta_-})} - \frac{2}{\beta_+}\frac{\psi_+^{(1)}}{\psi^{(0)}}, \tag{3.38}$$

$$\chi_-\left(\frac{\psi_-^{(1)}}{\psi_+^{(0)}}\right) = \frac{-\beta_+ e^{-\beta_-}}{\beta_+(e^{\beta_+}-1) + \beta_-(1-e^{-\beta_-})} - \frac{2}{\beta_-}\frac{\psi_-^{(1)}}{\psi^{(0)}}, \tag{3.39}$$

in which the Lagrange multipliers $\beta_\pm$ have to be eliminated.

The underlying distribution function for the minimum entropy half-moment model is never negative in contrast to the linear half-moment model. For many situations, e.g., in the case of Boltzmann-type transport equations, the nonlinear half-moment closure has, similar to the nonlinear full moment closure, many advantages compared to the linear closure. However, in the present setting the major problems, as, e.g., the appearance of terms in the balance equations, which depend on the distribution function and not only on the first moments, are present for both closures. This will be explained below in more detail. Thus we do not get into the details of the nonlinear numerical closure here.

4. A new approximation extending the method of moments. In this section we discuss the drawbacks of the above models and suggest a new moment model adapted to the original Fokker–Planck equations. The behavior of the spherical harmonics method, the diffusion approximation, and the minimum entropy method have been studied extensively. The half-moment method is not as well known as the other approximations. All methods have their individual advantages and drawbacks. We will sketch some of them.

The diffusion approximation leads to a scalar parabolic PDE, which makes it very fast and simple to solve. The three other methods give systems of hyperbolic equations, whose numerical analysis is more difficult. The key drawback of the diffusion approximation is its diffusivity, i.e., it smears out the solution. Furthermore, information is propagated infinitely fast, in contrast to the finite speed of propagation in hyperbolic equations and in reality.

The main disadvantage of the spherical harmonics approach is that it allows for negative particle distributions and has fixed characteristic speeds. The minimum entropy approximation, on the other hand, always has a positive distribution and adapts to the speed of propagation. However, it allows for discontinuous solutions [8, 12]. The individual drawbacks can be demonstrated in one example, the two beam case shown in Figure 5.1. The details of the test case are described in section 5.

The half-moment model was designed to remove the drawbacks of minimum entropy and spherical harmonics. This approach was successful in the case of radiative transfer [33]. The half-moment minimum entropy model guarantees positivity, adapts its speed of propagation, and eliminates the possibility of unphysical discontinuities. The half-moment P_N model is also an improvement over the P_N model but lacks the correct speed of propagation and does not guarantee positivity.

In the present case, the numerical results show that the half-moment model (both with polynomial and minimum entropy closure) fails dramatically. One hint at this

failure is the appearance of the terms $\partial_\mu \psi(0)$ in the derivation, which are depending on the distribution function and not directly on the moments of this function. A second indication that the half-moment model above is not a good approximation to the Fokker–Planck equation is that the "+" and "−" equations decouple. Thus the model cannot be a consistent discretization of the Fokker–Planck equation. The mathematical reason for the failure is that integration by parts is not allowed if the integrand is discontinuous. While full moment models (as opposed to half-moment models) and the diffusion approximation are invariant under the Fokker–Planck limit,

$$\sigma_{\text{tot}} \to \infty, \quad \frac{\sigma^{(1)} = \int_{-1}^{1} \mu\sigma(\mu)d\mu}{\sigma_{\text{tot}}} \to 1 \quad \text{with} \quad \sigma_{\text{tot}}\left(1 - \frac{\sigma^{(1)}}{\sigma_{\text{tot}}}\right) \text{ fixed}, \tag{4.1}$$

the half-moment models diverge in this limit. The reason for this divergence will be investigated in more detail in future work.

Driven by the success of the half-moment approach for radiative transfer, we seek to modify the approach in such a way that it is suitable for the Fokker–Planck equation. We want to use half-moments where possible but avoid ambiguous microscopic terms. We observe that the lowest order ansatz which satisfies these criteria is to test with P_0 and integrate over $[-1, 1]$ and then to test with P_1 and integrate over $[-1, 0]$ and $[0, 1]$ separately. We get

$$-S_{\text{M}}\partial_\varepsilon \psi^{(0)} + \partial_x(\psi_+^{(1)} + \psi_-^{(1)}) = (\partial_\varepsilon S_{\text{M}})\psi^{(0)}, \tag{4.2}$$

$$-S_{\text{M}}\partial_\varepsilon \psi_+^{(1)} + \partial_x\left(-\frac{1}{6}\psi_+^{(0)} + \psi_+^{(1)}\right) = \frac{T_{\text{M}} + T_{\text{Mott}}}{2}(\psi(0) - 2\psi_+^{(1)}) + (\partial_\varepsilon S_{\text{M}})\psi_+^{(1)}, \tag{4.3}$$

$$-S_{\text{M}}\partial_\varepsilon \psi_-^{(1)} + \partial_x\left(-\frac{1}{6}\psi_-^{(0)} - \psi_-^{(1)}\right) = \frac{T_{\text{M}} + T_{\text{Mott}}}{2}(-\psi(0) - 2\psi_-^{(1)}) + (\partial_\varepsilon S_{\text{M}})\psi_-^{(1)}. \tag{4.4}$$

If we close this system with an underlying distribution that is continuous, then the integration by parts which was performed to obtain the last two equations is justified. Furthermore, the microscopic term in the last two equations is unambiguously defined.

To close the model, we choose a modified minimum entropy (MME) closure function which is a mixture between half moment and full moment closure:

$$\psi_{MME} = \begin{cases} \alpha e^{\beta_+ \mu}, & \mu \in [0, 1], \\ \alpha e^{\beta_- \mu}, & \mu \in [-1, 0]. \end{cases} \tag{4.5}$$

Here, one could also consider a linear closure, but this has the drawback that it allows for negative energies and does not adapt to the correct speed of propagation similar to the radiative transfer case.

Whereas in the above minimum entropy closure functions all physically relevant moments could be attained, moment realizability is an issue here. Obviously,

$$\psi^{(0)} \geq 0, \quad 0 \leq \psi_+^{(1)} \leq \psi^{(0)}, \quad \text{and} \quad -\psi^{(0)} \leq \psi_-^{(1)} \leq 0 \tag{4.6}$$

are necessary conditions for the invertibility. Furthermore, we observe that

$$\begin{aligned} \psi_+^{(1)} - \psi_-^{(1)} &= \int_0^1 \mu e^{\beta_+ \mu} d\mu - \int_{-1}^0 \mu e^{\beta_- \mu} d\mu = \int_0^1 \mu e^{\beta_+ \mu} d\mu - \int_1^0 \mu e^{-\beta_- \mu} d\mu \\ &= \int_0^1 \mu(e^{\beta_+ \mu} + e^{-\beta_- \mu}) d\mu \leq \int_0^1 \mu(e^{\beta_+ \mu} + e^{-\beta_- \mu}) d\mu = \psi^{(0)}. \end{aligned} \tag{4.7}$$

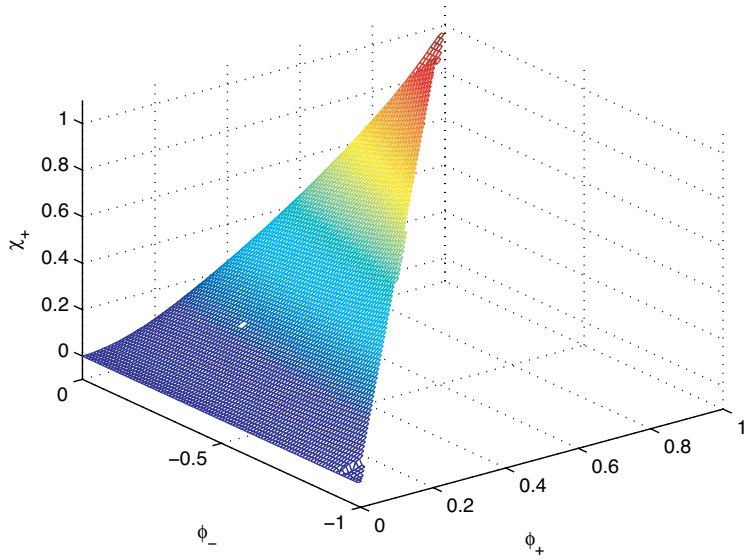


FIG. 4.1. *Minimum entropy Eddington factor* χ_+ *as a function of* $\phi_+ = \psi_+^{(1)}/\psi^{(0)}$ *and* $\phi_- = \psi_-^{(1)}/\psi^{(0)}$.

The numerical inversion of the system shows that conditions (4.6) and (4.7) are sufficient for realizability. This means that not all choices of the moments $(\psi^{(0)}, \psi_+^{(1)}, \psi_-^{(1)})$ in (4.6) are allowed here. In the test cases presented below, for realizable initial values, the moments always remained realizable during the evolution. A proof of this fact is crucial in the proof of the well-posedness of the system.

Again, the system cannot be closed analytically, but it can be written in the form

$$-S_{\mathrm{M}}\partial_\varepsilon\psi^{(0)} + \partial_x(\psi_+^{(1)} + \psi_-^{(1)}) = (\partial_\varepsilon S_{\mathrm{M}})\psi^{(0)},$$

$$-S_{\mathrm{M}}\partial_\varepsilon\psi_+^{(1)} + \partial_x\left(\chi_+\left(\frac{\psi_+^{(1)}}{\psi^{(0)}}, \frac{\psi_-^{(1)}}{\psi^{(0)}}\right)\psi^{(0)}\right)$$

$$= \frac{T_{\mathrm{M}} + T_{\mathrm{Mott}}}{2}(2\psi^{(0)} - 5\psi_+^{(1)} + 3\psi_-^{(1)}) + (\partial_\varepsilon S_{\mathrm{M}})\psi_+^{(1)},$$

$$-S_{\mathrm{M}}\partial_\varepsilon\psi_-^{(1)} + \partial_x\left(\chi_-\left(\frac{\psi_-^{(1)}}{\psi^{(0)}}, \frac{\psi_+^{(1)}}{\psi^{(0)}}\right)\psi^{(0)}\right)$$

$$= \frac{T_{\mathrm{M}} + T_{\mathrm{Mott}}}{2}(-2\psi^{(0)} + 3\psi_+^{(1)} - 5\psi_-^{(1)}) + (\partial_\varepsilon S_{\mathrm{M}})\psi_-^{(1)}.$$

The Eddington factors satisfy the symmetry relation $\chi_+(\phi, \psi) = \chi_-(-\psi, -\phi)$. Figure 4.1 shows χ_+.

5. Numerical results. In this section we consider several test problems in slab geometry. We have a domain between two infinite parallel plates. Thus the problem becomes one-dimensional in space. We will compare the Fokker–Planck solution with diffusion, spherical harmonics P_3, minimum entropy, half-moment P_1, and our new model, which we will refer to as MME.

The numerical results for Fokker–Planck and diffusion have been obtained with standard finite differences. The hyperbolic models were solved with kinetic schemes. For more details on the numerical methods see [15].

In the one-dimensional setting, the Fermi–Eyges solution becomes particularly simple. First we can average over the directions parallel to the surfaces of the plate (perpendicular to the direction of propagation). Second, if we average over the angular variable, we obtain for the pencil-beam problem

$$\psi^{(0)}(x,\varepsilon;\varepsilon_0) = \delta(\varepsilon - \varepsilon(x;\varepsilon_0)). \tag{5.1}$$

For an arbitrary energy distribution ψ_b at the left boundary, the solution can be obtained by convolution:

$$\psi^{(0)}(x,\varepsilon) = \int_0^\infty \int_{-1}^1 \delta(\varepsilon - \varepsilon(x,\varepsilon'))\psi_b(0,\varepsilon',\mu)d\mu d\varepsilon'. \tag{5.2}$$

Our first (artificial) test case is the two-beam test case mentioned above. We ignore the complicated physics of the electron-tissue interaction and use the values $S_{\mathrm{M}} = 1$, $T_{\mathrm{M}} + T_{\mathrm{Mott}} = 0.01$. The spatial domain is $x \in [0,9]$; the cutoff energy (maximum energy) is 19.6. At the left side and at the right side of the domain, we consider forward-peaked incoming beams, represented by

$$\psi_b(x=0,\varepsilon,\mu) = \psi_0 e^{-(\varepsilon-\bar{\varepsilon})^2} e^{-100(1-\mu)^2}, \tag{5.3}$$

$$\psi_b(x=9,\varepsilon,\mu) = \psi_0 e^{-(\varepsilon-\bar{\varepsilon})^2} e^{-100(1+\mu)^2}, \tag{5.4}$$

with $\psi_0 = 10^5$. The models behave qualitatively similar if we take the values of water for S_{M} and $T_{\mathrm{M}} + T_{\mathrm{Mott}}$. With the values chosen above, however, their properties can be visualized in a single figure. The lowest order moment $\psi^{(0)}$ which can be seen as a measure for the total number of electrons, computed by the different models at $\varepsilon = 0.6\bar{\varepsilon}$, is shown in Figure 5.1.

The drawbacks of the classical macroscopic model are clearly visible. The diffusion approximation is smeared out, the spherical harmonics solution oscillates into the negative, and minimum entropy produces an unphysical shock. The half-moment model oscillates strongly, independently from the choice of the closure. The new MME model is a good approximation to the Fokker–Planck solution.

In our second example, we test our model in a real physical application. We consider a water phantom (an infinite plate filled with water) with depth 9 cm. The effective atomic number is $Z = 7.51$; the density of electrons is $\rho_{\mathrm{e}} = 3340 \cdot 10^{20}$ cm^3, $\rho_{\mathrm{c}} = \rho_{\mathrm{e}}/Z$, $\rho = 1g/\mathrm{cm}^3$ (parameters taken from [17]). The phantom is irradiated from one side by a beam

$$\psi_b(x=0,\varepsilon,\mu) = \psi_0 e^{-1(\varepsilon-\bar{\varepsilon})^2} e^{-10(1-\mu)^2} \tag{5.5}$$

with mean energy $\bar{\varepsilon} = 10$ MeV and amplitude $\psi_0 = 10^5$ MeV^{-1}s^{-1}. We compare the MME, diffusion, and Fermi–Eyges approximations with the Fokker–Planck solution. Snapshots of the energy spectrum of the beam for several depths are shown in Figure 5.2. As above, the diffusion approximation strongly smears out and loses the beam structure. The Fermi–Eyges approximation, on the other hand, retains the same shape as the incoming radiation. The energy spectrum is just gradually shifted

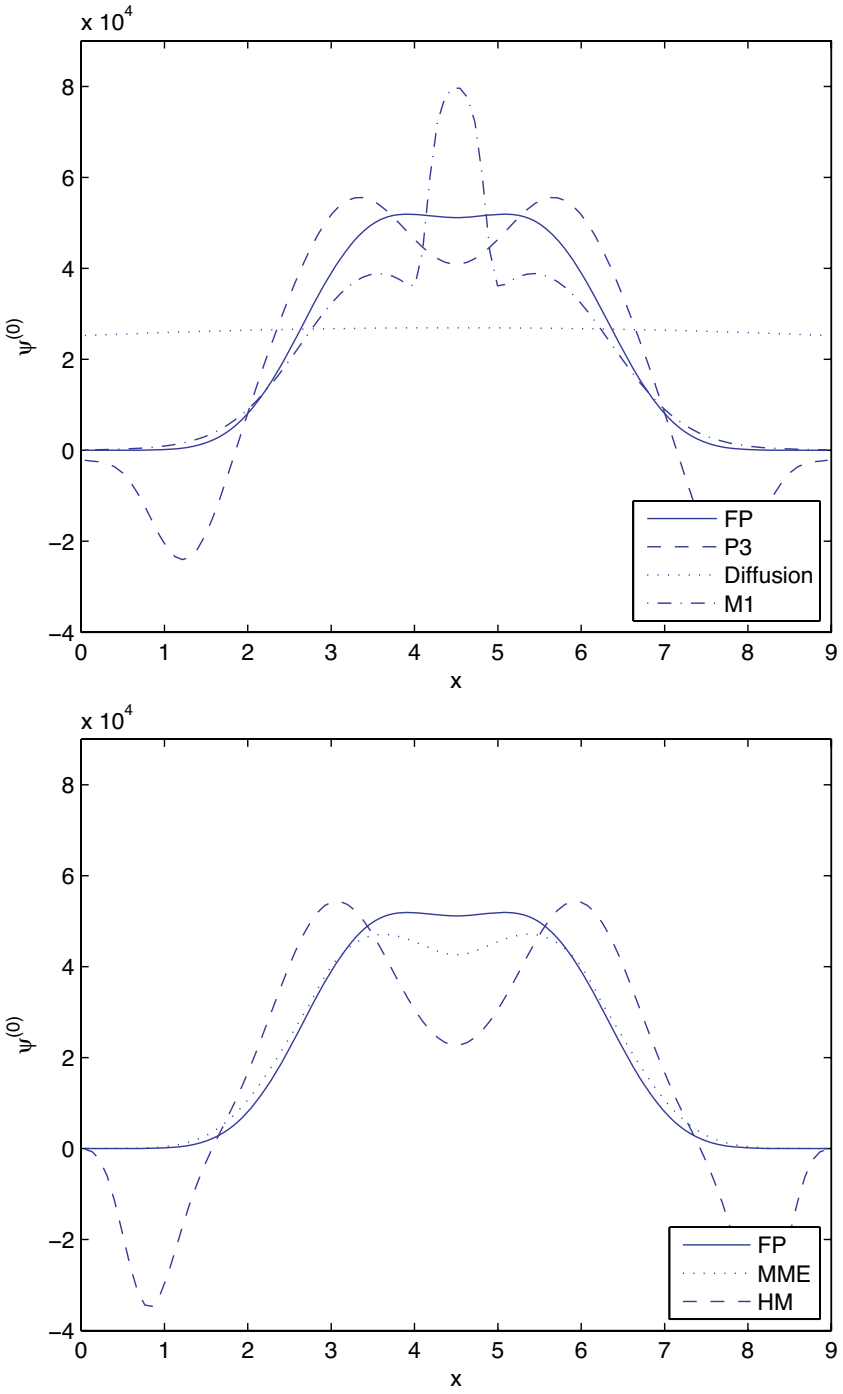


FIG. 5.1. *Distribution $\psi^{(0)}$ at $\varepsilon = 0.6\bar{\varepsilon}$ in two beam test case. Comparison between Fokker–Planck (FP), spherical harmonics P_3 (P3), diffusion, and minimum entropy M_1 (M1) solution (top). Comparison between Fokker–Planck (FP), MME, and half-moment (HM) solution (bottom).*

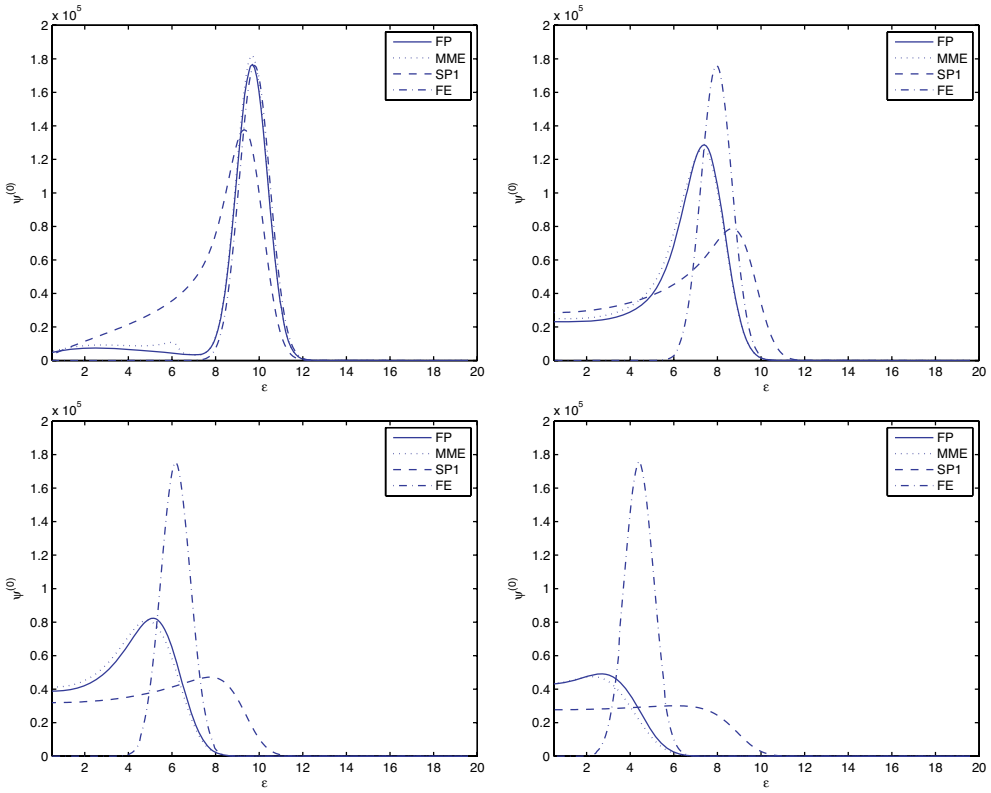


FIG. 5.2. *Snapshots of the energy spectrum. Distribution* $\psi^{(0)}$ *as a function of energy for different depths,* $x = 0.01$ *cm (top left),* $x = 1$ *cm (top right),* $x = 2$ *cm (bottom left), and* $x = 3$ *cm (bottom right). Comparison between Fokker–Planck solution (FP), MME, diffusion (SP*1*), and Fermi–Eyges (FE).*

according to the stopping power. The true behavior of the Fokker–Planck solution is inbetween the two extremes. The behavior is well matched by the MME approximation. The unmodified minimum entropy and the spherical harmonics solution (both not shown) give results similar to the MME model in the present case. The half-moment model oscillates strongly into negative energies.

For this test case, we also calculated the dose, using (2.1). The result is shown in Figure 5.3. As can be expected from the results above, the MME model is a good approximation to the dose computed using the Fokker–Planck solution. The diffusion is smeared out. The difference that can be seen, however, is less dramatic than for $\psi^{(0)}$. The Fermi–Eyges approximation is too simple to capture the absorption in the medium.

Finally, we should comment on the computation times the different models used. For the macroscopic models, the effort roughly scales with the number of equations: 1 for diffusion, 2 for minimum entropy, 3 for MME, 4 for spherical harmonics, and roughly 100 for Fokker–Planck, depending on the angular discretization. We did not optimize the convolution integrals in the Fermi–Eyges approximation, but it is expected that this method is faster than the diffusion approximation.

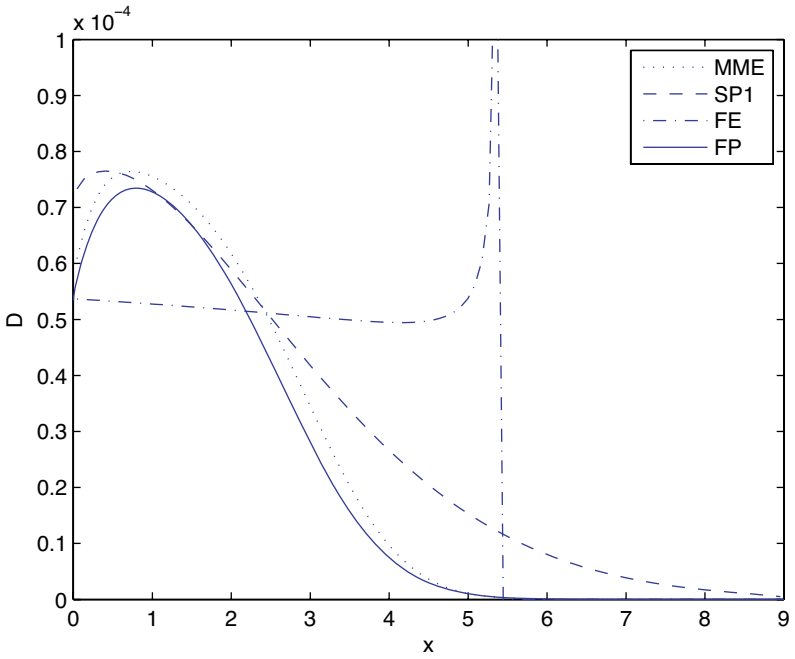


FIG. 5.3. *Dose deposited in tissue, computed with Fokker–Planck (FP), MME, diffusion (SP*1*), and Fermi–Eyges (FE).*

6. Conclusions. We derived and investigated a MME model as a computationally inexpensive but accurate model for dose calculation in electron radiotherapy. This model was derived from the Fokker–Planck equation and makes no assumption on the structure of the medium. In contrast to pencil-beam models, the treatment of inhomogeneities is model-inherent. Furthermore, we have demonstrated that the model preserves the positivity of the number of electrons and the shape of the energy spectrum better than several well-known macroscopic methods. We believe that the model is a first step toward a method which can compete with the techniques that are currently in clinical use.

Up to now numerical results have been obtained only in one dimension. The extension to two and three dimensions has been done with very good results in the case of radiative transfer, see, e.g., [15], using quarter space moments or other partial moment models. An extension of these methods to the present case is currently under investigation.

As mentioned in the beginning, the Fokker–Planck model lacks parts of the necessary physics for the description of electron radiotherapy. Some electrons will also experience hard collisions which have to be described by Boltzmann integral terms. Thus, the development of approximate methods for the original Boltzmann equation with forward-peaked scattering is a subject which has to be investigated as well.

Several open questions for further studies remain:

- The analytical properties of the new model, especially mathematical well-posedness, will be studied. We expect to obtain similar results as for the standard half-moment model [33, 16].

- The passage from the Boltzmann transport equation to the Fokker–Planck equation in the context of moment models should be studied in more detail. The half-moment model apparently breaks down in this limit. The exact reason for this breakdown is certainly an interesting subject of further study.
- The model will be generalized to two and three space dimensions. Results in the case of the radiative transfer equation [15] indicate that the advantage over diffusion and spherical harmonics methods becomes even more apparent.
- Moment models for the full Boltzmann equation will be developed and compared to the present approach.
- The model contains an accurate model of the microscopic interactions. In two and three dimensions, it should be validated using well-established codes for electron radiotherapy, experiments, and benchmark results.

Appendix A. Explicit formulas for the cross sections. This appendix lists all scattering cross sections that are used in the Boltzmann model. All cross sections are calculated for the laboratory system where the scattering centers are at rest before scattering. Except for elastic Mott scattering, all differential scattering cross sections are differential in energy and in solid angle. They can be decomposed into a product of a cross section, which is only differential in solid angle or energy, and a Dirac delta function, which guarantees energy and momentum conservation during the scattering event. The Mott cross section is only differential in solid angle. Total cross sections are calculated by integrating the double differential cross sections with respect to energy and solid angle. Because of the delta functions one integration is always trivial.

For all differential cross sections the following conventions are used: quantities with a prime belong to incoming particles, quantities without a prime to outgoing particles in a scattering event. The order of appearence in all differential cross sections is energy of incoming (ε'), energy of outgoing (ε), direction of incoming (Ω'), and direction of outgoing particles (Ω). To simplify notation and to keep the standard notation used in literature we use $\Omega' \cdot \Omega = \cos\vartheta \equiv \mu$, ϑ being the scattering angle in the laboratory system. Additionally we keep $\sin\vartheta$ and $\tan\vartheta$ in formulas to maintain a handy notation. The relationship to $\cos\vartheta$ is evident. Furthermore it should be kept in mind that $\epsilon_{\mathrm{B}} = \epsilon_{\mathrm{B}}(r)$ which is not explicitly written to keep notation short. In all formulas the classical electron radius $r_{\mathrm{e}} = 2.8179 \cdot 10^{-15}\,\mathrm{m}$ appears.

A.1. Differential cross section for Compton scattering of photons [10].

$$\tilde{\sigma}_{\mathrm{C},\gamma}(\epsilon'_\gamma, \epsilon_\gamma, \Omega'_\gamma \cdot \Omega_\gamma) = \sigma_{\mathrm{C},\gamma}(\epsilon'_\gamma, \Omega'_\gamma \cdot \Omega_\gamma)\delta_{\mathrm{C},\gamma}(\epsilon'_\gamma, \epsilon_\gamma)$$

with

$$\sigma_{\mathrm{C},\gamma}(\epsilon'_\gamma, \Omega'_\gamma \cdot \Omega_\gamma) = \frac{r_{\mathrm{e}}^2}{2}\left[\frac{1}{1+\epsilon'_\gamma(1-\cos\vartheta_\gamma)}\right]^2\left[1+\cos^2\vartheta_\gamma + \frac{\epsilon'^2_\gamma(1-\cos\vartheta_\gamma)^2}{1+\epsilon'_\gamma(1-\cos\vartheta_\gamma)}\right]$$

$$\delta_{\mathrm{C},\gamma}(\epsilon'_\gamma, \epsilon_\gamma) := \delta_\gamma\left(\epsilon_\gamma - \frac{\epsilon'_\gamma}{1+\epsilon'_\gamma(1-\cos\vartheta_\gamma)}g\right).$$

A.2. Total cross section for Compton scattering of photons [10].

$$\sigma^{\mathrm{tot}}_{\mathrm{C},\gamma}(\epsilon_\gamma) = 2\pi r_{\mathrm{e}}^2\left[\frac{1+\epsilon_\gamma}{\epsilon_\gamma^2}\left(\frac{2(1+\epsilon_\gamma)}{1+2\epsilon_\gamma} - \frac{1}{\epsilon_\gamma}\ln(1+2\epsilon_\gamma)\right) + \frac{1}{2\epsilon_\gamma}\ln(1+2\epsilon_\gamma) - \frac{1+3\epsilon_\gamma}{(1+2\epsilon_\gamma)^2}\right].$$

A.3. Differential cross section for Compton scattering of electrons [10].

$$\tilde{\sigma}_{\mathrm{C,e}}(\epsilon'_\gamma, \epsilon_\mathrm{e}, \Omega'_\gamma \cdot \Omega_\mathrm{e}) = \sigma_{\mathrm{C,e}}(\epsilon'_\gamma, \Omega'_\gamma \cdot \Omega_\mathrm{e})\delta_{\mathrm{C,e}}(\epsilon'_\gamma, \epsilon_\mathrm{e})$$

with

$$\begin{aligned}\sigma_{\mathrm{C,e}}(\epsilon'_\gamma, \Omega'_\gamma \cdot \Omega_\mathrm{e}) &= \frac{4r_\mathrm{e}^2(1+\epsilon'_\gamma)^2}{\cos^3\vartheta_\mathrm{e}} \frac{1}{(a(\epsilon'_\gamma, \vartheta_\mathrm{e}) + 2\epsilon'_\gamma)^2} \\ &\quad \times \left[1 - \frac{2}{a(\epsilon'_\gamma, \vartheta_\mathrm{e})} + \frac{2}{a^2(\epsilon'_\gamma, \vartheta_\mathrm{e})} + \frac{2\epsilon_\gamma^2}{a(\epsilon'_\gamma \vartheta_\mathrm{e})(a(\epsilon'_\gamma, \vartheta_\mathrm{e}) + 2\epsilon'_\gamma)}\right] \\ \delta_{\mathrm{C,e}}(\epsilon'_\gamma, \epsilon_\mathrm{e}) &:= \delta_\gamma\left(\epsilon_\mathrm{e} - \frac{2{\epsilon'_\gamma}^2}{2\epsilon'_\gamma + a(\epsilon'_\gamma, \vartheta_\mathrm{e})} g\right),\end{aligned}$$

where

$$a(\epsilon'_\gamma, \vartheta_\mathrm{e}) := (1+\epsilon'_\gamma)^2 \tan^2\vartheta_\mathrm{e} + 1.$$

A.4. Differential cross section for Møller scattering of primary electrons, i.e., $\epsilon_\mathrm{e} > (\epsilon'_\mathrm{e} - \epsilon_\mathrm{B})/2$ [23].

$$\tilde{\sigma}_\mathrm{M}(\epsilon'_\mathrm{e}, \epsilon_\mathrm{e}, \Omega'_\mathrm{e} \cdot \Omega_\mathrm{e}) = \sigma_\mathrm{M}(\epsilon'_\mathrm{e}, \epsilon_\mathrm{e})\delta_M(\mu_\mathrm{e}, \mu_p)\frac{1}{2\pi}, \quad \mu_\mathrm{e} = \Omega'_\mathrm{e} \cdot \Omega_\mathrm{e}$$

with

$$\begin{aligned}\sigma_\mathrm{M}(\epsilon'_\mathrm{e}, \epsilon_\mathrm{e}) &= \frac{2\pi r_\mathrm{e}^2(\epsilon'_\mathrm{e}+1)^2}{\epsilon'_\mathrm{e}(\epsilon'_\mathrm{e}+2)}\left[\frac{1}{\epsilon_\mathrm{e}^2} + \frac{1}{(\epsilon'_\mathrm{e}-\epsilon_\mathrm{e})^2} + \frac{1}{(\epsilon'_\mathrm{e}+1)^2} - \frac{2\epsilon'_\mathrm{e}+1}{(\epsilon'_\mathrm{e}+1)^2\epsilon_\mathrm{e}(\epsilon'_\mathrm{e}-\epsilon_\mathrm{e})}\right] \\ \delta_M(\mu_\mathrm{e}, \mu_p) &= \delta\left(\mu_\mathrm{e} - \sqrt{\frac{\epsilon_\mathrm{e}}{\epsilon'_\mathrm{e}}\frac{\epsilon'_\mathrm{e}+2}{\epsilon_\mathrm{e}+2}}\right), \quad \epsilon_\mathrm{e} > \frac{(\epsilon'_\mathrm{e} - \epsilon_\mathrm{B})}{2}.\end{aligned}$$

A.5. Differential cross section for Møller scattering of secondary electrons, i.e., $\epsilon_\mathrm{e} < (\epsilon'_\mathrm{e} - \epsilon_\mathrm{B})/2$ [23].

$$\tilde{\sigma}_{\mathrm{M},\delta}(\epsilon'_\mathrm{e}, \epsilon_\mathrm{e}, \Omega'_\mathrm{e} \cdot \Omega_\mathrm{e}) = \sigma_\mathrm{M}(\epsilon'_\mathrm{e}, \epsilon_\mathrm{e})\delta_{\mathrm{M},\delta}(\mu_\mathrm{e}, \mu_\delta)\frac{1}{2\pi}, \quad \mu_\mathrm{e} = \Omega'_\mathrm{e} \cdot \Omega_\mathrm{e}$$

with

$$\begin{aligned}\sigma_\mathrm{M}(\epsilon'_\mathrm{e}, \epsilon_\mathrm{e}) &= \frac{2\pi r_\mathrm{e}^2(\epsilon'_\mathrm{e}+1)^2}{\epsilon'_\mathrm{e}(\epsilon'_\mathrm{e}+2)}\left[\frac{1}{\epsilon_\mathrm{e}^2} + \frac{1}{(\epsilon'_\mathrm{e}-\epsilon_\mathrm{e})^2} + \frac{1}{(\epsilon'_\mathrm{e}+1)^2} - \frac{2\epsilon'_\mathrm{e}+1}{(\epsilon'_\mathrm{e}+1)^2\epsilon_\mathrm{e}(\epsilon'_\mathrm{e}-\epsilon_\mathrm{e})}\right] \\ \delta_{\mathrm{M},\delta}(\mu_\mathrm{e}, \mu_\delta) &= \delta\left(\mu_\mathrm{e} - \sqrt{\frac{\epsilon_\mathrm{e}}{\epsilon'_\mathrm{e}}\frac{\epsilon'_\mathrm{e}+2}{\epsilon_\mathrm{e}+2}}\right), \quad \epsilon_\mathrm{e} < \frac{(\epsilon'_\mathrm{e} - \epsilon_\mathrm{B})}{2}.\end{aligned}$$

A.6. Total cross section for Møller scattering of electrons [23].

$$\sigma_\mathrm{M}^\mathrm{tot}(\epsilon_\mathrm{e}) = \int_{\epsilon_\mathrm{B}}^{(\epsilon_\mathrm{e}-\epsilon_\mathrm{B})/2} \sigma_\mathrm{M}(\epsilon_\mathrm{e}, \epsilon'_\mathrm{e}) d\epsilon'_\mathrm{e}.$$

The lower limit of integration is due to the fact that the primary electron can be scattered only if at least the binding energy ϵ_B is transferred to the secondary electron

(of a tissue molecule). Besides the evident motivation of this choice based on our model, this is a standard way to avoid singularities in calculating total cross sections (see, e.g., [39]). The upper limit of integration is due to the fact that the primary electron has larger energy than the secondary electron and that the binding energy ϵ_{B} was introduced into the scattering processes (usually the upper limit is $\epsilon'_{\mathrm{e}}/2$). One gets

$$\sigma_{\mathrm{M}}^{\mathrm{tot}}(\epsilon_{\mathrm{e}}) = \frac{2\pi r_{\mathrm{e}}^2(\epsilon_{\mathrm{e}}+1)^2}{\epsilon_{\mathrm{e}}(\epsilon_{\mathrm{e}}+2)} \times \left\{ \frac{1}{\epsilon_{\mathrm{B}}} - \frac{3}{\epsilon_{\mathrm{e}}-\epsilon_{\mathrm{B}}} + \frac{2}{\epsilon_{\mathrm{e}}+\epsilon_{\mathrm{B}}} + \frac{\epsilon_{\mathrm{e}}-3\epsilon_{\mathrm{B}}}{2(\epsilon_{\mathrm{e}}+1)^2} + \frac{2\epsilon_{\mathrm{e}}+1}{\epsilon_{\mathrm{e}}(\epsilon_{\mathrm{e}}+1)} g\left[\ln\frac{\epsilon_{\mathrm{e}}+\epsilon_{\mathrm{B}}}{\epsilon_{\mathrm{e}}-\epsilon_{\mathrm{B}}} - \ln\frac{\epsilon_{\mathrm{e}}-\epsilon_{\mathrm{B}}}{\epsilon_{\mathrm{B}}} g \right] \right\}.$$

A.7. Differential cross section for Mott scattering of electrons [28, 26]. $\alpha \approx 1/137$ is the fine structure constant, and Z is the atomic number of the irradiated medium. Z depends on r to account for heterogeneous media.

$$\begin{aligned}\sigma_{\mathrm{Mott}}(r,\epsilon_{\mathrm{e}},\Omega'_{\mathrm{e}}\cdot\Omega_{\mathrm{e}}) &= \frac{Z^2(r) r_{\mathrm{e}}^2 (mc^2)^2}{4p^2c^2\beta^2 \sin^4\frac{\vartheta_{\mathrm{e}}}{2}} \left[1 - \beta^2 \sin^2\frac{\vartheta_{\mathrm{e}}}{2} + Z\pi\alpha\beta \sin\frac{\vartheta_{\mathrm{e}}}{2}\left(1-\sin\frac{\vartheta_{\mathrm{e}}}{2}\right)\right] \\ &\approx \frac{Z^2(r) r_{\mathrm{e}}^2 (mc^2)^2}{4p^2c^2\beta^2 \sin^4\frac{\vartheta_{\mathrm{e}}}{2}} \left[1 - \beta^2 \sin^2\frac{\vartheta_{\mathrm{e}}}{2}\right],\end{aligned}$$

with $\beta^2 = \frac{\epsilon_{\mathrm{e}}(\epsilon_{\mathrm{e}}+2)}{(\epsilon_{\mathrm{e}}+1)^2}$. The last approximation is justified, because in the energy range studied here and for typical low-Z media like water only small errors are made.

To avoid the singularity at $\vartheta_{\mathrm{e}} = 0$ a screening parameter η can be introduced [40] that models the screening effect of the electrons of the atomic shell:

$$\sigma_{\mathrm{Mott}}(r,\epsilon_{\mathrm{e}},\Omega'_{\mathrm{e}}\cdot\Omega_{\mathrm{e}}) = \frac{Z^2(r) r_{\mathrm{e}}^2 (1+\epsilon_{\mathrm{e}})^2}{4[\epsilon_{\mathrm{e}}(\epsilon_{\mathrm{e}}+2)]^2(1+2\eta(r,\epsilon_{\mathrm{e}})-\cos\vartheta_{\mathrm{e}})^2} \left[1 - \frac{\epsilon_{\mathrm{e}}(\epsilon_{\mathrm{e}}+2)}{(1+\epsilon_{\mathrm{e}})^2} \sin^2\frac{\vartheta_{\mathrm{e}}}{2}\right]$$

with

$$\eta(r,\epsilon_{\mathrm{e}}) = \frac{\pi^2\alpha^2 Z^{\frac{2}{3}}(r)}{\epsilon_{\mathrm{e}}(\epsilon_{\mathrm{e}}+2)}.$$

A.8. Total cross section for Mott scattering of electrons.

$$\sigma_{\mathrm{Mott}}^{\mathrm{tot}}(r,\epsilon_{\mathrm{e}}) = \frac{\pi(Z(r)r_{\mathrm{e}})^2}{\epsilon_{\mathrm{e}}(\epsilon_{\mathrm{e}}+2)} \times \left[\frac{(\epsilon_{\mathrm{e}}+1)^2}{(\pi\alpha)^2 Z^{2/3}(r)(1+\eta(r,\epsilon_{\mathrm{e}}))} + \frac{1}{1+\eta(r,\epsilon_{\mathrm{e}})} + \ln\eta(r,\epsilon_{\mathrm{e}}) - \ln(1+\eta(r,\epsilon_{\mathrm{e}}))\right]$$

REFERENCES

[1] A. Ahnesjö and M. M. Aspradakis, *Dose calculations for external photon beams in radiotherapy*, Phys. Med. Biol., 44 (1999), pp. R99–R155.

[2] P. Andreo, *Monte Carlo techniques in medical radiation physics*, Phys. Med. Biol., 36 (1991), pp. 861–920.

[3] A. M. Anile, S. Pennisi, and M. Sammartino, *A thermodynamical approach to Eddington factors*, J. Math. Phys., 32 (1991), pp. 544–550.

[4] E. D. Aydin, C. R. E. Oliveira, and A. J. H. Goddard, *A comparison between transport and diffusion calculations using finite element-spherical harmonics radiation transport method*, Med. Phys., 29 (2002), pp. 2013–2023.

[5] C. Börgers and E. W. Larsen, *The transversely integrated scalar flux of a narrowly focused particle beam*, SIAM J. Appl. Math., 55 (1995), pp. 1–22.
[6] C. Börgers and E. W. Larsen, *Asymptotic derivation of the Fermi pencil-beam approximation*, Nuclear Sci. Eng., 2123 (1996), pp. 343–357.
[7] C. Börgers and E. W. Larsen, *On the accuracy of the Fokker-Planck and Fermi pencil-beam equations for charged particle transport*, Med. Phys., 23 (1996), pp. 1749–1759.
[8] T. A. Brunner and J. P. Holloway, *One-dimensional Riemann solvers and the maximum entropy closure*, J. Quant. Spectrosc. Radiat. Transfer, 69 (2001), pp. 543–566.
[9] M. Caro and J. Ligou, *Treatment of scattering anisotropy of neutrons through the boltzmann-fokker-planck equation*, Nuclear Sci. Eng., 83 (1983), pp. 242–252.
[10] C. M. Davisson and R. D. Evans, *Gamma-ray absorption coefficients*, Rev. Modern Phys., 24 (1952), p. 79.
[11] B. Dubroca and J. L. Feugeas, *Entropic moment closure hierarchy for the radiative transfer equation*, C. R. Acad. Sci. Paris Ser. I, 329 (1999), pp. 915–920.
[12] B. Dubroca and A. Klar, *Half moment closure for radiative transfer equations*, J. Comput. Phys., 180 (2002), pp. 584–596.
[13] A. Eddington, *The Internal Constitution of the Stars*, Dover, New York, 1926.
[14] L. Eyges, *Multiple scattering with energy loss*, Phys. Rev., 74 (1948), pp. 1534–1535.
[15] M. Frank, B. Dubroca, and A. Klar, *Partial moment entropy approximation to radiative transfer*, J. Comput. Phys., 218 (2006), pp. 1–18.
[16] M. Frank and R. Pinnau, *Existence, uniqueness and bounds for the half moment minimum entropy approximation to rediative heat transfer*, Appl. Math. Lett., 20 (2007), pp. 189–193.
[17] H. Hensel, R. Iza-Teran, and N. Siedow, *Deterministic model for dose calculation in photon radiotherapy*, Phys. Med. Biol., 51 (2006), pp. 675–693.
[18] K. R. Hogstrom, M. D. Mills, and P. R. Almond, *Electron beam dose calculations*, Phys. Med. Biol., 26 (1981), pp. 445–459.
[19] H. Huizenga and P. Storchi, *Numerical calculation of energy deposition by broad high-energy electron beams*, Phys. Med. Biol., 34 (1989), p. 1371.
[20] J. Janssen, D. Riedeman, M. Morawska-Kaczynska, P. Storchi, and H. Huizenga, *Numerical calculation of energy deposition by broad high-energy electron beams:* III. *Three-dimensional heterogeneous media*, Phys. Med. Biol., 39 (1994), p. 1351.
[21] J. H. Jeans, *The equations of radiative transfer of energy*, Monthly Notices Roy. Astronom. Soc., 78 (1917), pp. 28–36.
[22] D. Jette, *Electron dose calculations using multiple-scattering theory. A new theory of multiple scattering*, Med. Phys., 23 (1996), pp. 459–477.
[23] I. Kawrakow and D. W. O. Rogers, *The EGSnrc code system: Monte Carlo simulation of electron and photon transport*, Technical report PIRS-701, National Research Council of Canada, Ottawa, CN, 2002.
[24] E. W. Larsen, M. M. Miften, B. A. Fraass, and I. A. D. Bruinvis, *Electron dose calculations using the method of moments*, Med. Phys., 24 (1997), pp. 111–125.
[25] E. Larsen and J. Keller, *Asymptotic solution of neutron transport problems for small mean free path*, J. Math. Phys., 15 (1974), p. 75.
[26] C. Lehmann, *Interaction of Radiation with Solids and Elementary Defect Production*, North–Holland, Amsterdam, 1977.
[27] M. Morawska-Kaczynska and H. Huizenga, *Numerical calculation of energy deposition by broad high-energy electron beams:* II. *Multi-layered geometry*, Phys. Med. Biol., 37 (1992), p. 2103.
[28] N. F. Mott and H. S. W. Massey, *The Theory of Atomic Collisions*, Clarendon Press, Oxford, 1965.
[29] I. Müller and T. Ruggeri, *Rational Extended Thermodynamics*, 2nd ed., Springer-Verlag, New York, 1993.
[30] G. C. Pomraning, *The Fokker-Planck operator as an asymptotic limit*, Math. Models Methods Appl. Sci., 2 (1992), pp. 21–36.
[31] K. Przybylski and J. Ligou, *Numerical analysis of the Boltzmann equation including Fokker-Planck terms*, Nuclear Sci. Eng., 81 (1982), pp. 92–109.
[32] B. Rossi and K. Greisen, *Cosmic-ray theory*, Rev. Modern Phys., 13 (1941), pp. 240–309.
[33] M. Schäfer, M. Frank, and R. Pinnau, *A hierarchy of approximations to the radiative heat transfer equations: Modeling, analysis and simulation*, Math. Models Methods Appl. Sci., 15 (2005), pp. 643–665.
[34] D. M. Shepard, M. C. Ferris, G. H. Olivera, and T. R. Mackie, *Optimizing the delivery of radiation therapy to cancer patients*, SIAM Rev., 41 (1999), pp. 721–744.
[35] P. Storchi and H. Huizenga, *On a numerical approach of the pencil beam model*, Phys. Med. Biol., 30 (1985), p. 467.

[36] J. Tervo, P. Kolmonen, M. Vauhkonen, L. M. Heikkinen, and J. P. Kaipio, *A finite-element model of electron transport in radiation therapy and related inverse problem*, Inverse Problems, 15 (1999), pp. 1345–1361.

[37] J. Tervo and P. Kolmonen, *Inverse radiotherapy treatment planning model applying Boltzmann-transport equation*, Math. Models Methods Appl. Sci., 12 (2002), pp. 109–141.

[38] R. Turpault, M. Frank, B. Dubroca, and A. Klar, *Multigroup half space moment approximations to the radiative heat transfer equations*, J. Comput. Phys., 198 (2004), pp. 363–371.

[39] M. M. R. Williams, *The role of the boltzmann transport equation in radiation damage calculations*, Prog. Nuclear Energy, 3 (1979), pp. 1–65.

[40] C. D. Zerby and F. L. Keller, *Electron transport theory, calculations and experiments*, Nuclear Sci. Eng., 27 (1967), pp. 190–218.

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 605–618

NONLINEAR ANALYSIS IN THE AW–RASCLE ANTICIPATION MODEL OF TRAFFIC FLOW*

ZHONG-HUI OU†, SHI-QIANG DAI†, PENG ZHANG†, AND LI-YUN DONG†

Abstract. In this paper, the Aw–Rascle anticipation (ARA) model is discussed from the perspective of the capability to reproduce nonlinear traffic flow behaviors observed in real traffic. For this purpose, a nonlinear traffic flow stability criterion is derived by using a wavefront expansion technique. The result of the nonlinear stability analysis can be used not only to judge the stability evolution of an initial traffic state but also to determine the pressure term in the ARA model. The KdV equation is derived from the ARA model added by the viscous term with the use of the reduction perturbation method. The soliton solution can be analytically obtained from the perturbed KdV equation only near the neutral stability line. Weighted essentially nonoscillatory schemes are employed to simulate the KdV soliton. The numerical results confirm the analytical KdV soliton solution.

Key words. traffic flow, soliton, WENO

AMS subject classifications. 35L65, 41A58

DOI. 10.1137/060656863

1. Introduction.

A macroscopic model of vehicular traffic started with the first-order fluid approximation of traffic flow dynamics proposed by Lighthill, Whitham [1], and Richards [2] independently, i.e., the Lighthill–Whitham–Richards (LWR) model [3], which assumes the conservation of the number of vehicles and the equilibrium relation between flow and density. However, besides the continuity equation, one needs an extra dynamic velocity equation in order to describe the emergent traffic jams and stop-and-go traffic. This kind of two-equation model includes the Payne–Whitham (PW) model [4, 5, 1], the Kühne model [6, 7], the Kerner–Konhäuser (KK) model [8, 9], the Lee–Lee–Kim (LLK) model [10, 11], the gas-kinetic-based (GKT) model [12, 13, 14, 15], etc. But Daganzo pointed out that one characteristic velocity greater than the macroscopic fluid velocity in the two-equation models would lead to nonphysical effects [16]. Aw and Rascle replaced the space derivative of the "pressure" with a convective derivative in PW-type models to resolve the theoretical inconsistencies and then constructed the Aw–Rascle anticipation (ARA) model [17, 18, 19]. They discussed the solution to the Riemann problem and the admissibility of the elementary waves by the hyperbolic conservation laws in detail. Moreover, the ARA model is the typical form of the anisotropic traffic flow model, which can be reduced to other models from different viewpoints [20, 21].

A traffic flow model usually needs stability analysis and numerical simulation in the stable and unstable density regions in order to investigate the evolution of a traffic initial state [8, 12, 22, 20, 23, 24, 25, 26]. Since in the linear stability analysis higher-order terms are neglected, we propose a nonlinear stability analysis by utilizing

*Received by the editors April 10, 2006; accepted for publication (in revised form) October 19, 2006; published electronically February 23, 2007. This work was financially supported by the National Basic Research Program of China (the 973 Program) under grant 2006CB705500, the National Natural Science Foundation of China under grant 10532060, the Shanghai Postdoctoral Scientific Program, and the Shanghai Leading Academic Discipline Project (Y0103).

http://www.siam.org/journals/siap/67-3/65686.html

†Shanghai Institute of Applied Mathematics and Mechanics, Shanghai University, Shanghai 200072, People's Republic of China (ouzhonghui@vip.sina.com, sqdai@shu.edu.cn, pzhang@mail.shu.edu.cn, lydong@mail.shu.edu.on).

a wavefront expansion technique under large traffic disturbances. Moreover, some typical nonlinear waves such as the KdV soliton, triangular shock, and kink have been found through investigating the car-following models with the reduction perturbation method [27, 28, 29], but only Kurtze and Hong [30] and Berg and Woods [31] did similar work in the continuum model. The reduction perturbation method is more difficult to conduct in the continuum model than in the car-following model because the former has a much more complicated form than the latter. Therefore it is valuable for the anisotropic ARA model to make the nonlinear stability and wave analysis in traffic flow.

This paper is arranged as follows. We take the nonlinear stability analysis on the ARA model, and one stability criterion is used to determine the pressure term in section 2. The KdV equation is derived from the ARA model with the viscous term in section 3. Weighted essentially nonoscillatory schemes (WENO) schemes are used to simulate the KdV soliton in section 4. Concluding remarks are presented in section 5.

2. Nonlinear stability analysis. The ARA model is

$$\partial_t \rho + \partial_x(\rho v) = 0, \tag{1}$$

$$\partial_t(v + p(\rho)) + v\partial_x(v + p(\rho)) = \tau^{-1}(V(\rho) - v), \tag{2}$$

where $\rho(x,t)$ is the density at point x and time t, $v(x,t)$ is the velocity, τ is the relaxation time, $V(\rho)$ is the equilibrium function, $p(\rho) = \rho^\gamma$ is the pressure (the anticipation factor is more accurate), and γ is a positive constant which needs to be determined later [17]. The ARA model under conservative form is

$$\partial_t Y + \partial_x(f(Y)) = g(Y), \tag{3}$$

where the conservative vectors $Y = (\rho, y) = (\rho, \rho(v+p(\rho)))$, $f(Y) = (\rho v, \rho v(v+p(\rho)))$, and $g(Y) = \big(0, \tau^{-1}\rho(V(\rho) - v)\big)$.

The eigenvalues of (1) and (2) are

$$\lambda_1 = v - \rho p^{'}(\rho) \le \lambda_2 = v. \tag{4}$$

Formula (4) shows that all the waves propagate at a speed at most equal to the velocity v of the corresponding state. Daganzo has pointed out that continuum models (especially the hydrodynamic models) with one characteristic speed greater than the macroscopic fluid velocity encounter difficulties showing nonphysical effects in certain situations [16]. But the ARA model avoids these difficulties.

When the traffic jam happens, the density of the traffic jam is much higher than that of the neighboring section, so most traffic jams belong to large disturbances. When the disturbance is fairly large, the linearization method may produce incorrect results because of neglecting higher-order terms as pointed out by Whitham [1], which is the primary cause we propose a nonlinear stability analysis for traffic flow [1, 22, 23, 24]. If a disturbance starts at position x_0 in the homogeneous state of traffic flow, the wavefront is the propagation curve of the disturbance along the homogeneous traffic flow [1, 22, 23, 24]. The magnitude of the initial disturbance will not increase during its propagation if the traffic system is stable in propagation; otherwise, a disturbed density or velocity wave may increase in magnitude as it propagates upstream and ultimately form a shock wave or traffic jam on the highway. If the form of the initial density disturbance is given, the initial velocity disturbance can be determined by the equilibrium function. Furthermore, the profiles of density and velocity disturbances

are symmetrical to some degree along the x-axis in the simulation with the continuum model (e.g., monotone increase vs. monotone decrease, concavity vs. convexity, etc.) [8, 9, 15], which is still considered valid in the ARA model, and this fact will also be demonstrated by the following numerical simulation of the KdV soliton. Therefore, when we merely demand some mathematical assumptions for the density, they are automatically needed for the velocity, and most results are presented for either the density or the velocity. Assume that a disturbance initiates from the equilibrium state (ρ_0, v_0), the solutions of (1) and (2), in the homogeneous traffic flow. If the mth derivatives of ρ around the wavefront WF are the first ones to be discontinuous, the expanded Taylor series from ρ_0 starts with the term in the mth power of a small parameter, which is equally assumed for v_0. Without losing generality, we might as well assume that the first derivative of the density around the wavefront WF is discontinuous.

It is convenient to expand the solution of the system around the wavefront WF in powers of

$$\xi = x - X(t), \tag{5}$$

where $X(t)$ is the location of the wavefront WF at time t. Since the wavefront is the boundary of the disturbance in the homogeneous state, the characteristic method is still feasible in the near neighborhood of the wavefront WF. Therefore the wavefront has the characteristic velocity $v_{c1,2}$ in the equilibrium states, i.e.,

$$\dot{X}(t) = v_{c1,2}(\rho_0, v_0) = v_0 - \lambda_{1,2}. \tag{6}$$

Using (5), we can expand the flow variables ρ, v, their partial derivatives, $V(\rho)$ and $p_\rho(\rho)$, etc., in the power series of ξ as

$$\rho(x,t) = \rho_0 + \xi\rho_1(t) + \frac{1}{2}\xi^2\rho_2(t) + \cdots, \tag{7}$$

$$v(x,t) = v_0 + \xi v_1(t) + \frac{1}{2}\xi^2 v_2(t) + \cdots, \tag{8}$$

where

$$\rho_i(t) = \frac{\partial^i \rho}{\partial x^i}\bigg|_{(X(t)^-,t)}, \quad v_i(t) = \frac{\partial^i v}{\partial x^i}\bigg|_{(X(t)^-,t)}, \quad i = 1, 2, 3, \ldots,$$

$$\rho_t = -\dot{X}(t)\rho_1(t) + \xi\dot{\rho}_1(t) + \xi\left[-\dot{X}(t)\right]\rho_2(t) + \frac{1}{2}\xi^2\dot{\rho}_2(t) + \cdots, \tag{9}$$

$$\rho_x = \rho_1(t) + \xi\rho_2(t) + \frac{1}{2}\xi^2\rho_3(t) + \cdots, \tag{10}$$

$$v_t = -\dot{X}(t)v_1(t) + \xi\dot{v}_1(t) + \xi\left[-\dot{X}(t)\right]v_2(t) + \frac{1}{2}\xi^2\dot{v}_2(t) + \cdots, \tag{11}$$

$$v_x = v_1(t) + \xi v_2(t) + \frac{1}{2}\xi^2 v_3(t) + \cdots. \tag{12}$$

$$V(\rho) = V^0 + \xi V^0_\rho \rho_1(t) + \cdots, \tag{13}$$

where

$$V^0 = V(\rho_0) \quad \text{and} \quad V^0_\rho = \frac{\partial V}{\partial \rho}\bigg|_{(\rho_0, v_0)}.$$

$$p_\rho(\rho) = p^0_\rho + \xi p^0_{\rho\rho}\rho_1(t) + \cdots. \tag{14}$$

Substituting (7)–(14) into (1) and (2), for the coefficients of the first two terms ξ^0 and ξ^1, we obtain

$$u_0\rho_1 + \rho_0 v_1 = 0, \tag{15}$$

$$\dot{\rho}_1 + 2\rho_1 v_1 + u_0\rho_2 + \rho_0 v_2 = 0, \tag{16}$$

$$v_1 + p^0_\rho \rho_1 = 0, \tag{17}$$

$$u_0 v_2 + u_0 p^0_\rho \rho_2 + \dot{v}_1 + v_1^2 + p^0_\rho \dot{\rho}_1 + u_0 p^0_{\rho\rho} {\rho_1}^2 + v_1 p^0_\rho \rho_1 + \tau^{-1}\left(v_1 - V^0_\rho \rho_1\right) = 0, \tag{18}$$

where $u_0 = \lambda_{1,2}$.

Substituting (15) into (17) yields

$$u_0 - \rho_0 p^0_\rho = 0, \tag{19}$$

which shows that the coefficients of terms ρ_2 and v_2 are linearly dependent and can be eliminated from (16) and (18). Inserting $\rho_1 = -\rho_0 v_1/u_0$ obtained from (15) into (16) and (18) leads to the Bernoulli equation

$$\dot{v}_1 + \alpha v_1 + \beta v_1^2 = 0, \tag{20}$$

where

$$\alpha = \tau^{-1}\left(1 + \frac{V^0_\rho}{p^0_\rho}\right) \quad \text{and} \quad \beta = 2 + \frac{\rho_0 p^0_{\rho\rho}}{p^0_\rho} = \gamma + 1.$$

$\beta > 0$, if $p(\rho) = \rho^\gamma, \gamma > 0$.

If $\alpha = 0$, the solution of (20) is

$$v_1(t) = \frac{v_1(0)}{\beta v_1(0) t + 1}, \tag{21}$$

whose monotonicity is determined by

$$v_1'(t) = -\frac{\beta v_1^2(0)}{\left(\beta v_1(0) t + 1\right)^2}. \tag{22}$$

This situation is depicted in Figure 1.

If $\alpha \neq 0$, (20) has two constant solutions,

$$v_1^1(t) \equiv 0 \quad \text{and} \quad v_1^1(t) \equiv -\frac{\alpha}{\beta}, \tag{23}$$

and a general solution,

$$v_1(t) = \frac{\alpha}{\beta}\frac{e^{-\alpha t}}{\left[1 + \frac{\alpha}{\beta v_1(0)}\right] - e^{-\alpha t}}, \tag{24}$$

whose monotonicity is determined by

$$v_1'(t) = -\frac{\alpha^2}{\beta}\frac{\left[1 + \frac{\alpha}{\beta v_1(0)}\right] e^{-\alpha t}}{\left\{\left[1 + \frac{\alpha}{\beta v_1(0)}\right] - e^{-\alpha t}\right\}^2}. \tag{25}$$

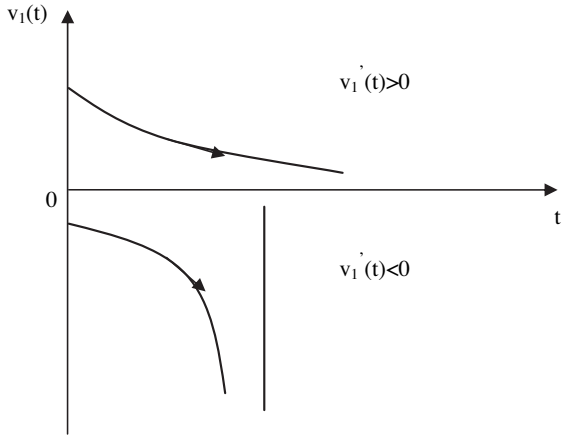


FIG. 1. $\alpha = 0$, $\beta > 0$. $v_1(t)$ *is the partial derivative of velocity.*

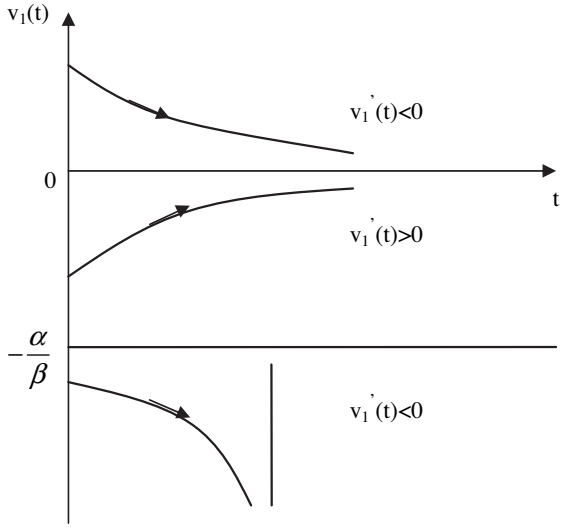


FIG. 2. $\alpha > 0$, $\beta > 0$.

This situation is depicted in Figures 2 and 3. Therefore we can judge the trend of $v_1(t)$ in the light of the parameter α and the initial equilibrium state according to Table 1. Without loss of generality, if the density disturbance is upward, the velocity perturbation is downward, i.e., $v_1(0) < 0$. If $v_1(0) \in [-\alpha/\beta, 0]$, $\alpha > 0$, i.e.,

$$V_\rho > -p_\rho. \tag{26}$$

The system is stable for traffic flow disturbances; otherwise, the system is unstable under $v_1(0) < 0$ and $\alpha < 0$, which is identical to the linear stability criterion. The linear stability result is usually only one condition obtained from the nonlinear analysis [22, 23]. Furthermore, the nonlinear stability results predict the ultimate trend of the slope of disturbance around the wavefront WF, i.e., converging to zero or diverging to negative infinity, which cannot be obtained from the linear stability analysis.

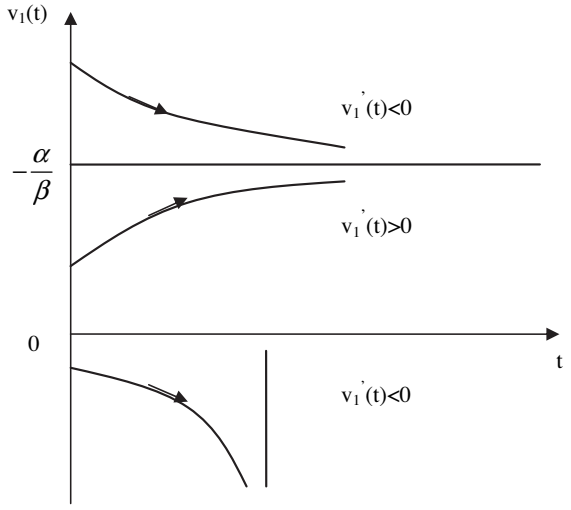


FIG. 3. $\alpha < 0$, $\beta > 0$.

TABLE 1
Stability conditions of (20).

α	Stable region	Unstable region
$\alpha > 0$	$v_1(0) \in (-\frac{\alpha}{\beta}, +\infty), v_1(t) \to 0$	$v_1(0) \in (-\infty, -\frac{\alpha}{\beta}), v_1(t) \to -\infty$
$\alpha = 0$	$v_1(0) \in (0, +\infty), v_1(t) \to 0$	$v_1(0) \in (-\infty, 0), v_1(t) \to -\infty$
$\alpha < 0$	$v_1(0) \in [0, +\infty), v_1(t) \to -\frac{\alpha}{\beta}$	$v_1(0) \in (-\infty, 0), v_1(t) \to -\infty$

Hydrodynamic models with equilibrium functions have been used to obtain reliable results which can reflect the real traffic [8, 9, 12, 15]. Although the ARA model doesn't belong to the hydrodynamic model, they are both higher-order continuum models, and we substitute the abstract form in the ARA model with the equilibrium functions given in previous hydrodynamic models, which are related to the time-independent and homogeneous traffic flow and are fitted by empirical data [4, 5, 8]. Some equilibrium functions established recently actually approach each other in the fundamental diagram [23, 24], and we choose a representative one, i.e., the KK equilibrium function first used in the KK model [8, 9], to displace the abstract form in the ARA model:

$$V(\rho) = V_f \left[\left(1 + \exp\left\{ \frac{\frac{\rho}{\rho_j} - 0.25}{0.06} \right\} \right)^{-1} - 3.72 \times 10^{-6} \right], \tag{27}$$

where V_f is the free flow velocity and ρ_j is the jam density. The KK equilibrium function is a monotonously decreasing function with respect to ρ, reaches the minimum 0 at the jam density, and possesses a turning point which is necessary for the derivation of the modified KdV equation [28, 29]. Moreover, stability criteria dependent on the equilibrium function need to contain the unstable condition because the traffic flow is unstable in one density or velocity subinterval.

The stability distribution of continuum traffic flow models along the whole density range abides by the following rule: stable at low density $\longrightarrow$ metastable $\longrightarrow$ unstable at medium density $\longrightarrow$ metastable $\longrightarrow$ stable at higher density, and in general the

dimensionless unstable density region is about $(0.15, 0.4)$ [8, 9, 23, 24]. According to this rule, we can determine the range of the pressure parameter by (26) with the KK equilibrium function:

$$p(\rho) = \rho^\gamma, \quad 0.1 < \gamma < 1.2. \tag{28}$$

Up to now, the propagation stability for (1) and (2) can thus be analyzed in terms of the initial condition $v_1(0)$ and the parameter α, and the pressure term has been determined by the stability criterion and the KK equilibrium function. We have programmed the completed ARA model with WENO schemes and succeeded in simulating the cluster and KdV soliton (the latter will be described in detail in section 4), which shows that the ARA model possesses the same numerical-simulation capability as other continuum models and the method of determining the pressure term is reasonable.

3. KdV equation. In order to obtain the unique weak solution of the nonlinear hyperbolic equation according to the weak solution theory and smoothing numerical solution, the continuum models always contain viscous terms, e.g., the KK model and the LLK model [8, 9, 10, 11]. Therefore it is necessary especially for the derivation of the KdV equation that the right-hand side of (2) be added by a higher-order viscous term, e.g., $\mu\partial_x^2 v$, $\mu > 0$, for a stable solution [9, 15, 30, 31]. We decompose the traffic flow into a linear combination of Fourier modes, each of which grows or decays with its own growth rate [30]. Thus we write

$$\rho(x,t) = \rho_0 + \sum_k \hat{\rho}_k \exp(ikx + \sigma_k t), \tag{29}$$

$$v(x,t) = v_0 + \sum_k \hat{v}_k \exp(ikx + \sigma_k t). \tag{30}$$

Substitute (29) and (30) into (1) and (2) with the viscous term, and linearize in $\hat{\rho}_k$ and $\hat{v}_k$. We find that each linear growth rate σ_k must satisfy the quadratic equation

$$0 = (ikv_0 + \sigma_k)^2 + \left(\frac{\mu}{\rho_0}k^2 - i\gamma{\rho_0}^\gamma k + \tau^{-1}\right)(ikv_0 + \sigma_k) + i\tau^{-1}V'\rho_0 k. \tag{31}$$

Both roots of (31) have negative real parts provided

$$V_\rho^0 + \gamma\rho_0^{\gamma-1} > 0, \tag{32}$$

while otherwise one root has a positive real part. (32) is an equivalent form of (26). (29)–(32) are the products of one kind of linear stability method, and the traffic flow is stable against all infinitesimal disturbances if they satisfy the linear stability condition (32). The neutral stability condition is

$$\eta \equiv V_\rho + \gamma\rho^{\gamma-1}|_{\rho=\rho_0} = 0. \tag{33}$$

The nonlinear stability analysis in section 2 is to take Taylor series expansions at the location of the wavefront, retain the constant and the first-order terms, obtain an ordinary differential equation, and finally determine the stability conditions by the convergence of solutions. However, the linear stability analysis in this section is to take Fourier series expansions with respect to the density and the velocity, linearize in the small density and velocity disturbances, obtain a quadratic equation of the growth rate

of Fourier modes, and ultimately determine the stability conditions only by judging the sign of the real parts of roots of the quadratic equation. The linear stability analysis cannot accurately distinguish each stability condition or present the stability evolution like Table 1. Moreover, there is another easier method to judge the stability according to the wave propagation rule: a higher-order partial differential equation with respect to small disturbances of the density and velocity can be obtained after the linearization of (1) and (2), the propagation speeds in the highest-order derivatives always determine the fastest and slowest signals, and the kinematic wave speed must intervene between the speeds of the fastest and slowest signals [1]. This is the so-called subcharacteristic condition, which is exactly another linear stability criterion [32]. Compared with the subcharacteristic conditions, the linear stability analysis listed in (29)–(33) also has its own advantage, i.e., some further results about the frequency and amplitude of the complex function can be worked out as follows [30].

Expanding (31) with ik near the neutral stability point yields

$$\sigma_k = -c(\rho_0)ik + \tau {\rho_0}^2 V_\rho^0 \eta k^2 + \mu\tau V_\rho^0 ik^3 - \mu\tau^2 V_\rho^0 \left(2\rho_0 V_\rho^0 + \gamma {\rho_0}^\gamma\right) k^4 + O(k^5), \tag{34}$$

where $c(\rho_0) = v_0 + \rho_0 V_\rho^0$ is the wave velocity, i.e., (6). Suppose the density of traffic flow near the neutral stability point is slightly perturbed. We quantify this supposition by writing

$$\eta = V'(\rho_0 + \delta\rho) + p'(\rho_0 + \delta\rho) = (V_{\rho\rho}^0 + p_{\rho\rho}^0)\delta\rho \equiv \theta\xi^2. \tag{35}$$

We consider the slowly varying behavior at long wavelengths near the neutral stability line. We wish to extract slow scales for space variable x and time variable t. The real part of (34) is $\tau {\rho_0}^2 V_\rho^0 \eta k^2$ and $-\mu\tau^2 V_\rho^0 \left(2\rho_0 V_\rho^0 + \gamma {\rho_0}^\gamma\right) k^4$. In order to balance the two terms, k scales as $k \propto \xi$, which leads to the scaling relation $x \propto \xi^{-1}$. The imaginary part of (34) is $-c(\rho_0)ik$ and $\mu\tau V_\rho^0 ik^3$. Since $-c(\rho_0)ik$ can be eliminated by a reference frame moving with the velocity $c(\rho_0)$, t scales as $t \propto \xi^{-3}$. Therefore we define the slow variables X and T [30, 27, 28, 29]:

$$X = \xi(x - ct) \qquad \text{and} \qquad T = \xi^3 t. \tag{36}$$

Finally we expect that an amplitude equation would balance the linear growth term of order $\xi^4 A$ with a stabilizing nonlinear term of order A^3; thus, we expect that the disturbance saturates at a size of order ξ^2. We implement the scalings by writing

$$\rho(x,t) = \rho_0 + \xi^2 \hat{\rho}(X,T), \tag{37}$$

$$v(x,t) = v_0 + \xi^2 \hat{v}(X,T). \tag{38}$$

Expanding each term in (1) and (2) added by $\mu\partial_x^2 v$ to the fifth order of ξ leads to the following nonlinear partial differential equations:

$$\xi^3 \left(-c\frac{\partial\hat{\rho}}{\partial X} + \rho_0\frac{\partial\hat{v}}{\partial X} + v_0\frac{\partial\hat{\rho}}{\partial X}\right) + \xi^5 \left(\frac{\partial\hat{\rho}}{\partial T} + \hat{\rho}\frac{\partial\hat{v}}{\partial X} + \hat{v}\frac{\partial\hat{\rho}}{\partial X}\right) = 0, \tag{39}$$

$$\begin{aligned}
&\xi^2\left(\rho_0\hat{v} - \rho_0 V_\rho^0\hat{\rho}\right) + \xi^3\left(-c\tau\rho_0\frac{\partial\hat{v}}{\partial X} + V\tau\rho_0\frac{\partial\hat{v}}{\partial X} - c\tau\gamma{\rho_0}^\gamma\frac{\partial\hat{\rho}}{\partial X} + V\tau\gamma{\rho_0}^\gamma\frac{\partial\hat{\rho}}{\partial X}\right)\\
&+\xi^4\left(\hat{\rho}\hat{v} - V_\rho^0\hat{\rho}^2 - \frac{1}{2}\rho_0 V_{\rho\rho}^0\hat{\rho}^2 - \mu\tau\frac{\partial^2\hat{v}}{\partial X^2}\right) + \xi^5\left(\tau\rho_0\frac{\partial\hat{v}}{\partial T} + \tau\gamma{\rho_0}^\gamma\frac{\partial\hat{\rho}}{\partial X} + \tau\rho_0\hat{v}\frac{\partial\hat{v}}{\partial X}\right.\\
&\left.-\tau c\hat{\rho}\frac{\partial\hat{v}}{\partial X} + \tau V^0\hat{\rho}\frac{\partial\hat{v}}{\partial X} + \tau\gamma{\rho_0}^\gamma\hat{v}\frac{\partial\hat{\rho}}{\partial X} - c\tau\gamma^2{\rho_0}^{\gamma-1}\hat{\rho}\frac{\partial\hat{\rho}}{\partial X} + V^0\tau\gamma^2{\rho_0}^{\gamma-1}\hat{\rho}\frac{\partial\hat{\rho}}{\partial X}\right) = 0.
\end{aligned} \tag{40}$$

The third-order term of ξ can be rewritten as

(41)
$$\xi^3\left(-c\tau\rho_0\frac{\partial\hat{v}}{\partial X}+V\tau\rho_0\frac{\partial\hat{v}}{\partial X}-c\tau\gamma{\rho_0}^{\gamma}\frac{\partial\hat{\rho}}{\partial X}+V\tau\gamma{\rho_0}^{\gamma}\frac{\partial\hat{\rho}}{\partial X}\right)=-\xi^5\left(\theta\tau V_\rho^0\rho_0^2\frac{\partial\hat{\rho}}{\partial X}\right).$$

From the second-order term of ξ, we obtain

$$\hat{v}=V_\rho^0\hat{\rho}+O(\xi^2). \tag{42}$$

From the fourth-order and fifth-order terms of ξ, we obtain

$$\begin{aligned}\hat{q}\equiv\hat{\rho}\hat{v}=&\left(V_\rho^0+\frac{1}{2}\rho_0V_{\rho\rho}^0\right)\hat{\rho}^2+\mu\tau\frac{\partial^2\hat{v}}{\partial X^2}+\xi\tau\rho_0\left\{\tau\mu{V_\rho^0}^2\frac{\partial^3\hat{\rho}}{\partial X^3}+\left(\theta V_\rho^0\rho_0-\gamma{\rho_0}^{\gamma-1}\right)\frac{\partial\hat{\rho}}{\partial X}\right.\\&\left.+\left[2{V_\rho^0}^2+\rho_0V_\rho^0V_{\rho\rho}^0+(\gamma-\gamma^2)V_\rho^0{\rho_0}^{\gamma-1}\right]\hat{\rho}\frac{\partial\hat{\rho}}{\partial X}\right\}.\end{aligned} \tag{43}$$

Substituting (43) into the fifth-order term of ξ in (39) leads to the KdV equation with the perturbed term:

$$\begin{aligned}&\frac{\partial\hat{\rho}}{\partial T}+\left(2V_\rho^0+\rho_0V_{\rho\rho}^0\right)\hat{\rho}\frac{\partial\hat{\rho}}{\partial X}+\tau\mu V_\rho^0\frac{\partial^3\hat{\rho}}{\partial X^3}=-\xi\tau\rho_0\frac{\partial^2}{\partial X^2}\left\{\tau\mu{V_\rho^0}^2\frac{\partial^2\hat{\rho}}{\partial X^2}\right.\\&\left.+\left(\theta V_\rho^0\rho_0-\gamma{\rho_0}^{\gamma-1}\right)\hat{\rho}+\frac{1}{2}\left[2{V_\rho^0}^2+\rho_0V_\rho^0V_{\rho\rho}^0+(\gamma-\gamma^2)V_\rho^0{\rho_0}^{\gamma-1}\right]\hat{\rho}^2\right\}.\end{aligned} \tag{44}$$

In order to derive the regularized equation, we make the following transformations:

$$\hat{\rho}=\frac{-h}{2V_\rho^0+\rho_0V_{\rho\rho}^0}\hat{\rho}',\quad X=-\sqrt{\frac{-\tau\mu V_\rho^0}{h}}X',\quad\text{and}\quad T=\sqrt{\frac{-\tau\mu V_\rho^0}{h^3}}T', \tag{45}$$

where h is a constant. With the use of (45), one obtains the regularized equation:

$$\frac{\partial\hat{\rho}'}{\partial T'}+\hat{\rho}'\frac{\partial\hat{\rho}'}{\partial X'}+\frac{\partial^3\hat{\rho}'}{\partial X'^3}=-\xi A_1\frac{\partial^2}{\partial X'^2}\left[A_2\frac{\partial^2\hat{\rho}'}{\partial X'^2}+A_3\hat{\rho}'+A_4\hat{\rho}'^2\right], \tag{46}$$

where A_1, A_2, A_3, and A_4 are constant coefficients. If one ignores the $O(\varepsilon)$ terms in (46), it is just the KdV equation with a soliton solution as the desired solution:

$$\hat{\rho}'(X',T')=A\,\mathrm{sech}^2\left[\sqrt{\frac{A}{12}}\left(X'-\frac{A}{3}T'\right)\right]. \tag{47}$$

Amplitude A of soliton solutions of the KdV equation is a free parameter. The disturbance term $O(\xi)$ of the perturbed KdV equation (46) selects a unique member of the continuous family of KdV solitons.

Next, assuming that $\hat{\rho}'(X',T')=\hat{\rho}_0'(X',T')+\xi\hat{\rho}_1'(X',T')$, we take into account the $O(\xi)$ correction. In order to determine the selected value of A for the soliton solution (47), it is necessary to satisfy the solvability condition:

$$(\hat{\rho}_0',M[\hat{\rho}_0'])\equiv\int_{-\infty}^{\infty}\hat{\rho}_0'M[\hat{\rho}_0']dX'=0, \tag{48}$$

where $M[\hat{\rho}_0']$ is the $O(\xi)$ term of (46).

By performing the integration in the solvability condition (48), one obtains the selected value

$$A = -\frac{7\theta\rho_0\left(2V_\rho^0 + \rho_0 V_{\rho\rho}^0\right)}{4h(1-\gamma)V_\rho^0}. \tag{49}$$

Rewriting each variable to the original one leads to the soliton solution of the density:

$$\rho = \rho_0 + \frac{7\eta\rho_0}{4(1-\gamma)V_\rho^0}\operatorname{sech}^2\left\{\sqrt{\frac{7\eta\rho_0\left(2V_\rho^0+\rho_0V_{\rho\rho}^0\right)}{48\tau\mu(1-\gamma){V_\rho^0}^2}}\left[x - ct + \frac{7\eta\rho_0\left(2V_\rho^0+\rho_0V_{\rho\rho}^0\right)}{12(1-\gamma)V_\rho^0}t\right]\right\}. \tag{50}$$

4. Numerical schemes. In this section, we will use the KdV soliton simulation to examine the analytical solution (50) and also to demonstrate that the ARA model can be used to reproduce the nonlinear traffic behaviors in real traffic. Because the solitary wave is a constant-shape traveling wave solution, we use WENO schemes to conduct numerical simulation. WENO schemes are high-order accurate finite difference schemes designed for the problems with piecewise smooth solutions containing discontinuities for hyperbolic conservation laws. The key idea lies at the approximation level, where a nonlinear adaptive procedure is used to automatically choose the locally smoothest stencil, hence avoiding crossing discontinuities in the interpolation procedure as much as possible. WENO schemes have been quite successful in applications, especially for problems containing both shocks and complicated smooth solution structures [33, 34, 35]. Because (3) with the viscous term is in the hyperbolic conservative form, λ_1 admits either shock waves or rarefactions, and $v_1(t)$ in (21) may be divergent, WENO schemes fit for the simulation of a unique smooth solution of nonlinear hyperbolic equations. Therefore, WENO schemes developed in [33, 34, 35] will be used in the following.

Let $\hat{f}$ be the numerical flux function corresponding to the flux f of (3) with the viscous term. Then, a standard conservative scheme of (3) with the viscous term reads as follows:

$$\frac{dY_i}{dt} + \frac{1}{\triangle x}(\hat{f}_{i+1/2} - \hat{f}_{i-1/2}) = g(Y_i). \tag{51}$$

In the following, the numerical flux $\hat{f}_{i+1/2}$ is reconstructed by the WENO method through the Lax–Friedrichs flux splitting. The third-order accurate WENO finite difference scheme applies the cell point values $\{Y_j\}_{j=i-1}^{i+1}$ to reconstruct $Y_{i+1/2}^-$, which is the cell boundary value of $x_{i+1/2}$ on the left-hand side. With $\{Y_j\}_{j=i}^{i+2}$, $Y_{i+1/2}^+$ is similarly constructed and is the cell boundary value of $x_{i+1/2}$ on the right-hand side. Thus, we use the Lax–Friedrichs numerical flux as follow:

$$\hat{f}_{i+1/2} = \frac{1}{2}\left[f\left(Y_{i+1/2}^-\right) + f\left(Y_{i+1/2}^+\right) - \alpha\left(Y_{i+1/2}^+ - Y_{i+1/2}^-\right)\right]. \tag{52}$$

(51) and (52) constitute a complete semidiscretized scheme. We apply the third-order accurate TVD Runge–Kutta time discretization, for which the semidiscrete scheme (51) is written as the ODEs: $Y_t = L(Y)$.

The equilibrium function in (1) and (2) with the viscous term is displaced by (27), and the dimensionless unstable region of the traffic system (3) with the viscous term, $(0.105, 0.414)$, can be obtained from (26) or (32). We design a density disturbance in

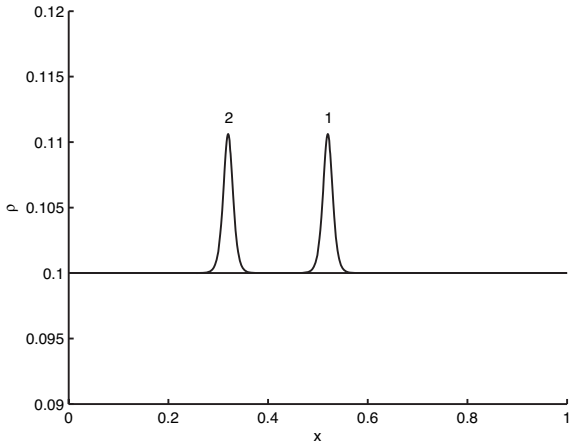


FIG. 4. *The analytical solution with* $\mu = 0.05$, (50).

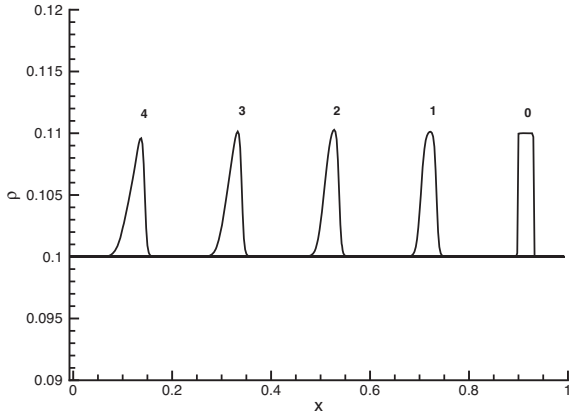


FIG. 5. *The numerical density solution in* 0, $0\triangle t$; 1, $3000\triangle t$; 2, $6000\triangle t$; 3, $9000\triangle t$; *and* 4, $12000\triangle t$.

the steady state as an initial condition:

$$\rho(x,0) = \begin{cases} \rho_0 + \hat{\rho} & x_0 - l \le x \le x_0 + l, \\ \rho_0 & \text{otherwise}, \end{cases} \tag{53}$$

where $\hat{\rho}$ is a small perturbation. The initial velocity is given by (27). The basic parameters are the pressure parameter $\gamma = 0.8$, the free velocity $v_f = 30$ m/s, the road length $L = 15000$ m, the relaxation time $\tau = 12$ s, the space interval $\triangle x = 37.5$ m, the perturbed radius $l = 0.03L$, and the dimensionless viscous coefficient $\mu = 0.05$ (see [10, 11]). The time interval $\triangle t$ must be less than 0.084 s for the numerical convergence, and we choose $\triangle t = 0.042$ s [11].

We have indeed obtained the numerical and analytical solutions of the KdV soliton near the neutral stability line as shown in Figures 4–6. In Figure 5, curve-1 is still affected by the initial condition (53) in $3000\triangle t$, and curve-2 in $6000\triangle t$ and curve-3 in $9000\triangle t$ are basically consistent with the analytic results in Figure 4. The amplitude, wave propagation velocity, and the basic shape of the numerical results are close to

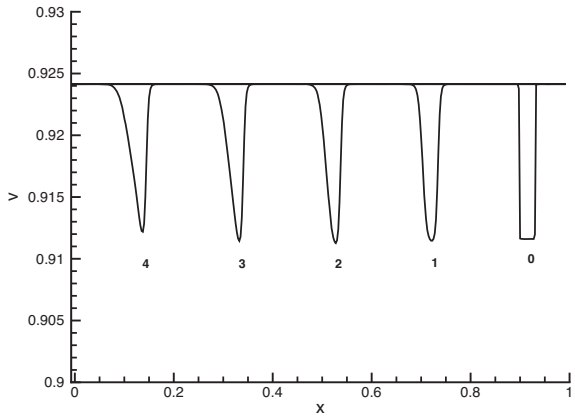


FIG. 6. *The numerical velocity solution in* 0, $0\triangle t$; 1, $3000\triangle t$; 2, $6000\triangle t$; 3, $9000\triangle t$; *and* 4, $12000\triangle t$.

those of the analytical solution, but the former are smoother than the latter, which may be caused by the numerical viscosity or the omission of the terms of the sixth and higher orders in the asymptotic expansion. Therefore the reduction perturbation method is superior to the semianalytical explanation in [31]. However, curve-4 in $12000\triangle t$ in Figure 5 shows that the dissipation term may decrease the amplitude, broaden the distribution, and contribute to an asymmetric effect. The propagation situation of the velocity is shown in Figure 6: the perturbed velocity profile is nearly symmetric to the perturbed density profile along the x-axis except the amplitude, which has examined our viewpoints in section 2. The consistency between the analytical and numerical results demonstrates that the reduction perturbation method in the continuum model originated by Kurtze and Hong [30] and further developed in this paper can really give the reasonable KdV soliton solution for traffic flow.

5. Summary. Considering that the previous two-equation models mimicked the gas dynamics equations with an unrealistic dependence on the acceleration with respect to the space derivative of the traffic pressure and consequently led to nonphysical effects, Aw and Rascle proposed the ARA model by replacing the space derivative with a convective derivative and rigidly discussed the solution to the Riemann problem and the admissibility of the elementary waves by the hyperbolic conservation laws. In this paper, we have discussed the application of the ARA model to investigating the traffic flow by the nonlinear stability and wave analyses. We have taken the nonlinear stability analysis on the ARA model through the wavefront expansion method. In comparison with the linear stability analysis, the nonlinear stability analysis additionally gives the analytical solution of the slope of the wavefront, and then the evolution of disturbances with time can be illuminated by stability parameters and initial conditions. We used the stability results to determine the anticipation factor. We obtained the KdV equation and analytical soliton solution from the "viscous" ARA model near the neutral stability line by extracting slow scales for space and time variables with the reduction perturbation method. The derivation of the KdV equation in the viscous continuum ARA model is similar to that in the car-following models but is more difficult because the continuum model has two equations. We applied WENO schemes to simulating the KdV soliton, and the simulation result is consistent with the analytical solution of the soliton density wave.

Acknowledgment. The authors are deeply appreciative to the referees for the insight represented in their constructive comments and suggestions.

REFERENCES

[1] G. B. Whitham, *Linear and Nonlinear Waves*, John Wiley and Sons, Inc., New York, 1974.
[2] P. I. Richards, *Shockwaves on the highway*, Oper. Res., 4 (1956), pp. 42–51.
[3] M. J. Lighthill and G. B. Whitman, *On kinematic waves:* II. *A theory of traffic flows on long crowded roads*, in Proceedings of the Royal Society, Series A, 229 (1955), pp. 317–345.
[4] H. J. Payne, *Models of freeway traffic and control*, in Mathematical Models of Public Systems Simulation Council Proc. Ser. 28, Vol. 1, Simulation Council, New York, 1971, pp. 51–61.
[5] H. J. Payne, *FREFLO: A macroscopic simulation model of freeway traffic*, Transport. Res. Record, 772 (1979), pp. 68–75.
[6] R. D. Kühne, *Macroscopic freeway model for dense traffic—Stop-start waves and incident detection*, in Proceedings of the 9th International Symposium on Transportation and Traffic Theory, I. Volmuller and R. Hamerslag, eds., VNU Science Press, Utecht, The Netherlands, 1984, pp. 21–42.
[7] R. D. Kühne, *Freeway speed distribution and acceleration noise—Calculations from a stochastic continuum theory and comparison with measurements*, in Proceedings of the 10th International Symposium on Transportation and Traffic Theory, N. H. Garter and N. H. M. Wilson, eds., Elsevier, New York, 1987, pp. 119–137.
[8] B. S. Kerner and P. Konhäuser, *Cluster effect in initially homogeneous traffic flow*, Phys. Rev. E, 48 (1993), pp. R2335–R2338.
[9] B. S. Kerner and P. Konhäuser, *Structure and parameters of clusters in traffic flow*, Phys. Rev. E, 50 (1994), pp. 54–83.
[10] H. Y. Lee, H. W. Lee, and D. Kim, *Origin of synchronized traffic flow on highways and its dynamic phase transitions*, Phys. Rev. Lett., 81 (1998), pp. 1130–1133.
[11] H. Y. Lee, H. W. Lee, and D. Kim, *Dynamic states of a continuum traffic equation with on-ramp*, Phys. Rev. E, 59 (1999), pp. 5101–5111.
[12] D. Helbing, *Gas-kinetic derivation of Navier-Stokes-like traffic equation*, Phys. Rev. E, 53 (1996), pp. 2366–2381.
[13] D. Helbing, *Derivation and emprical validation of a refined traffic flow model*, Phys. A, 233 (1996), pp. 253–282.
[14] M. Treiber, A. Hennecke, and D. Helbing, *Derivation, properties, and simulation of a gas-kinetic-based, non-local traffic model*, Phys. Rev. E, 59 (1999), pp. 239–253.
[15] D. Helbing, *Micro- and macro-simulation of freeway traffic*, Math. Comput. Modelling, 35 (2002), pp. 517–547.
[16] C. Daganzo, *Requiem for second-order fluid approximation to traffic flow*, Trans. Res. B, 29 (1995), pp. 277–286.
[17] A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow*, SIAM J. Appl. Math., 60 (2000), pp. 916–938.
[18] M. Rascle, *An improved macroscopic model of traffic flow: Derivation and links with the Lighthill-Whitham model*, Math. Comput. Modelling, 35 (2002), pp. 581–590.
[19] A. Klar and R. Wegener, *Kinetic derivation of macroscopic anticipation models for vehicular traffic*, SIAM J. Appl. Math., 60 (2000), pp. 1749–1766.
[20] R. Jiang, Q. S. Wu, and Z. J. Zhu, *A new continuum model for traffic flow and numerical tests*, Transport. Res. B, 36 (2002), pp. 405–419.
[21] Y. Xue and S. Q. Dai, *Continuum traffic model with the consideration of two delay time scales*, Phys. Rev. E, 68 (2003), pp. 066123.
[22] J. G. Yi, H. Lin, L. Alvarez, and R. Horowitz, *Stability of macroscopic traffic flow modeling through wavefront exapansion*, Transport. Res. B, 37 (2003), pp. 661–679.
[23] Z. H. Ou, *Equlibrium functions of traffic flow*, Phys. A, 351 (2005), pp. 620–636.
[24] Z. H. Ou, S. Q. Dai, L. Y. Dong, Z. Wu, and M. D. Tao, *New equilibrium function of traffic flow*, Phys. A, 362 (2006), pp. 525–531.
[25] H. M. Zhang, *Analyses of the stability and wave properties of a new continuum traffic theory*, Transport. Res. B, 33 (1999), pp. 299–415.
[26] H. M. Zhang, *A non-equilibrium traffic model devoid of gas-like behavior*, Transport. Res. B, 36 (2002), pp. 275–290.
[27] T. Komatsu and S. Sasa, *Kink soliton characterizing traffic congestion*, Phys. Rev. E, 52 (1995), pp. 5574–5582.
[28] M. Muramatsu and T. Nagatani, *Soliton and kink jams in traffic flow with open boundaries*, Phys. Rev. E, 60 (1999), pp. 180–187.

[29] Z. H. Ou, S. Q. Dai, and L. Y. Dong, *Density waves in the full velocity difference model*, J. Phys. A, 39 (2006), pp. 1251–1263.
[30] D. A. Kurtze and D. C. Hong, *Traffic jams, granular flow, and soliton selection*, Phys. Rev. E, 52 (1995), pp. 218–221.
[31] P. Berg and A. W. Woods, *On-ramp simulations and solitary waves of a car-following model*, Phys. Rev. E, 64 (2001), pp. 035602.
[32] A. Aw, A. Klar, T. Materne, and M. Rascle, *Derivation of continuum traffic flow models from microscopic follow-the-leader models*, SIAM J. Appl. Math., 63 (2000), pp. 259–278.
[33] A. Harten, B. Engquish, S. Osher, and S. Chakravarthy, *Uniformly high order essentially non-oscillatory schemes*, III, J. Comput. Phys., 71 (1987), pp. 231–303.
[34] C.-W. Shu, *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, in Advanced Numerical Approximation of Numerical Hyperbolic Equations, B. Cockburn, C. Johnson, C.-W. Shu, and E. Tadmor, eds., Lecture Notes in Math. 1697, A. Quarteroni, ed., Springer, New York, 1998, pp. 325–432.
[35] P. Zhang, S. C. Wong, and C.-W. Shu, *A weighted essentially non-oscillatory numerical scheme for a multi-class traffic flow model on an inhomogeneous highway*, J. Comput. Phys., 212 (2006), pp. 739–756.

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 619–629

AN EXACT EQUATION FOR THE FREE SURFACE OF A FLUID IN A POROUS MEDIUM*

WILLIAM ARTILES† AND ROBERTO A. KRAENKEL†

Abstract. We study the problem of the evolution of the free surface of a fluid in a saturated porous medium, bounded from below by a flat impermeable bottom, and described by the Laplace equation with moving-boundary conditions. By making use of a convenient conformal transformation, we show that the solution to this problem is equivalent to the solution of the Laplace equation on a fixed domain, with new variable coefficients, the boundary conditions. We use a kernel of the Laplace equation which allows us to write the Dirichlet-to-Neumann operator, and in this way we are able to find an *exact* differential-integral equation for the evolution of the free surface in one space dimension. Although not amenable to direct analytical solutions, this equation turns out to allow an easy numerical implementation. We give an explicit illustrative case at the end of the article.

Key words. free surface evolution, flow in porous media, mathematical modeling, conformal transformation, Dirichlet-to-Neumann, groundwater flow

AMS subject classifications. 35S30, 35Q35, 76S05

DOI. 10.1137/050644835

1. Introduction.

In this work we shall address a conceptually simple, yet until now not fully solved, question: given a fluid totally contained in a homogeneous, saturated, porous medium, bounded from below by a flat impermeable bottom and with a free deformable surface above, write down the evolution equation for the free surface in the case where the fluid can be considered two-dimensional and unbounded in the horizontal direction.

This is a classical problem. It is mathematically expressed by the Laplace equation in two dimensions, with boundary conditions on an unknown, time-dependent, boundary. As will be clear from the equations in the next section, its solution corresponds to the determination of a Dirichlet-to-Neumann operator. The usual way to tackle it is by a perturbative approach. Small parameters are introduced, measuring the relative amplitude of the motion and the longness of the perturbation:

$$\alpha = \frac{a}{h_0}, \qquad \beta = \left(\frac{h_0}{\lambda}\right)^2, \tag{1.1}$$

where a is the amplitude of the surface displacement, h_0 is the unperturbed depth, and λ is the typical wavelength of the perturbation. When $\beta \ll 1$ we have the *Dupuit approximation*, corresponding to the physical assumption of hydrostatic motion. Its use, together with the Darcy law, leads through an asymptotic expansion to the Boussinesq equation for the total thickness of the fluid [1], which in convenient nondimensional variables (see below) reads simply

$$h_t = (h_x h)_x. \tag{1.2}$$

*Received by the editors November 10, 2005; accepted for publication (in revised form) November 28, 2006; published electronically February 23, 2007.
http://www.siam.org/journals/siap/67-3/64483.html

†Instituto de Física Teórica, Universidade Estadual Paulista – UNESP, R. Pamplona 145, 01405-900 São Paulo, Brazil (william@ift.unesp.br, kraenkel@ift.unesp.br). The research of the first author was supported by FAPESP, São Paulo. The research of the second author was partially supported by CNPq/Brazil.

Equation (1.2) has been widely studied, not only in the context of porous media dynamics, but in such areas as high temperature gas dynamics [2] or convective instabilities [3]. Although not an integrable equation, its long-time behavior is known, being dominated by the self-similar solution, for localized initial data [4]. These self-similar solutions exhibit shocks, that is, propagating regions where the first derivative is singular. One speaks of diffusive waves to characterize these solutions.

Extensions of the Boussinesq equation have been proposed, in the same perturbation theoretic spirit, by several authors, encompassing higher order expansions in the longness parameter [5, 6, 7] together with a small α [8]. Still other works introduce a new perturbative parameter, the steepness $\alpha\sqrt{\beta}$ [9]. For localized initial conditions, the above-mentioned shocks become smoothed out, and we have propagating fronts.

A great analogy exists between the present problem and the determination of the evolution of a free surface of an inviscid fluid, the water-wave problem with the same two-dimensional geometry. Again, one studies the two-dimensional Laplace equation with a free boundary, but with different boundary conditions, and the determination of the evolution of the free surface is again equivalent to the determination of a Dirichlet-to-Neumann operator. Perturbative expansions have been widely used, dating back to the nineteenth century [10]. The same parameters α and β as above come into play. Assuming $\alpha \ll 1$ and $\beta \ll 1$ with $\mathcal{O}(\alpha) \approx \mathcal{O}(\beta)$ results in the asymptotic theory named *long waves in shallow water* and described by the Boussinesq system of equations, or the Benney–Luke equations [11], or the Korteweg–de Vries equation (for waves in a given direction) and its asymptotic equivalents, like the Kaup–Boussinesq [12] or the Benjamin–Bona–Mahoney–Peregrine equations [13, 14]. Alternatively, one can also directly expand the Dirichlet-to-Neumann operator in a Fourier series (due to a result on the analyticity of the Dirichlet-to-Neumann operator given in [15]), leading to numerically efficient integration schemes [16, 17]. So-called fully dispersive waves, where no assumption on β is made, have been obtained by making use of properties of harmonic functions, which is natural in the context of the two-dimensional Laplace equation, in [18]. Extensions to waves over variable topography have been obtained in [19] by using results on an analytical representation of the Dirichlet-to-Neumann operator given in [20]. All these results explore the smallness of one or two parameters in order to obtain an approximate expression for evolution of the free surface. On a different path, an important nonperturbative result was obtained in [21], where an exact integral-differential equation for the evolution of the free surface was obtained after the introduction of convenient conformal mappings. This equation was numerically studied by an FFT pseudospectral method in [22].

In the present work we will present an exact integral-differential equation for the evolution of a free surface in a porous medium which is analogous to the results obtained in [21, 22] for the water-wave problem, although we will follow some different steps from these papers. We will take advantage of a conformal map, mapping the region filled with fluid to a straight strip, thus transforming the free surface problem to a fixed domain problem for the Laplace equation, but with transformed boundary conditions, which, however, will be explicitly solvable. Although not promptly amenable to analytical calculations, the equation will lend itself to the implementation of an efficient numerical method.

The study of free surface dynamics in a porous medium finds its main applications in the investigations of groundwater oscillations in unconfined aquifers in coastal regions. In such regions, the fluctuations of the sea surface, in the form of either surface waves or tidal oscillations, induce watertable oscillations. These oscillations, in turn, affect the environmental dynamics in the region. Many works have addressed this

question [5, 6, 7, 8, 9, 23] on the theoretical side, providing equations to be used in larger integrated models for coastal environments. In particular, we should mention the effect caused by the periodic, tidal-induced, variation of the sea level, which is to induce a watertable over height with respect to the mean sea level. Our numerical calculations at the end of this article will illustrate this point.

2. Governing equations. The formulation of the problem is standard and can be found in textbooks [1, 24]. We place ourselves in a two-dimensional plane geometry. Let us call y the vertical axis, defined by gravity's direction, and x the perpendicular, horizontal, direction. Consider a fluid filling a porous medium, lying over a flat impermeable bottom, up to a total height limited by a free surface given by a curve described by $y = h(x)$. The relevant dynamical variable is the piezometric head $\Phi(x, y, t)$, defined as

$$\Phi = \frac{P}{\gamma} + y,$$

where P is the pressure and $\gamma = \rho g$ the specific weight. We assume Darcy's law; that is, we suppose that the seepage velocity is proportional to the gradient of the piezometric head, i.e,

$$\mathbf{u} = -K\nabla\phi,$$

where K is the permeability of the medium. Darcy's law is valid for the situation we have in mind, which is the flow of water percolating in rocks and soils. Theoretically it can be obtained from Stokes flow together with asymptotic expansions in a parameter measuring the ratio of microscopic (pores) length scales to macroscopic ones. Non-Darcian effects would typically arise if the flow in pores became turbulent (e.g., in high-rate gas wells).

Supposing the validity of Darcy's law and taking the flow as incompressible, we come to our dynamical equation

$$\Phi_{xx} + \Phi_{yy} = 0, \qquad 0 < y < h(x,t), \tag{2.1}$$

with the boundary conditions at the free surface given by

$$\Phi = h - h_0 \qquad \text{at} \quad y = h(x,t), \tag{2.2}$$

$$h_t - \frac{K}{n_e}\Phi_x h_x + \frac{K}{n_e}\Phi_y = 0 \quad \text{at} \quad y = h(x,t), \tag{2.3}$$

where n_e is the effective porosity and the displacement surface $h - h_0$ is an integrable function. At the bottom, we have a Neumann condition:

$$\Phi_y = 0, \qquad y = 0. \tag{2.4}$$

The problem is posed with an initial condition for the free surface, $h(x, 0) = \varphi(x)$. The reader will appreciate here that the above equations are directly connected to the Dirichlet-to-Neumann operator. In rescaled variables, (2.3) says that the time-derivative of $h(x, t)$ is proportional to the normal derivative of $\Phi(x, y, t)$ at the surface $y = h(x, t)$. We have thus the Laplace equation with a Dirichlet condition at the free boundary, (2.2) (in terms of the unknown function $h(x, t)$), and we have to find the normal derivative of the solution at this boundary (again in terms of $h(x, t)$) to insert

it into (2.4), implying an evolution equation for the free surface, $h(x,t)$. Therefore, the solution to our problem goes through a Dirichlet-to-Neumann operator. We will, however, avoid explicitly introducing it here, for the sake of simplicity and as we would not really gain much in doing so. Let us also point out here that the main difference between the equations governing the classical water-wave problem and those that govern the porous medium problem under consideration can be seen in (2.2), which in the last case is much simpler than in the former case, where it involves time-derivatives and nonlinear terms.

Our strategy to broach the problem will be the following: (i) first introduce nondimensional variables; (ii) next, define a conformal transformation from the strip $\Re \times (0, h(x))$ to $\Re \times (0, \mu)$, where μ is a constant to be defined below; (iii) this transformation eliminates the free-boundary problem, replacing it by a Laplace equation with new boundary conditions, involving the Jacobian of the transformation; (iv) we then solve the Laplace equation, with mixed Neumann-Dirichlet conditions, in terms of the unknown function describing the free surface, resulting in an equation for this surface, in conformal coordinates; (v) once this equation is obtained, we will develop an asymptotic analysis in a parameter measuring the longness of the wave with respect to the depth and obtain classical results on the problem; (iv) we close the paper with some numerical results on the full equations for the free surface.

3. Nondimensional equations and conformal transformation. We first write (2.1)–(2.4) in a nondimensional form. To do so, we introduce the following nondimensional variables:

$$x = \lambda\, x', \qquad y = \lambda\, y', \qquad h = h_0 h'(x', t'),$$
$$\Phi = h_0 \Phi', \qquad t = \frac{n_e}{K}\frac{\lambda^2}{h_0} t',$$

where λ is the typical wavelength of the free-surface perturbation and h_0 is the undisturbed depth of the fluid. In these new variables, we come to the following system of equations:

$$\Delta\Phi = 0, \qquad 0 < y < \mu h(x,t), \tag{3.1}$$
$$\Phi = h - 1, \qquad y = \mu h(x,t), \tag{3.2}$$
$$0 = h_t - \Phi_x h_x + \frac{1}{\mu}\Phi_y, \qquad y = \mu h(x,t), \tag{3.3}$$
$$\Phi_y = 0, \qquad y = 0, \tag{3.4}$$

where all primes have been omitted for notational convenience. A dimensionless parameter appears in these equations, $\mu = h_0/\lambda$. This would be the usual perturbative parameter for long-wave asymptotics (Dupuit approximation), where $\mu \ll 1$, meaning that the wavelength is much larger than the depth. We will not make this approximation from the beginning. Instead, we will obtain an exact equation for the free surface and only then take $\mu \ll 1$ in order to rederive previously known equations.

Further, we should note that we used a nondimensional variable so as to preserve the Laplacian, a fact of which we will make good use in what follows.

The crucial step in our procedure is the introduction of a conformal mapping. Consider a strip in the w-plane, $w = \xi + i\zeta$, given by $\Re \times [0, \mu]$. A mapping of this strip to the undulated strip in the z-plane, $z = x + iy$, given by $\Re \times [0, \mu h(x(\xi, \mu), t)]$,

is defined as a harmonic function, given as the solution of the Dirichlet problem

$$y_{\xi\xi} + y_{\zeta\zeta} = 0, \qquad 0 < y < \mu h(x(\xi,\mu),t), \tag{3.5}$$

$$y(\xi,\mu) = \mu h(x(\xi,\mu),t), \tag{3.6}$$

$$y(\xi,0) = 0 \tag{3.7}$$

if we suppose that we know the function $x(\xi,\mu)$ in the time t.

Time t plays the role of a parameter in these equations: for each t we have different functions $x(\xi,\zeta)$ and $y(\xi,\zeta)$. Equations (3.5)–(3.7) can be solved explicitly. Indeed, the solution is given by the imaginary part of

$$z(w) = \frac{1}{2}\int_{-\infty}^{\infty} \tanh\left[\frac{\pi}{2\mu}(w-\xi')\right] h(x(\xi',\mu),t)d\xi'. \tag{3.8}$$

Equation (3.5) is verified trivially, as well as (3.7). To show that (3.6) is also satisfied, we first obtain explicitly the imaginary part of (3.8). This is easily done with the help of the trigonometric identity

$$\tanh\left[\frac{\pi}{2\mu}(w-\xi')\right] \frac{\sinh[\frac{\pi}{\mu}(\xi-\xi')] + i\sin[\frac{\pi}{\mu}\zeta]}{\cosh[\frac{\pi}{\mu}(\xi-\xi')] + \cos[\frac{\pi}{\mu}\zeta]}, \tag{3.9}$$

which implies that

$$y(w) = \frac{1}{2}\int_{\Re} \frac{\sin[\frac{\pi}{\mu}\zeta]\, h(\xi',t)}{\cosh[\frac{\pi}{\mu}(\xi-\xi')] + \cos[\frac{\pi}{\mu}\zeta]} d\xi'. \tag{3.10}$$

We now use the fact that the convolution between two functions is equal to the inverse Fourier transform of the product of their Fourier transform, $\mathcal{F}^{-1}[\mathcal{F}(f)\star\mathcal{F}(g)] = f\star g$, and obtain that, after some algebra,

$$\begin{aligned} y(w) &= \mu\int_{\Re} \frac{\sinh[2\pi k\zeta]}{\sinh[2\pi k\mu]}\mathcal{F}[h]e^{2\pi i\kappa\xi}d\kappa \\ &= \mu\int_{\Re} \frac{\sinh[2\pi k\zeta]}{\sinh[2\pi k\mu]}\mathcal{F}[h-1]e^{2\pi i\kappa\xi}d\kappa + \zeta. \end{aligned} \tag{3.11}$$

Evaluated at $\zeta = \mu$, (3.11) gives (3.6) immediately. Therefore, we now have transformed our physical space, moving-boundary, domain to a fixed one through a time-dependent conformal mapping which is explicitly given by either (3.10) or (3.11).

4. Transformed equations and their solution: Free-surface evolution. Although the conformal transformation introduced in the last section leaves the Laplacian in (3.1) invariant, this is not so for the boundary conditions. In the new coordinates (ξ,ζ) the system given by (3.1)–(3.4) takes, nevertheless, a simple and convenient form. If we use that, at the upper surface, $y = \mu h(x(\xi,\mu),t)$, $\partial_\zeta = x_\xi(\partial_y - \mu h_x\partial_x)$, which follows from the Cauchy–Riemann conditions, $x_\xi = y_\zeta$ and $x_\zeta = -y_\xi$ and (3.6), we come to the transformed equations

$$\Delta\Phi = 0, \qquad 0 < \zeta < \mu, \tag{4.1}$$

$$\Phi = h - 1, \qquad \zeta = \mu, \tag{4.2}$$

$$0 = h_t + \frac{\Phi_\zeta}{\mu x_\xi}, \qquad \zeta = \mu, \tag{4.3}$$

$$\Phi_\zeta = 0, \qquad \zeta = 0. \tag{4.4}$$

This is now a system of equations defined on a fixed domain, with a coordinate-dependent coefficient in (4.3). The system of equations formed by (4.1), (4.2), (4.4) may now be seen as a Laplace equation to be solved with a Neumann condition at $\zeta = 0$ and a Dirichlet condition at $\zeta = \mu$, where $h - 1$ is the prescribed boundary value of $\Phi(\xi, \zeta)$. A solution to this problem reads

$$\Phi(\xi, \zeta, t) = \int_{-\infty}^{\infty} \mathcal{F}[\Phi(\xi, \mu, t)] \frac{\cosh[2\pi\kappa\zeta]}{\cosh[2\pi\kappa\mu]} \, e^{2\pi i\kappa\xi} d\kappa, \tag{4.5}$$

where $\mathcal{F}[\Phi](\kappa, \mu, t)$ is the Fourier transform of the piezometric head Φ at the surface $\zeta = \mu$, given in terms of $h - 1$.

Equation (4.5) solves (4.1), (4.2), (4.4). We may use it to obtain a relation between Φ_ζ and Φ at the surface (having thus implicitly constructed a Dirichlet-to-Neumann operator):

$$\Phi_\zeta(\xi, \mu, t) = \int_{-\infty}^{\infty} 2\pi\kappa \tanh[2\pi\kappa\mu] \mathcal{F}[\Phi] \, e^{2\pi i\kappa\xi} d\kappa \tag{4.6}$$

$$= \int_{-\infty}^{\infty} -i \tanh[2\pi\kappa\mu] \mathcal{F}[\Phi_\xi] \, e^{2\pi i\kappa\xi} d\kappa \equiv \mathbf{T}\partial_\xi[\Phi], \tag{4.7}$$

where $\mathbf{T}[-]$ is an integral operator defined by the above equation. It will be quite useful in the numerical calculations. For the moment, it is introduced for notational convenience. Inserting the above equation into (4.3) gives

$$h_t + \frac{1}{\mu x_\xi} \mathbf{T}\partial_x[h] = 0, \tag{4.8}$$

where use was made of the fact that $h_\xi = \Phi_\xi$ at $\zeta = \mu$. Equation (4.8) gives the time evolution of the free surface in the conformal coordinates (ξ, ζ). We should, however, note that x_ξ, which can be derived from (3.8), depends also on $h(x, t)$, making (4.8) nonlinear.

We can see the system formed by (3.8) and (4.8) as determining the time evolution of the free surface exactly. Let us, however, write it in a more compact form. If we note that the real part of (3.8) may be expressed as

$$\begin{aligned} x(\xi, \zeta) &= -i\mu \int_{-\infty}^{\infty} \frac{\cosh[2\pi\kappa\zeta]}{\sinh[2\pi\kappa\mu]} \mathcal{F}[h] e^{2\pi i\kappa\xi} d\kappa \\ &= -i\mu \int_{-\infty}^{\infty} \frac{\cosh[2\pi\kappa\zeta]}{\sinh[2\pi\kappa\mu]} \mathcal{F}[h - 1] e^{2\pi i\kappa\xi} d\kappa + \xi, \end{aligned} \tag{4.9}$$

where we again used (3.9) and the properties of the Fourier transform of the convolution of two functions, we can obtain x_ξ in the limit $\zeta \to \mu$:

$$x_\xi = \mu \int_{-\infty}^{\infty} 2\pi\kappa \coth[2\pi\kappa\mu] \, \mathcal{F}[h] e^{2\pi i\kappa\xi} d\kappa \tag{4.10}$$

or

$$x_\xi = -\mu \mathbf{T}^{-1} \partial_\xi[h]. \tag{4.11}$$

The evolution equation for the free surface is thus given in the conformal coordinates by

$$h_t = \frac{1}{\mu^2} \frac{\mathbf{T}\partial_\xi[h]}{\mathbf{T}^{-1}\partial_\xi[h]} = 0. \tag{4.12}$$

The above equation displays the time evolution of the free surface elegantly, although one could object that it could be difficult to use it in actual analytical calculations. We will, therefore, go further and show two distinct developments originating from (4.12): asymptotics and numerics.

5. Long-wave asymptotics. The theory of long-wave perturbation for a fluid in a porous medium is a classical subject, which has been extensively studied in many different aspects. Here we will systematically rederive this approximation from the exact equation (4.12), or equivalently, from (4.8), (4.11). It corresponds to the limit $\mu \ll 1$.

To proceed, we note first the identities

$$\mathbf{T}\partial_\xi[h] = \int_{-\infty}^{\infty} 2\pi\kappa \tanh[2\pi\kappa\mu]\mathcal{F}[h]\, e^{2\pi i\kappa\xi} d\kappa = -\tan(\mu\partial_\xi)\partial_\xi[h], \tag{5.1}$$

$$\mathbf{T}^{-1}\partial_\xi[h] = -\int_{-\infty}^{\infty} 2\pi\kappa \coth[2\pi\kappa\mu]\, \mathcal{F}[h] e^{2\pi i\kappa\xi} d\kappa = -\cot(\mu\partial_\xi)\partial_\xi[h], \tag{5.2}$$

where the tan and cot are defined by their series. Equations (4.8), (4.11) become respectively

$$\mu h_t x_\xi - \tan(\mu\partial_\xi)\partial_\xi[h] = 0, \tag{5.3}$$

$$x_\xi = \mu \cot(\mu\partial_\xi)\partial_\xi[h]. \tag{5.4}$$

Introduce now expansions up to order μ^2 of both equations. This implies

$$0 = h_t x_\xi - h_{\xi\xi} - \frac{\mu^2}{3} h_{\xi\xi\xi\xi} + \cdots, \tag{5.5}$$

$$x_\xi = h - \frac{\mu^2}{3} h_{\xi\xi} + \cdots. \tag{5.6}$$

The derivatives h_ξ may be rewritten iteratively as terms of h_x in an asymptotic sense:

$$h_\xi = h_x x_\xi = h_x h - \frac{\mu^2}{3} h_x h_{\xi\xi} + O(\mu^4), \tag{5.7}$$

and then

$$h_{\xi\xi} = (h_x h)_\xi - \frac{\mu^2}{3}(h_x h_{\xi\xi})_\xi + O(\mu^4) \tag{5.8}$$

$$= (h_x h)_x x_\xi - \frac{\mu^2}{3}(h_x h_{\xi\xi})_x x_\xi + O(\mu^4), \tag{5.9}$$

$$h_{\xi\xi\xi\xi} = (h_{\xi\xi\xi})_x x_\xi + O(\mu^2), \tag{5.10}$$

from which we obtain, by substituting into (5.5), the following:

$$\begin{aligned} h_t &= (h_x h)_x - \frac{\mu^2}{3}(h_x h_{\xi\xi})_x + \frac{\mu^2}{3}(h_{\xi\xi\xi})_x + \cdots \\ &= (h_x h)_x + \frac{\mu^2}{3}[-h_x h_{\xi\xi} + h_{\xi\xi x} x_\xi]_x + \cdots \\ &= (h_x h)_x + \frac{\mu^2}{3}[-2h_x h_{\xi\xi} + h_x h_{\xi\xi} + h_{\xi\xi x} h]_x + \cdots \end{aligned}$$

$$= (h_x h)_x + \frac{\mu^2}{3}[-2h_x h_{\xi\xi} + (h_{\xi\xi} h)_x]_x + \cdots$$
$$= (h_x h)_x + \frac{\mu^2}{3}[-2h_x (h_x h)_x h + ((h_x h)_x h^2)_x]_x + \cdots$$
$$= (h_x h)_x + \frac{\mu^2}{3}[-2h_x h (h_x h)_x + (h_{xx} h^3)_x + (h_x^2 h^2)_x]_x + \cdots .$$

Thus, to order μ^2, we have the equation

$$h_t = (h_x h)_x + \frac{\mu^2}{3}[h_{xx} h^3]_{xx} + \cdots , \tag{5.11}$$

which had been derived in [7]. It is quite evident that we could consistently continue the expansion to any desired order. Also, one notes that the approximation of small amplitude fluctuations was not made, but could consistently be introduced, as long as we previously state the order relations between μ^2 and the order of magnitude of the amplitude fluctuations. A further point here is to again mention the analogous problem for water waves. Equation (5.11) shows us that the problem at hand is, phenomenologically speaking, intrinsically diffusive. The first term in (5.11) represents a nonlinear diffusion, as if the diffusion coefficient were proportional to h, and the next terms are higher order and nonlinear diffusion ones. Water waves offer a comparison if one exchanges diffusion for dispersion.

6. The linear problem. In the last section we saw that it is possible to obtain a perturbative expansion in the wavelength parameter μ, giving rise to a nonlinear partial differential equation, even in the lowest order. In this section we will explore another possibility, which is to leave μ free, and obtain a new expansion based on the smallness of the amplitude of the surface elevation. The corresponding lowest order equation is a linear partial integro-differential equation, whose solution we will also present.

Let us go back to (4.12), and let us write $h = 1 + \eta$, where η is the displacement of the free surface from its undisturbed position. In the case where $\eta \ll 1$, as a first approximation we can obtain a differential equation for η by noting that

$$\begin{aligned} \eta_t &= -\frac{1}{\mu}\frac{\mathbf{T}[\eta_\xi]}{1 - \mu\mathbf{T}^{-1}[\eta_\xi]} \\ &= -\frac{1}{\mu}\mathbf{T}[\eta_\xi] - \mathbf{T}[\eta_\xi]\mathbf{T}^{-1}[\eta_\xi] + \cdots , \end{aligned} \tag{6.1}$$

and thus that the lowest order linear equation reads

$$\eta_t = -\frac{1}{\mu}\mathbf{T}[\eta_\xi]. \tag{6.2}$$

In order to give the solution of (6.2) in a compact way, define the dispersion relation as

$$w_k = \frac{2\pi k}{\mu}\tanh[2\pi k\mu], \tag{6.3}$$

and the function $G(\xi, t)$ as the inverse Fourier transform of $e^{-w_k t}$,

$$G(\xi, t) = \int_{\Re} e^{-w_k t}\, e^{2\pi i k\xi} dk, \tag{6.4}$$

where G satisfies $\lim_{t\to 0} G(\xi, t) = \delta(\xi)$.

This allows us to write the solution of (6.2), for $\xi \in \Re$, as

$$\begin{aligned} \eta(\xi,t) &= \int_{\Re} e^{-w_k t}\mathcal{F}[\varphi]\, e^{2\pi i k\xi} dk \\ &= \int_{\Re} G_t(\xi-\xi')\varphi(\xi')\, d\xi', \end{aligned} \tag{6.5}$$

where $\mathcal{F}[\varphi]$ is the Fourier transform of the initial free surface position $\eta(\xi,0)=\varphi(\xi)$. In the case $\mu \to 0$ then $w_k = (2\pi k)^2$, $G_t(\xi)$ is the Gaussian exponential, and the solution for η is

$$\eta(\xi,t) = \frac{1}{2\sqrt{\pi t}} \int_{\Re} e^{-\frac{(\xi-\xi')^2}{4t}} \varphi(\xi')\, d\xi', \tag{6.6}$$

a well-known formula for the small amplitude long-wave approximation. From the last two sections, it is clear that (4.8) may be used as starting point for other perturbative expansions involving relations between two perturbative parameters. As an example, one can substitute $h = 1+\eta$ into (5.11), write $\eta = \epsilon\overline{\eta}$ with $\epsilon \ll 1$, and have a two-parameter asymptotic expansion.

7. Numerics. In this section we will briefly describe a pseudospectral numerical method used to integrate (4.12) and display an example calculation for the classical problem of the tide-induced over-height in unconfined aquifers.

A very important point is that we were able to reduce the dynamics of a bi-dimensional boundary problem with a free surface to a one-dimensional problem given by a differential-integral equation in an exact way. Although (4.12) is somewhat odd for analytical calculations, it is quite convenient for numerical implementation. We do not need tools like, for instance, boundary integral methods involving singular operators. The Fourier-like transforms that appear in the integral operators are by no means a problem, as they can be easily managed by FFTs, resulting in a method with spectral accuracy. Let us now give a definite example of implementation of the method. We shall solve (4.12) with a periodic boundary condition $h(0,t) = 1 + \alpha\sin(wt)$ at $x=0$. This simulates the effect of ocean tides in contact with groundwater in a coastal aquifer, through an idealized vertical beach. We take $h_x(L,t)=0$ when $L \gg h_0$. This last condition allows us to make a periodic extension to the interval $2L$ by introducing an adjunct forced boundary condition as

$$h(0,t) = 1+\alpha\sin(wt) \quad \text{and} \quad h(2L,t) = 1+\alpha\sin(wt)$$

and, consequently, allowing the use of Fourier transform methods. In this model the parameter α gives a measure of the nonlinearity of the problem. As initial condition we take $h(x,0)=1$.

At each time step, the periodic functions (x, x_ξ, h) are expanded as discrete Fourier series in ξ using the FFT, and the T-transform is computed in Fourier space. For example, $\mathbf{T}[h_\xi]$ of a function can be found via FFT after multiplying the Fourier coefficients of h by $2\pi\kappa\tanh[2\pi\kappa\mu]$, as it follows from (4.7). In a similar way we may compute $\mathbf{T}^{-1}[h_\xi]$. After evaluating nonlinear terms in physical space, we advance the solution of (4.12) in time with a 4th order Runge–Kutta method.

We have worked with 516, 512, or 1024 spatial points. The spatial step size is chosen in a range between 0.01 and 0.1. Usually we work with the time step $\Delta t = 0.01$. We do not need a high-pass filter.

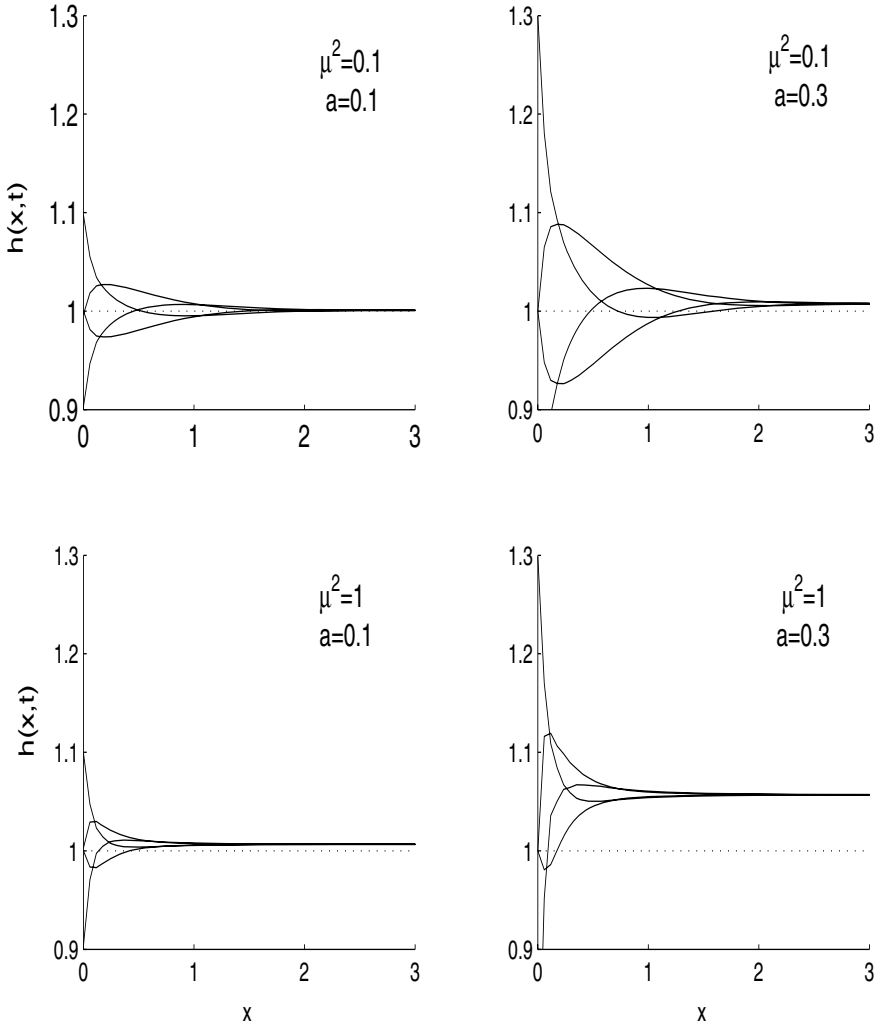


FIG. 7.1. *Different profiles for the free surface when* $t \to \infty$ *and phase* $wt = 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$.

In Figure 7.1 we plot the surface profile for distinct values of α and μ. The figure presents the numerical solution for a combination of two values of α and three values of μ when $t \to \infty$, and for four values the phase $wt = 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$.

We see the free surface displacement decaying with the distance from the boundary $x = 0$, while oscillating in both space and time. The decay rate is of the same order of magnitude of the parameter μ, and the free surface elevation for large x is of the order of α.

8. Conclusions. We have introduced a new differential-integral equation exactly describing the evolution of a free surface of a fluid totally immersed in a saturated porous medium and bounded from below by an impermeable bottom. Our equation, (4.12), is a porous-media analogue of the exact equation found for water waves in [21] and numerically studied in [22]. We have also shown that the asymptotic long-wave

expansion for this equations leads to known equations [7, 8]. The numerical implementation of (4.12) is easy, and we have provided a simple example.

The method that we have used is based on properties of harmonic functions and makes full use of a conformal transformation. Generalization to the case of an uneven bottom may follow quite easily and will be the object of future work. This should allow implementing calculations for more realistic cases of interaction of aquifers and tides through sloping beaches. A less evident generalization would be the extension of the results of this work for three-dimensional problems.

REFERENCES

[1] A. Fowler, *Mathematical Methods in the Applied Sciences*, Cambridge University Press, Cambridge, UK, 1997.

[2] Ya. B. Zeldovich e Yu. P., *Physics of Shock Waves and High Temperature Hydrodynamic Phenomena*, Vol. II, Academic Press, New York, 1967.

[3] R. A. Kraenkel, S. M. Kurcbart, J. G. Pereira, and M. A. Manna, *Nonlinear diffusion process in a Bénard system at the critical point for the onset of convection*, Phys. Rev. E, 47 (1993), pp. 3303–3306.

[4] E. S. Sabinina, *Class of nonlinear degenerate parabolic equation*, Doklady Acad. Nausk SSSR, 143 (1961), pp. 794–797.

[5] G. Dagan, *Second order theory of shallow free surface flow in porous media*, Quart. J. Mech. Appl. Math., 20 (1967), pp. 517–526.

[6] J.-Y. Parlange, F. Stagnitti, and J. L. Starr, *Free surface flow in porous media and periodic solution of the shallow flow approximation*, J. Hidrol., 70 (1984), pp. 251–263.

[7] P. Nielsen, R. Aseervatham, J. D. Fenton, and P. Perrochet, *Groundwater waves in aquifers of intermediate depths*, Adv. Water Resour., 20 (1996), pp. 37–43.

[8] P. Liu and J. Wen, *Nonlinear diffusive surface waves in porous media*, J. Fluid Mech., 347 (1997), pp. 119–139.

[9] D.-S. Jenga, B. R. Seymour, D. A. Barry, J.-Y. Parlange, D. A. Lockingtone, and L. Li, *Steepness expansion for free surface flows in coastal aquifers*, J. Hydrology, 309 (2005), pp. 85–92.

[10] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley, New York, 1974.

[11] D. J. Benney and J. C. Luke, *Interactions of permanent waves of finite amplitude*, J. Math. Phys., 43 (1964), pp. 309–313.

[12] D. J. Kaup, *Higher order water-wave equation and method for solving it*, Prog. Theoret. Phys., 54 (1975), pp. 396–400.

[13] T. B. Benjamin, J. L. Bona, and J. J. Mahony, *Model equations for long waves in nonlinear dispersive systems*, Phil. Trans. Roy. Soc., A272 (1972), pp. 47–78.

[14] D. H. Peregrine, *Long waves on a beach*, J. Fluid Mech., 27 (1967), pp. 815–827.

[15] R. R. Coifman and Y. Meyer, *Nonlinear harmonic analysis and analytic dependence*, in Pseudodifferential Operators and Applications, Proceedings of Symposia in Pure Mathematics 43, F. Treves, ed., AMS, Providence, RI, 1985, pp. 71–79.

[16] W. Craig and C. Sulem, *Numerical simulation of gravity waves*, J. Comput. Phys., 108 (1993), pp. 73–83.

[17] D. P. Nicholls and F. Reitich, *Stability of high-order perturbative methods for the computation of Dirichlet-Neumann operators*, J. Comput. Phys., 170 (2001), pp. 276–298.

[18] Y. Matsuno, *Nonlinear evolutions of surface gravity-waves on a fluid of finite depth*, Phys. Rev. Lett., 69 (1992), pp. 609–611.

[19] W. Artiles and A. Nachbin, *Nonlinear evolution of surface gravity waves over highly variable depth*, Phys. Rev. Lett., 93 (2004), paper 234501.

[20] P. Guidotti, *A first-kind boundary integral formulation for the Laplace Dirichlet-to-Neumann map in* $2D$, J. Comput. Phys., 190 (2003), pp. 325–345.

[21] A. I. Dyachenko, V. E. Zakharov, and E. A. Kuznetsov, *Nonlinear dynamics of the free surface of an ideal fluid*, Plasma Phys. Rep., 22 (1996), pp. 829–840.

[22] Ya. A. Li, J. M. Hyman, and W. Y. Choi, *A numerical study of the exact evolution equations for surface waves in water of finite depth*, Stud. Appl. Math., 113 (2004), pp. 303–324.

[23] N. Su, F. Liu, and V. Anh, *Tides as phase-modulated waves inducing periodic groundwater flow in coastal aquifers overlaying a sloping impervious base*, Environ. Model. Softw., 18 (2003), pp. 937–942.

[24] J. Bear, *Dynamics of Fluids in Porous Media*, Dover, New York, 1988.

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 630–666

ON THE FORMATION OF GLASS MICROELECTRODES*

HUAXIONG HUANG†, JONATHAN J. WYLIE‡, ROBERT M. MIURA§, AND PETER D. HOWELL¶

Abstract. Glass microelectrodes are used widely in experimental studies of the electrophysiology of biological cells and their membranes. However, the pulling of these electrodes remains an art, based on trial and error. Following Huang et al. [*SIAM J. Appl. Math.*, 63 (2003), pp. 1499–1519], we derive a one-dimensional model for the stretching of a hollow glass tube that is being radiatively heated. Our framework allows us to consider two commonly used puller designs, that is, horizontal (constant force) and vertical (variable force) pullers. We derive explicit solutions and use these solutions to identify the principal factors that control the final shape of the microelectrodes. The design implications for pullers also are discussed.

Key words. free-boundary problem, glass microelectrode, heat transfer, incompressible fluids, long-wave approximation, partial differential equations, temperature-dependent viscosity

AMS subject classifications. 76D27, 80A20, 35L60

DOI. 10.1137/050640722

1. Introduction.

Glass microelectrodes have played an essential role in cell electrophysiology for decades and will continue to be an important tool in the future. These micropipettes are used to measure membrane potentials and inject electric current and dyes into cells. This is done by inserting the electrode tips through cellular membranes or by "patching" the electrode tip to the membrane. The data collected by these techniques provide crucial information about the electrical properties of the membrane, e.g., the voltage-gated and receptor-gated ion channels, under various conditions, including during drug applications. Laboratories generally produce these microelectrodes on a daily basis using commercially available glass tubes and mechanical microelectrode pullers. For more descriptions of the medical applications of these electrodes, we refer interested readers to [16, 15], and references therein.

Electrode pullers vary in design, but all have the same basic features. They start with a uniform glass tube and heat a small section using a radiative heating element. As the glass is heated, a pulling force is applied along the axis of the tube. When the glass temperature becomes sufficiently high, its viscosity decreases dramatically, and the glass tube stretches rapidly. The tube then becomes extremely thin, ultimately breaks, and each half of the tube can be used as an electrode. Pullers are designed to

*Received by the editors September 19, 2005; accepted for publication (in revised form) October 17, 2006; published electronically February 26, 2007.
http://www.siam.org/journals/siap/67-3/64072.html

†School of Mathematics, Fudan University, Shanghai, China 200433, and Department of Mathematics and Statistics, York University, Toronto, Ontario M3J 1P3, Canada (hhuang@yorku.ca). This author's research was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada and the Mathematics of Information Technology and Complex Systems of Canada.

‡Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong (mawylie@cityu.edu.hk). This author's work was supported by a grant from the City University of Hong Kong (projects 7001560 and 7001714).

§Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102 (miura@njit.edu). This author's research was supported by the Department of Mathematical Sciences, New Jersey Institute of Technology.

¶Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Oxford OX1 3LB, UK (howell@maths.ox.ac.uk).

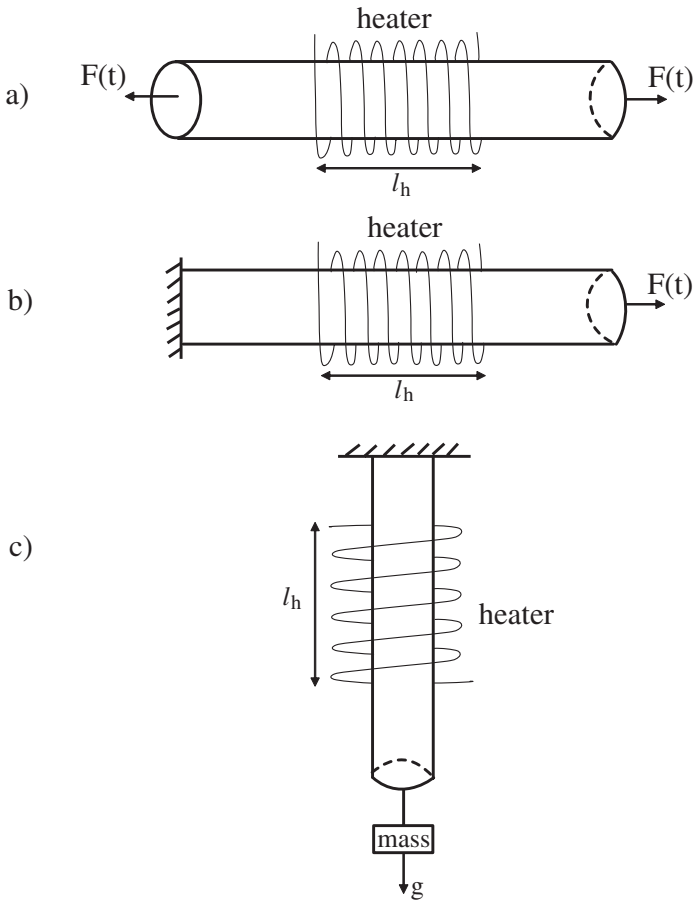


FIG. 1.1. *Three possible pullers.* (a) *Horizontal puller with free-moving ends and equal forces.* (b) *Horizontal puller with one fixed end and a single force.* (c) *Vertical puller with mass under gravity.*

achieve this result in different ways. For example, the pulling force may be achieved using electromagnets or by simply attaching a weight to one end of the tube. Electrodes may be formed in a single pull, or the glass can be stretched, allowed to cool, and subjected to a second pull that then breaks the tube. Designs for three possible pullers are shown in Figure 1.1.

Glass microelectrodes can be characterized by four experimentally relevant parameters. These are the tip length, tip diameter, electrode resistance, and electrode capacitance. Tip length is significant because it determines the physical strength of the electrode and the ease with which it can penetrate tissue and cells. A short steeply tapered tip is robust but does not penetrate tissue easily. The converse is true of long gently tapered tips. The tip shape also affects its electrical resistance and capacitance. More information on the relevance of the tip shape to physical parameters can be found in [8] and [18].

During the actual manufacturing process, the exact relationship between the variables (heater geometry, rate of pulling the glass tube, length of first pull (for a patch electrode), rate of the second pull, etc.) and electrode properties is usually determined empirically by trial and error. Although some work has been done towards

understanding how the final electrode shape is influenced by the heater geometry and width [8], in general the process is not well determined. In a previous paper [12], we developed a basic mathematical model for the formation process of these glass microelectrodes and, through computer simulations, showed that the model was capable of predicting, relatively closely, the breakup process observed in the laboratory.

The results in [12] illustrate several features that are fundamental in understanding the pulling process. First, the glass tube initially stretches very little due to the large value of the viscosity at room temperature. As the portion of the tube that is being heated by the heating element increases in temperature, the viscosity of the glass tube decreases dramatically, and the extension and breakup of the glass tube occur very rapidly. During this time, the temperature of the glass tube locally remains approximately constant; i.e., the effects of thermal diffusion and radiation are negligible.

In this paper, we derive a simplified model that captures the principal physics underlying the electrode formation. Using a dimensional analysis argument, we show that the conductive heat transfer is small compared to the radiative and convective heat transfer, and therefore conduction can be neglected in the temperature equation. We develop a general method to solve the model equations and, in some special cases, compute explicit solutions to the time-dependent equations. We carefully investigate the effects of the parameters on the final shape of the microelectrodes. Our results are relevant to existing pullers and have important implications for the future design of devices to fabricate microelectrodes.

There are certain similarities between the pulling of the glass microelectrodes and the drawing of optical and polymer fibers, which have been studied extensively in the literature [3, 4, 5, 6, 7]. For example, the governing equations for the extension of the tube or fiber can be obtained by taking the long-wave limit of the Navier–Stokes equation for incompressible fluids. However, most of the fiber drawing literature focuses on the steady state solution and its stability (drawing resonance) under isothermal conditions; cf. [6, 7, 3] and references therein. Nonisothermal cases also have been considered [11, 10, 9], but the focus is on the effect of the temperature variation on the drawing resonance. On the other hand, in the case of making glass microelectrodes, the problem is inherently transient. Another distinctive feature in the electrode pulling case is that the puller normally imposes a constant or variable force, instead of a fixed drawing speed.

Finally, the pulling of microelectrodes also has similar characteristics to extensional flow and break-off of viscous drops, which has been studied extensively; cf. [19] and references therein. However, these studies generally assume that the viscosity of the fluid remains constant, whereas in electrode production temperature-induced viscosity variations are critical.

The rest of the paper is organized as follows. In section 2, we state the basic assumptions and give the model equations. Details of the control-volume approach used in this derivation are similar to those used in [12] and [6] and therefore will be omitted. However, some issues, which have not previously been considered, such as the effects of surface tension on the inner and outer radii during stretching, will be carefully addressed on the basis of dimensional analysis. A general methodology for solving the model equations based on the method of characteristics is given.

In section 3, we obtain analytical solutions for a horizontal puller with fixed pulling force and uniform heating. For more general cases, approximate solutions are obtained using a simple numerical method. The role of parameter values is thoroughly investigated, and a particularly appealing and simple approximate theory is developed

to predict the radius of the tip at breaking. The consequences for puller design and other applications are discussed in section 4.

2. Model for glass microelectrode formation. The basic equations used in this paper to describe the stretching of the glass tube are similar to those derived in [12] for heated tubes and in [6] for isothermal tubes. In this section, we address the assumptions made in [12] and show that the flow equations reduce to those given in [12] when surface tension is negligible and can be further simplified using dimensional analysis. We also will show that the simplified model leads to a Lagrangian formulation, which can be used to compute the solution much more efficiently than the method used in [12].

2.1. Model formulation. We let ρ be the density of the glass and assume that variations in the density with temperature are negligible. We assume the tube is axisymmetric with length ℓ and inner and outer radii r and R, respectively. In the one-dimensional approximation, the velocity of the glass along the axis of the tube, which we denote by u, is independent of the radial position. When the pressure inside and outside the tube are equal, the momentum, mass, and energy conservation laws lead to the following equations:

$$\rho(R^2-r^2)\left(\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}-g\right)=\frac{\partial}{\partial x}\left(3\mu(R^2-r^2)\frac{\partial u}{\partial x}+\gamma(R+r)\right), \tag{2.1}$$

$$\frac{\partial R^2}{\partial t}+u\frac{\partial R^2}{\partial x}+R^2\frac{\partial u}{\partial x}=-\frac{\gamma Rr}{\mu(R-r)}, \tag{2.2}$$

$$\frac{\partial r^2}{\partial t}+u\frac{\partial r^2}{\partial x}+r^2\frac{\partial u}{\partial x}=-\frac{\gamma Rr}{\mu(R-r)}, \tag{2.3}$$

$$\rho c_p\left(\frac{\partial \theta}{\partial t}+u\frac{\partial \theta}{\partial x}\right)=E_R, \tag{2.4}$$

where g is the gravitational constant, γ is the surface tension coefficient, t is the time, x is the coordinate measured along the axis of the cylinder, μ is the viscosity of the glass, θ is the temperature, c_p and k are the specific heat and thermal conductivity of the glass, respectively, and E_R represents the transport of thermal energy to the glass tube by radiation. This thermal radiation term is given by

$$\begin{aligned}E_R=2k_B\sqrt{\frac{\pi}{s(1-\beta^2)}}\Bigg[&E_h\frac{\varepsilon_h\alpha}{1-(1-\alpha)(1-\varepsilon_h)}\left(\theta_h^4-\theta^4\right)\\&+E_b\frac{\varepsilon_b\alpha}{1-(1-\alpha)(1-\varepsilon_b)}\left(\theta_b^4-\theta^4\right)\Bigg],\end{aligned} \tag{2.5}$$

where $s=\pi(R^2-r^2)$ is the cross-sectional area, $\beta=r/R$ is the ratio of the radii, k_B is the Boltzmann constant, α is the absorptivity of the glass to radiative thermal energy, ε_h and ε_b are the emissivities of the heater and background, respectively, and $\theta_h(x,t)$ and $\theta_b(x,t)$ are the temperatures of the heater and the background, respectively. The quantities E_h and E_b are geometric factors between the heater and the glass tube and between the background and the glass tube, respectively. These geometric factors can be derived by integrating over the surface of the heater (and surrounding body) visible to the element of the glass tube. The details can be found in [12]. The heating is usually applied to a highly localized region of the tube of length $\ell_h \gg R_0$.

These equations are valid if the glass tube is long and thin with a small radius to length aspect ratio, as shown in Table 2.1. In addition, they also require that

the viscosity variation in the radial direction be small compared to that in the axial direction, which is justified since the viscosity is temperature-dependent and the radial variation of temperature is small, compared to that in the axial direction, as shown in Appendix A. The derivation of (2.1)–(2.3) can be found in [6].

The above equations are subject to the following initial and boundary conditions. Initially, we assume that the glass tube has a uniform temperature θ_0, length ℓ_0, and inner and outer radii r_0 and R_0. The tube is being pulled at one end with force $F(t)$, i.e.,

$$3\mu\pi(R^2 - r^2)\frac{\partial u}{\partial x} + \gamma\pi(R + r) = F(t) \tag{2.6}$$

at $x = \ell(t)$, which is a moving boundary with speed

$$\frac{d\ell}{dt} = v, \tag{2.7}$$

where $v = u(\ell, t)$ is the velocity of the glass at $x = \ell$. For symmetric pulling, we apply the condition of symmetry at $x = 0$. For asymmetric pulling with a fixed end, we simply have

$$u = 0 \tag{2.8}$$

at $x = 0$.

For pulling of glass electrodes, we also apply a terminal condition. If the pulling process is successful, the glass tube breaks in the location where the stress exceeds the "breaking stress." The breaking stress is a material-dependent parameter that also depends on the temperature. For example, for the glass used in this study, the breaking stress, $\mathcal{S}_b$, is given by the empirical formula [17]

$$\mathcal{S}_b = \frac{B}{\sqrt{\theta}}, \tag{2.9}$$

where B is an empirically determined constant.[1] The glass tube breaks when the stress in the tube is greater than $\mathcal{S}_b$. We note that this empirical law indicates that it becomes easier to break this type of glass as the temperature increases.

The difference between horizontal and vertical pullers appears only in the boundary conditions. For vertical pullers, one end of the tube is attached to a fixed location while a weight is attached to the other end. In this case, as the tube stretches, the

[1]The concept of breaking stress used in this paper is the same as the strength of glass, explored by Coenen [1]. From observations of the processes in a glass melt during the formation of cavities and of new surfaces, the formula

$$\sigma \approx 27 \times 10^6 \sqrt{\gamma^3/\theta}$$

was given in [17, p. 272]. Here σ is the applied stress (in N/m^2), γ is the surface tension coefficient (in N/m), and θ is the temperature measured in K. The weakening of the glass melt strength is partially caused by the decrease of elastic modulus when the temperature increases. The dominant factor, however, is due to the surface damage that occurs upon heating, which leads to spontaneous fracture. This mechanism is different from the pinch-off of viscous jets, where surface tension is the dominant factor and normally the inner radius collapses. On the other hand, the functional glass microelectrodes produced by the pullers have open annuli at the tips formed by breaking.

By neglecting the temperature-dependence of the surface tension coefficient [17] and using $\gamma = 0.33$ N/m for a soda-lime glass melt at 800° C, we arrive at (2.9), with B given in Table 2.1.

weight accelerates, and so the force experienced by the end of the tube is decreased. We denote the weight by F_0, gravity by g, the distance along the tube from the fixed end by x, and the location of the end attached to the weight by $x = \ell(t)$. We obtain an expression for the force applied to the free end of the tube, $F(t)$, given by

$$F(t) = F_0 - \frac{F_0}{g}\frac{\mathrm{d}^2\ell}{\mathrm{d}t^2}. \tag{2.10}$$

For horizontal pullers, the situation is simpler, and the force applied to the ends of the tube is simply $F(t)$, a specified function of time.

Equations (2.2) and (2.3) can be combined to give an equation for the cross-sectional area, s, which is essentially the equation of mass conservation,

$$\frac{\partial s}{\partial t} + u\frac{\partial s}{\partial x} + s\frac{\partial u}{\partial x} = 0. \tag{2.11}$$

The equations (2.2) and (2.3) also imply that the ratio of the radii, $\beta = r/R$, satisfies

$$\frac{\partial \beta}{\partial t} + u\frac{\partial \beta}{\partial x} = -\frac{\gamma(1+\beta)}{2\mu R}. \tag{2.12}$$

In the rest of the paper, we will use (2.11) and (2.12) instead of (2.2) and (2.3).

In a certain range of glass temperatures, the viscosity varies rapidly, and this plays a fundamental role in controlling the dynamics. Empirical data for soda-lime [2] shows that for temperatures below 900 K, which generally will be the case for electrode formation, the viscosity has an exponential dependence on temperature given by

$$\mu(\theta) = \mu_0 \exp\left[-\frac{(\theta - \theta_0)}{\theta_a}\right], \tag{2.13}$$

where μ_0 is the viscosity at the ambient temperature, θ_0, and θ_a is the "activation temperature change" required to change the viscosity by a factor of e^{-1}.

The parameters for the glass tube and heater are given in Tables 2.1 and 2.2, respectively. The other parameter that is relevant to the puller is the maximum length that the glass can be extended. This is generally constrained by the physical size of the device. If the device does not allow the tube to be extended sufficiently, then the tube may not break, and so no electrode will be formed. In later sections, we will carefully examine how the required amount of extension is related to the other parameters.

In the following sections, we will consider the most complicated case of the vertical puller, where the force on the glass tube is due to a weight that is accelerating under the influence of gravity. The other cases of specified time-dependent forces are simpler, and we will explain how to treat these cases in section 2.5.

TABLE 2.1
List of the physical parameters relating to the glass tube.

ρ g cm^{-3}	c_p Erg K^{-1} g^{-1}	k Erg cm^{-1} s^{-1} K^{-1}	k_B Erg cm^{-2} s^{-1} K^{-4}	ε_h	ε_b	α
2.23	7.538×10^6	1.130×10^5	5.67×10^{-5}	1	1	0.4

ℓ_0 cm	R_0 cm	r_0 cm	μ_0 g cm^{-1} s^{-1}	θ_a K	B dyn cm^{-3} K$^{1/2}$	γ g s^{-2}
7.56	8.66×10^{-2}	4.33×10^{-2}	10^9	50	5×10^{10}	300

TABLE 2.2
List of the geometrical parameters relating to the puller.

ℓ_h	F_0	$\theta_0 = \theta_b$	θ_h
cm	g cm s^{-2}	K	K
0.3	2×10^5	300	1000

2.2. Dimensional analysis. For the vertical puller, we nondimensionalize the variables using the following scales:

$$u = \frac{\ell_0 F_0}{3\mu_0 s_0} u', \quad s = s_0 s', \quad x = \ell_0 x', \quad t = \frac{3\mu_0 s_0}{F_0} t', \quad F = F_0 F',$$
$$R = R_0 R', \quad \ell = \ell_0 \ell', \quad \theta = \theta_0 + \theta_a \theta', \quad \text{and} \quad \mu(\theta) = \mu_0 \mu'(\theta'), \tag{2.14}$$

where the dimensionless variables are labeled with primes. Here ℓ_0, s_0, R_0, and r_0 are the initial length, cross-sectional area, and the outer and inner radii of the glass tube. Note that we have used

$$u_0 = \frac{\ell_0 F_0}{3\mu_0 s_0}$$

as the velocity scale, by balancing the pulling weight F_0 with the viscous force in the glass tube using the elongation viscosity. After substitution and dropping primes, the momentum equation (2.1), mass conservation equation (2.11), equation for the radii ratio (2.12), and heat equation (2.4) become

$$Re\left(\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} - \frac{1}{F_r}\right) = \frac{1}{s}\frac{\partial}{\partial x}\left(\mu s \frac{\partial u}{\partial x} + \lambda R(1+\beta)\right), \tag{2.15}$$

$$\frac{\partial s}{\partial t} + u\frac{\partial s}{\partial x} + s\frac{\partial u}{\partial x} = 0, \tag{2.16}$$

$$\frac{\partial \beta}{\partial t} + u\frac{\partial \beta}{\partial x} = -\frac{3}{2}(1-\beta_0^2)\lambda\frac{(1+\beta)}{\mu R}, \tag{2.17}$$

$$\theta_t + u\theta_x = \mathcal{H}\frac{H(x,\theta)}{s^{1/2}}\sqrt{\frac{1-\beta_0^2}{1-\beta^2}}. \tag{2.18}$$

(The derivation of (2.18) is given in Appendix A.) The dimensionless boundary conditions for the vertical puller are

$$u = 0 \ \text{ at } \ x = 0 \tag{2.19}$$

and

$$\mu s\frac{\partial u}{\partial x} + \lambda R(1+\beta) = 1 - F_r\frac{d^2\ell}{dt^2}, \qquad u = \frac{d\ell}{dt} \ \text{ at } \ x = \ell. \tag{2.20}$$

The initial conditions are

$$u = 0, \quad s = 1, \quad \theta = 0, \quad \ell = 1 \ \text{ at } \ t = 0. \tag{2.21}$$

The terminal condition is the nondimensional breaking criterion

$$\max_{0\le x\le \ell}\left\{\mu s\frac{\partial u}{\partial x} + \lambda R(1+\beta) - \frac{C_b s}{\sqrt{\theta + \frac{\theta_0}{\theta_a}}}\right\} = 0, \tag{2.22}$$

where

$$C_b = \frac{Bs_0}{F_0\sqrt{\theta_a}}.$$

The dimensionless parameters

$$Re = \frac{\rho F_0 \ell_0^2}{9\mu_0^2 s_0}, \quad F_r = \frac{u_0^2}{g\ell_0}, \quad \lambda = \frac{\pi\gamma R_0}{F_0}$$

are the ratio of inertia to viscous forces, the inertia to the gravity forces, and surface tension forces to the external pulling force, respectively. $\beta_0 = r_0/R_0$ is the initial value of β,

$$\mathcal{H} = \frac{6\mu_0\sqrt{s_0}k_B\epsilon_h\alpha\theta_h^4\sqrt{\pi}}{\rho c_p\theta_a F_0\sqrt{(1-\beta_0^2)}[1-(1-\alpha)(1-\epsilon_h)]}$$

is the dimensionless heater strength, and

(2.23)
$$H(x,\theta) = E_h(x)\left(1-\left(\frac{\theta_0+\theta\theta_a}{\theta_h}\right)^4\right) + \frac{\epsilon_b[1-(1-\alpha)(1-\epsilon_h)]E_b(x)(\theta_b^4-(\theta_0+\theta_a\theta)^4)}{\epsilon_h[1-(1-\alpha)(1-\epsilon_b)]\theta_h^4}$$

is the dimensionless radiation distribution. In the case when heater temperature is much higher than that of the glass and the background, then $H(x,\theta)$ can be approximated by a given function of x. The dimensionless heater strength, $\mathcal{H}$, can be thought of as the heat absorbed by a thread being pulled with constant force F_0 as it passes through the heater, divided by the heat required to significantly change the viscosity. Small values of $\mathcal{H}$ imply that the viscosity remains almost constant, and so the solution will be similar to the isothermal case. Large values of $\mathcal{H}$ imply that significant viscosity gradients will occur in the thread. The dimensionless radiation distribution is the normalized radiative heat flux.

We now can further simplify the governing equations based on the parameter estimates at two of the most relevant stages: the beginning of the pulling and when the glass tube breaks. Initially, the tube is cold and the viscosity is large, whereas near breaking, the tube will have absorbed a significant amount of heat, and the viscosity can be reduced by several orders of magnitude. As a consequence, dimensionless variables describing the flow can vary dramatically. Therefore, we need to consider the relative sizes of inertia, gravity, surface tension, and viscous and pulling forces at both stages.

Using the typical parameters in Tables 2.1 and 2.2, we see that dimensionless parameters

$$\lambda \approx 4\times 10^{-4}, \quad Re \approx 1.6\times 10^{-10}, \quad \text{and} \quad Re/F_r \approx 10^{-3}$$

are small, and therefore initially surface tension and inertial and gravitational forces can be ignored. Typically, $\mathcal{H} = O(200)$; therefore initially the heating of the tube dominates the advective term, and the tube temperature increases with very little motion. However, when the tube is heated, the viscosity decreases dramatically, the tube stretches quickly, and the tube diameter decreases rapidly. These large changes mean that the above dimensionless parameters may not adequately characterize the sizes of

the surface tension, inertial and gravitational forces when the tube is close to breaking. We therefore must also compare the inertial, surface tension, and gravitational forces with the size of the imposed force near breaking.

Assuming that the stress in the tube can be approximated by the pulling force divided by the cross-sectional area, the dimensionless breaking criterion for the vertical puller is given by

$$\frac{1}{s}\left(1 - F_r \frac{d^2\ell}{dt^2}\right) = \frac{C_b}{\sqrt{\theta + \frac{\theta_0}{\theta_a}}}. \tag{2.24}$$

We first must obtain order of magnitude estimates for the diameter at which the tube breaks, the highest temperature that the tube reaches, and the viscosity of the tube near breaking. The tube starts to stretch significantly when the advection and radiative heating terms are of the same order of magnitude. This means that the viscosity must drop by a factor of order $\mathcal{H} = 200$. Hence, the viscosity near breaking μ_b will be of order $1/200$, and the dimensionless temperature must rise by approximately 6 (which corresponds to a dimensional temperature change of approximately 300 K). Knowing the temperature, θ, and using the breaking stress formula (2.24), we can obtain an order of magnitude estimate for the dimensionless cross-sectional area at which breaking occurs, s_b, yielding $s_b = 10^{-2}$ (which corresponds to a dimensional cross-sectional area of order 10^{-4} cm^2).

We are now in a position to estimate the dimensionless ratios near the breaking time. Using (2.15) and (2.20), we see that the characteristic sizes of the surface tension and inertial and gravitational terms compared to the imposed force are given by

$$s_b^{1/2}\lambda = O(10^{-5}), \quad \frac{1}{\mu_b^2 s_b} Re = O(10^{-3}), \quad \text{and} \quad s_b \frac{Re}{F_r} = O(10^{-5}),$$

respectively. Therefore, we can conclude that surface tension, inertia, and gravity can be neglected during the entire pulling process. The acceleration term, $F_r\ d^2\ell/dt^2$, is negligible initially because it is $O(10^{-7})$, but near the breaking time, this term may be O(1), and so we must retain this term. Similarly, the ratio of the initial stress to the breaking stress, C_b, is $O(10^2)$, indicating that the tube is initially far from breaking, but eventually the tube will become sufficiently thin that the stress will approach the breaking stress.

We note that the above discussion considers only the case of successful pulling of microelectrodes; i.e., the breaking criterion is met during the extension of the glass tube. In practice, this is not always the case, and it is possible that the glass continues to extend without breaking. When this occurs, the glass may become very heavily extended, and the acceleration term $F_r\ d^2\ell/dt^2$ may approach unity. In this case, the inertia, surface tension, and the gravity may become important. Even though this is an interesting problem, it is not really relevant to successful production of glass microelectrodes and will not be pursued in this paper.

2.3. Eulerian formulation. If the surface tension terms are neglected in (2.17), we immediately see that β is conserved following material elements. We will make the natural assumption that the initial radii, r_0 and R_0, of the tube are uniform, and therefore, β will be a constant throughout the pulling process.

As a result of neglecting inertia, surface tension, and gravity, (2.15) becomes

$$(\mu(\theta) s u_x)_x = 0. \tag{2.25}$$

Since β is constant, the heat equation (2.18) reduces to

$$\theta_t + u\theta_x = \mathcal{H}\frac{H(x,\theta)}{s^{1/2}}. \tag{2.26}$$

Equations (2.25) and (2.26), combined with the mass equation

$$s_t + us_x + su_x = 0, \tag{2.27}$$

form a closed system. The glass tube does not break as long as the inequality

$$\frac{1}{s}\left(1 - F_r\frac{d^2\ell}{dt^2}\right) < \frac{C_b}{\sqrt{\theta + \frac{\theta_0}{\theta_a}}} \tag{2.28}$$

is valid for all x. If this criterion is violated, then the tube will break and the stretching process is terminated.

For vertical pullers, a mass is attached to one end of the glass tube. Thus, the pulling force $F(t)$ is determined by Newton's second law, in nondimensional form as

$$F(t) = 1 - F_r\frac{d^2\ell}{dt^2}. \tag{2.29}$$

The boundary condition is

$$\mu(\theta)su_x|_{x=\ell} = F(t). \tag{2.30}$$

For horizontal pullers, the pulling force $F(t)$ is externally prescribed. For a constant force puller, this is equivalent to setting $F_r = 0$ in (2.29).

The momentum equation (2.25) can be integrated as

$$\mu(\theta)su_x = 1 - F_r\frac{\mathrm{d}^2\ell}{\mathrm{d}t^2}. \tag{2.31}$$

Dividing this equation by $s\mu(\theta)$ and integrating from $x = 0$ to ℓ gives an expression for the velocity of the free end, denoted by v,

$$v(t) = u(\ell,t) = \frac{d\ell}{dt} = \left(1 - F_r\frac{\mathrm{d}v}{\mathrm{d}t}\right)\int_0^\ell \frac{\mathrm{d}\eta}{s(\eta,t)\mu(\theta(\eta,t))}. \tag{2.32}$$

This can be rewritten as a differential equation for v,

$$F_r\frac{\mathrm{d}v}{\mathrm{d}t} = 1 - v\left(\int_0^\ell \frac{\mathrm{d}\eta}{s(\eta,t)\mu(\theta(\eta,t))}\right)^{-1}. \tag{2.33}$$

Using (2.31) and (2.33), the mass equation (2.27) can be reduced to

$$s_t + us_x = -\frac{v}{\mu}\left(\int_0^\ell \frac{\mathrm{d}\eta}{s(\eta,t)\mu(\theta(\eta,t))}\right)^{-1}. \tag{2.34}$$

Finally, the temperature equation is unchanged as

$$\theta_t + u\theta_x = \mathcal{H}\frac{H(x,\theta)}{s^{1/2}}. \tag{2.35}$$

Equations (2.33)–(2.35) can be solved subject to the initial conditions

$$s = 1,\ \theta = 0,\ \ell = 1,\ u = 0,\ v = 0 \quad \text{at} \quad t = 0. \tag{2.36}$$

2.4. Lagrangian formulation. As shown in [20, 13, 19], the system of equations becomes significantly simpler if expressed in Lagrangian coordinates (ξ, τ). The relationship between the Lagrangian and Eulerian coordinates is given by $x = X(\xi, \tau)$ and $t = \tau$, and

$$x_\tau = \frac{\partial X(\xi, \tau)}{\partial \tau} = u. \tag{2.37}$$

When there is no ambiguity, we will use x as both Eulerian coordinate and the Lagrangian variable X, which is the spatial coordinate of a material point which was at the location $x = \xi$ at the initial time $\tau = 0$.

For a function $f(x, t)$ defined using Eulerian coordinates, its Lagrangian derivatives are

$$f_\tau = f_t + f_x x_\tau = f_t + u f_x, \qquad f_\xi = f_x x_\xi. \tag{2.38}$$

It follows immediately that

$$u_x = \frac{u_\xi}{x_\xi} = \frac{x_{\tau\xi}}{x_\xi}. \tag{2.39}$$

Using (2.38) and (2.39), the conservation-of-mass equation (2.16) can be rewritten as

$$s_\tau + s\frac{x_{\tau\xi}}{x_\xi} = 0 \rightarrow (s x_\xi)_\tau = 0.$$

Integrating and applying the initial conditions, $s(\xi, 0) = 1$ and $x(\xi, 0) = \xi$, gives

$$x_\xi = \frac{1}{s}. \tag{2.40}$$

Writing $\mu(\theta) = e^{-\theta}$ and using Lagrangian coordinates, (2.33)–(2.35) become

$$F_r\, v_\tau(\tau) = 1 - v(\tau)\left(\int_0^1 s(\eta, \tau)^{-2} e^{\theta(\eta,\tau)} \mathrm{d}\eta\right)^{-1}, \tag{2.41}$$

$$s_\tau(\xi, \tau) = -v(\tau) e^{\theta(\xi,\tau)}\left(\int_0^1 s(\eta, \tau)^{-2} e^{\theta(\eta,\tau)} \mathrm{d}\eta\right)^{-1}, \tag{2.42}$$

and

$$\theta_\tau(\xi, \tau) = \mathcal{H}\frac{H(x(\xi, \tau), \theta(\xi, \tau))}{s(\xi, \tau)^{1/2}}. \tag{2.43}$$

2.5. Horizontal puller with a specified time-dependent force. For the horizontal puller, we need specify only the time-dependent force, $F(t)$, applied to the ends of the tube. We use the same nondimensionalizations specified in (2.14), except that now F_0 would be the maximal value of $F(t)$ over the entire time period of application.

The dimensionless (2.22) then is replaced by the simpler expression

$$\frac{F(t)}{s} = \frac{C_b}{\sqrt{\theta + \frac{\theta_0}{\theta_a}}},$$

where on the left-hand side $F(t)$ is a dimensionless specified function of time. We note that for a horizontal puller with a constant force, the dimensionless force is given by $F(t) = 1$.

In the constant force case, setting $F_r = 0$ in (2.41) leads to an equation for the velocity at the free end, which is given by the algebraic equation

$$v(t) = \int_0^\ell \frac{\mathrm{d}\eta}{s(\eta, t)\mu(\theta(\eta, t))}. \tag{2.44}$$

In Lagrangian coordinates with $\mu(\theta) = e^{-\theta}$, the velocity at the free end is given by

$$v(\tau) = \int_0^1 s(\eta, \tau)^{-2} e^{\theta(\eta,\tau)} \mathrm{d}\eta. \tag{2.45}$$

Then (2.42) and (2.43) reduce to

$$s_\tau = -\frac{1}{\mu(\theta)} = -e^\theta \tag{2.46}$$

and

$$\theta_\tau = \frac{\mathcal{H}}{\sqrt{s}} H(x(\xi, \tau), \theta). \tag{2.47}$$

2.6. Numerical method. For a general heating profile or for time-dependent pulling forces, no explicit solutions can be obtained, and we must resort to numerical methods. For an arbitrary heating profile, $H(x, \theta)$, the equations (2.41)–(2.43) can be used as the basis for a simple numerical method. The Lagrangian description of the system allows us to implement a very simple numerical method that completely avoids the problem of numerical diffusion, which arises in finite difference methods.

In order to solve this problem, we discretize the domain in both ξ and τ. At any given time τ, we use the trapezoidal rule to compute the integral $\int_0^1 s^{-2} e^\theta \mathrm{d}\xi$. Having done this, we then integrate the system of equations given in (2.41)–(2.43) using an ODE solver, e.g., a simple Euler method. Since the heater profile is given in Eulerian coordinates, we need to know the location of the material point, which can be computed using a numerical quadrature of (2.40). We note that this method can cope easily with generalized heating profiles, asymmetrical pulling, variable pulling forces, and the inclusion of heat exchange terms, θ_a^4/θ_h^4.

3. Results. In this section, we consider a number of possible configurations for the puller and heating profiles. We begin with a symmetric puller with a heater that supplies spatially constant heating. In this case, we can derive an analytical solution of the equations, and this allows us to understand many of the important features leading to the shape formation of the electrode. We then go on to consider a symmetric puller with nonuniform heating, an asymmetric puller with nonuniform heating, and finally a vertical puller.

3.1. Symmetrical pulling with constant force and constant heating. In this section, we consider the case when the pulling force is a constant, that is, a horizontal puller. We assume that the pulling and heating are symmetric about $x = 0$, and so the velocity at $x = 0$ is zero by symmetry. Therefore, the point that is initially at the centerline, $\xi = 0$, will always correspond to the point $x = 0$. Since there is no direct coupling between the breaking criterion and the shape evolution, we

will derive the solution by initially computing the shape profile and then determining the time at which the breaking criterion is first satisfied.

In order to obtain an explicit solution, we make a physically relevant assumption about the heater profile that provides a number of important insights into the problem. We will assume that the tube is initially located in the region $-1/2 \le x \le 1/2$ and that the heater is localized to the region $-\ell_h/2 \le x \le \ell_h/2$, where $0 < \ell_h < 1$ is the length of the heater. We also assume that the geometric factor, $E_h(x)$, is a constant in the heater region and zero outside of the heater region. This is appropriate if the radius of the heater element is not much larger than the outer radius of the tube. Since the glass achieves a peak temperature that typically is significantly less than the heater temperature, we will neglect the terms that are $O(\theta_a^4/\theta_h^4)$. Hence, we approximate the heating in the heater region by a constant. After the tube absorbs a significant amount of heat, it is stretched rapidly before it breaks, and so there is very little time for cooling to occur outside of the heater region. This means that we can safely neglect the cooling terms.

Therefore, the heater profile is assumed to be piecewise constant, that is,

$$H(x) = \begin{cases} 1 & 0 \le x \le \ell_h/2, \\ 0 & x > \ell_h/2. \end{cases} \tag{3.1}$$

We now discuss the solution of (2.40), (2.46), and (2.47) when $H(x)$ is given by (3.1). For any time τ, there are three parts of the solution that need to be considered separately.

Region (i). We first consider material points that are initially in the heater region and remain in the heater region at time τ. We define $\tau_h(\xi)$ to be the solution of $x(\xi, \tau_h(\xi)) = \ell_h/2$, which represents the time that a material element that starts at $x = \xi$ exits the heater region (see Figure 3.1). When $\tau < \tau_h(\xi)$, the material element

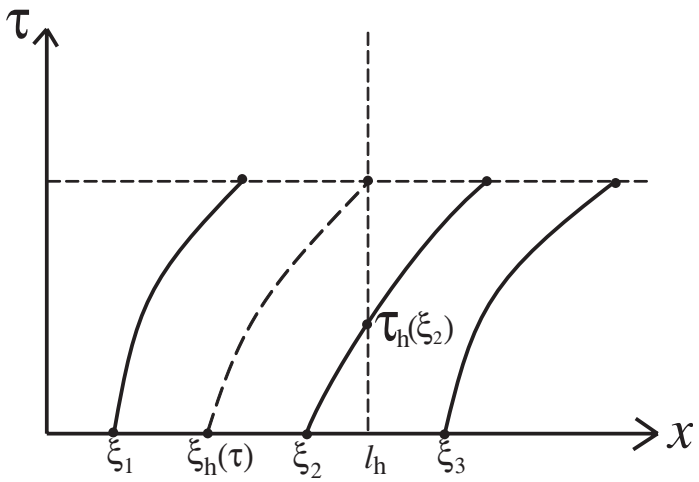


FIG. 3.1. *Symmetrical pulling with constant force and constant heating. Lagrangian trajectories.*

is subject to constant heating; thus, we need to solve the following equations:

$$s_\tau = -e^\theta, \quad \theta_\tau = \frac{\mathcal{H}}{\sqrt{s}}, \quad \text{and} \quad s x_\xi = 1, \tag{3.2}$$

subject to $s = 1$, $\theta = 0$, and $x = \xi$ at $\tau = 0$.

From (3.2), we obtain

$$\frac{\partial s}{\partial \theta} = -\frac{\sqrt{s} e^\theta}{\mathcal{H}}. \tag{3.3}$$

Integrating and applying the initial conditions yields

$$\sqrt{s} = \frac{1 + 2\mathcal{H} - e^\theta}{2\mathcal{H}}. \tag{3.4}$$

Substituting (3.4) into (3.2), we obtain

$$s_\tau = 2\mathcal{H}\sqrt{s} - (1 + 2\mathcal{H}) \quad \text{and} \quad \theta_\tau = \frac{2\mathcal{H}^2}{2\mathcal{H} + 1 - e^\theta}. \tag{3.5}$$

These two equations can be integrated again, and after applying the boundary conditions, we obtain

$$(2\mathcal{H} + 1)\theta - e^\theta + 1 = 2\mathcal{H}^2 \tau, \tag{3.6}$$

$$\sqrt{s} - 1 + \frac{2\mathcal{H} + 1}{2\mathcal{H}} \ln\left(2\mathcal{H} + 1 - 2\mathcal{H}\sqrt{s}\right) = \mathcal{H}\tau. \tag{3.7}$$

These equations can be solved explicitly in terms of the Lambert-W function that satisfies $W(x)e^{W(x)} = x$ to give

$$\theta = \frac{2\mathcal{H}^2\tau - 1}{2\mathcal{H} + 1} - W\left(-\frac{\exp\left(\frac{2\mathcal{H}^2\tau - 1}{2\mathcal{H}+1}\right)}{2\mathcal{H} + 1}\right) \tag{3.8}$$

and

$$s = \left(\frac{2\mathcal{H} + 1}{2\mathcal{H}}\right)^2 \left[1 + W\left(-\frac{\exp\left(\frac{2\mathcal{H}^2\tau - 1}{2\mathcal{H}+1}\right)}{2\mathcal{H} + 1}\right)\right]^2. \tag{3.9}$$

The Lambert function is defined only for values $x \geq -e^{-1}$. At the point $x = -e^{-1}$, its value is $W = -1$, and its gradient becomes singular. If the glass tube is allowed to extend, this singularity occurs at the finite time

$$\tau_{pinch} = \frac{(2\mathcal{H} + 1)\ln(2\mathcal{H} + 1) - 2\mathcal{H}}{2\mathcal{H}^2} \tag{3.10}$$

and corresponds to a pinch-off. However, in the case of pulling electrodes, pinch-off does not happen for the following reason. Note that at pinch-off, the cross-sectional area tends to zero, while the extension tends to infinity. In Figure 3.2, we plot the time at which pinch-off occurs as a function of the heating rate $\mathcal{H}$. We see that increasing the heating rate $\mathcal{H}$ causes the viscosity to decrease, and so the tube pinches off more quickly. In fact, this pinch-off will never occur because the stress in the tube will tend

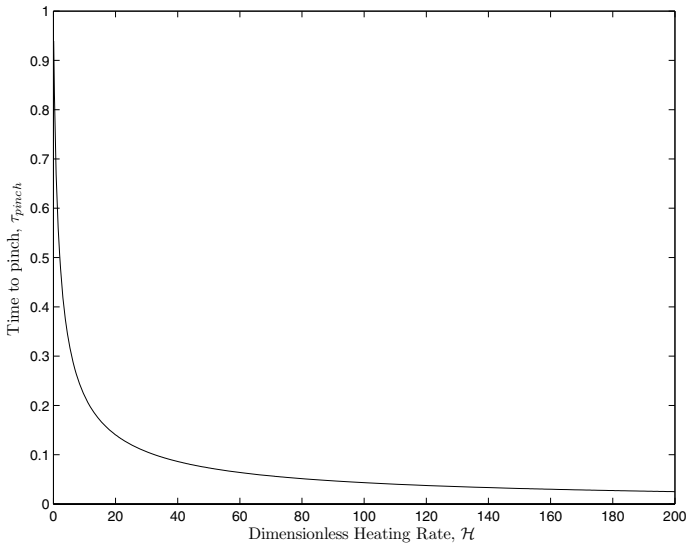


FIG. 3.2. *Dimensionless time to pinch-off as a function of the dimensionless heating rate.*

to infinity, hence exceeding the breaking stress. This means that the tube will always break before a pinch-off can occur. The time for pinch-off represents an upper bound on the duration of the pull. We will return with a more detailed discussion on the breaking of the glass tube after presenting the solutions for regions (ii) and (iii).

We note that in Lagrangian coordinates, the cross-sectional area, s, and the temperature, θ, are functions of τ only. Therefore, at time τ, we can use (3.2) to find the location of a material element that started at $x = \xi$ to obtain

$$x(\xi, \tau) = \int_0^{\xi} \frac{1}{s(\tau)} d\eta = \frac{\xi}{s(\tau)}. \tag{3.11}$$

Hence, a material element that started at $x = \xi$ will exit the heater region at a time $\tau_h(\xi)$, which can be found by solving

$$s(\tau_h(\xi)) = \frac{2\xi}{\ell_h}. \tag{3.12}$$

Substituting the above equation into (3.7), we find that

$$\tau_h(\xi) = \frac{1}{\mathcal{H}} \left[\sqrt{\frac{2\xi}{\ell_h}} - 1 + \frac{1 + 2\mathcal{H}}{2\mathcal{H}} \ln \left(2\mathcal{H} + 1 - 2\mathcal{H} \sqrt{\frac{2\xi}{\ell_h}} \right) \right]. \tag{3.13}$$

Note that points that were initially arbitrarily close to the center will exit the heater region when the time is sufficiently close to the pinching time.

In order to obtain the solution in the next region, we need the temperature of the material element that exits the tube at $\tau = \tau_h(\xi)$. This can be obtained by solving (3.4) and (3.12) to yield

$$\theta(\tau_h(\xi)) = \ln \left(2\mathcal{H} + 1 - 2\mathcal{H} \sqrt{\frac{2\xi}{\ell_h}} \right). \tag{3.14}$$

Region (ii). We now consider material points that are initially inside the heater region but exit the heater region before time τ. In other words, this region contains particles ξ for which $\tau_h(\xi) < \tau$. In this case, we can obtain the solution in a way similar to that used for region (i). Over the time interval $(0, \tau_h(\xi)]$, the solution for these particles is as given in region (i), while for the time interval $(\tau_h(\xi), \tau)$, the solution can be obtained as follows. Since the heating rate is zero, the temperature remains constant at $\theta(\tau_h(\xi))$. Therefore, the equations reduce to

$$s_\tau = -e^{\theta(\tau_h(\xi))} \quad \text{and} \quad s x_\xi = 1, \tag{3.15}$$

subject to $s = 2\xi/\ell_h$ and $x = \ell_h/2$ at $\tau = \tau_h(\xi)$. These can be solved easily to obtain

$$\theta(\xi, \tau) = \ln\left(2\mathcal{H} + 1 - 2\mathcal{H}\sqrt{\frac{2\xi}{\ell_h}}\right), \tag{3.16}$$

$$s(\xi, \tau) = \frac{2\xi}{\ell_h} - \left(2\mathcal{H} + 1 - 2\mathcal{H}\sqrt{\frac{2\xi}{\ell_h}}\right)(\tau - \tau_h(\xi)), \tag{3.17}$$

where $\tau_h(\xi)$ is given in (3.13). At time τ, we can find the location of a material element that started at $x = \xi$ using

$$x(\xi, \tau) = \frac{\ell_h}{2} + \int_{\xi_h(\tau)}^{\xi} \frac{1}{s(\eta, \tau)} d\eta, \tag{3.18}$$

where s is given in (3.17) and $\xi_h(\tau)$ is the original location of the material point that exits the heater at time τ. An explicit expression for $\xi_h(\tau)$ can be obtained by solving (3.13) to yield

$$\xi_h(\tau) = \frac{\ell_h}{2}\left(\frac{2\mathcal{H}+1}{2\mathcal{H}}\right)^2 \left(1 + W\left[-\frac{\exp\left(\frac{2\mathcal{H}^2\tau - 1}{2\mathcal{H}+1}\right)}{2\mathcal{H}+1}\right]\right)^2. \tag{3.19}$$

Region (iii). We finally consider material points that are initially outside the heater region, that is, $\ell_h/2 < \xi \le 1/2$. These points are never exposed to the heater; therefore, the temperature remains at zero. Thus, the equation for the cross-sectional area (2.42) becomes $s_\tau = -1$, which can be integrated once to yield

$$s = 1 - \tau.$$

Finally, (2.40) can be integrated to yield

$$x(\xi, \tau) = x\left(\frac{\ell_h}{2}, \tau\right) + \frac{\xi - \ell_h/2}{1 - \tau}.$$

Breaking criterion. In order to successfully make a microelectrode, the glass tube must break before the end of the tube reaches the maximum travel distance, ℓ_{max}. We begin by computing the time at which breaking occurs, τ_b. We use the solution obtained above along with the breaking criterion

$$\frac{1}{s_b}\sqrt{\theta_b + \frac{\theta_0}{\theta_a}} = C_b.$$

In the symmetric pulling case, the minimum cross-sectional area and maximum temperature occur at all the material points that have not yet exited the heater region at τ_b (region (i)). Thus, in principle, the tube can break at any of these material points. Since this region has uniform radii and the length of the region is small and of the same length as the heater, the ambiguity has little effect on the final tip radius of the electrode. For simplicity, we assume that the breaking will occur at $x = 0$, even though our model predicts that the glass tube can break anywhere in the heater region since the stress is the same. This is due to the fact that we have assumed that the heater strength is the same and the tube has uniform initial radii. In practice, of course, the middle of the heater is most likely to be the hottest; therefore, the glass will probably break in the middle. Using the solution obtained earlier, (3.4), the cross-sectional area at which breaking occurs, s_b, can be found by solving

$$\left[C_b^2 s_b^2 - \frac{\theta_0}{\theta_a}\right] = \ln\left[2\mathcal{H}(1-\sqrt{s_b})+1\right]. \tag{3.20}$$

If the tube is preheated such that $\theta_0 \gg \theta_a$, then this equation can be solved explicitly. Otherwise, it must be solved numerically, and this does not pose any serious challenges.

Having obtained s_b, we can compute the electrode profile at breaking. We first compute the time at which the tube breaks, τ_b,

$$\tau_b = \frac{1}{\mathcal{H}}\left[\sqrt{s_b} - 1 + \frac{2\mathcal{H}+1}{2\mathcal{H}}\ln\left(2\mathcal{H}+1-2\mathcal{H}\sqrt{s_b}\right)\right]. \tag{3.21}$$

Given τ_b, the initial location of the material element that is exiting the heater as the breaking occurs, $\xi_h(\tau_b)$, is obtained from (3.19). At breaking, the cross-sectional area profiles in the three different regions can be obtained using the formulas given earlier.

(i) In the heater region, $0 < \xi \le \xi_h(\tau_b)$, the cross-sectional area is independent of location and is given by $s = s_b$.

(ii) For points that were initially in the heater region but exited before breaking, $\xi_h(\tau_b) < \xi \le \ell_h/2$, the solution is given parametrically by

$$s(\xi,\tau_b) = \frac{2\xi}{\ell_h} - \left(2\mathcal{H}+1-2\mathcal{H}\sqrt{\frac{2\xi}{\ell_h}}\right)\left[\tau_b - \tau_h(\xi)\right], \tag{3.22}$$

$$x(\xi,\tau_b) = \frac{\ell_h}{2} + \int_{\xi_h(\tau_b)}^{\xi} \frac{1}{s(\eta,\tau_b)}d\eta, \tag{3.23}$$

where $\tau_h(\xi)$ is given in (3.13) and $\xi_h(\tau_b)$ is given in (3.19).

(iii) For material elements that were initially outside the heater region, $\ell_h/2 < \xi \le 1/2$, the cross-sectional area is also independent of location and is given by

$$s(\xi,\tau_b) = 1 - \tau_b.$$

The total extension of the glass tube is given by

$$x(1/2,\tau_b) = x\left(\frac{\ell_h}{2},\tau_b\right) + \frac{1-\ell_h}{2(1-\tau_b)}.$$

With the analytical expressions for the solution, we are in a position to discuss how to control the final electrode shape. First, we must ensure that the apparatus is long enough to allow sufficient extension so that breaking can occur, that is, $\ell_{max} > x(1, \tau_b)$. If breaking occurs, then each resulting electrode is composed of three parts: a region of length $\ell_h/2$ near the tip of the electrode with constant cross-sectional area s_b, a region from $x(\ell_h/2, \tau_b)$ to the end of the electrode with constant cross-sectional area $1 - \tau_b$, and a region of width $x(\ell_h/2, \tau_b) - \ell_h/2$ that connects these two regions. All of these quantities can be easily computed, and this allows one to determine puller settings to control the shape of the resulting tip.

In the following figures, we consider a heater with dimensionless length, $l_h = 0.5$, with a constant heating rate. In Figures 3.3 and 3.4, we plot the outer radius of the glass tube and the temperature as functions of the distance along the axis at various times before breaking and at the time of breaking. Initially, the glass is cool and the viscosity is relatively high, and so the glass tube deforms very little. However, after some time, the glass in the heater region becomes heated, and therefore the viscosity in this region drops. The dots in the figure show the evolution of material points that initially were spaced uniformly along the tube. From this, it is clear that the vast majority of the deformation occurs to material elements that initially were in the heater region.

We also see that as the breaking point is approached, the stretching occurs very quickly. This can be seen even more clearly in Figure 3.5, where we plot the minimum radius, which in this case occurs at the centerline, as a function of time. For the majority of the time, the material thins slowly, but once stretching begins, it occurs extremely rapidly. If we had ignored the breaking criteria, a pinching event would have occurred when the radius became zero. From this, it is clear that the breaking time can be well approximated by the time at which pinching occurs.

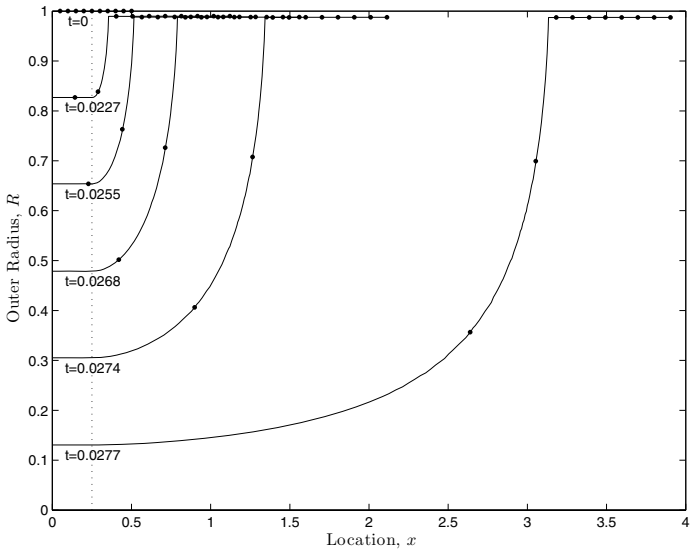


FIG. 3.3. *Outer radius of the glass tube as a function of position. The heater is located between* $x = 0$ *and* $x = 0.25$ *as indicated by the dotted line. The dots in this and subsequent figures show the evolution of material points that initially were spaced uniformly along the tube.*

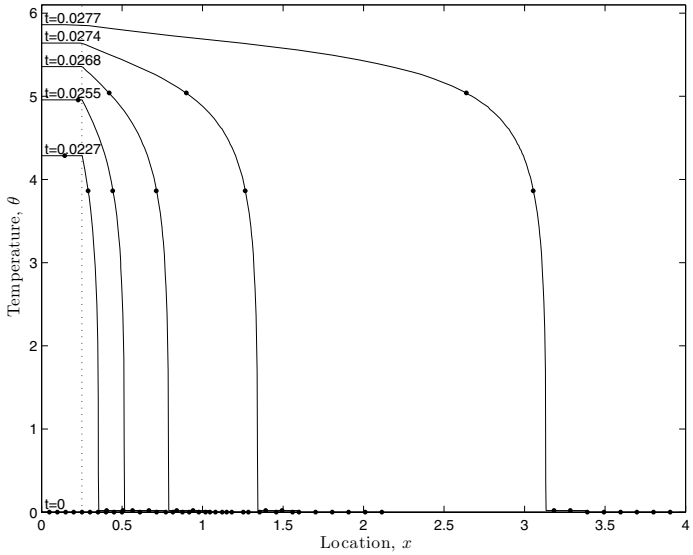


FIG. 3.4. *Temperature of the glass tube as a function of position. The heater is located between* $x = 0$ *and* $x = 0.25$ *as indicated by the dotted line.*

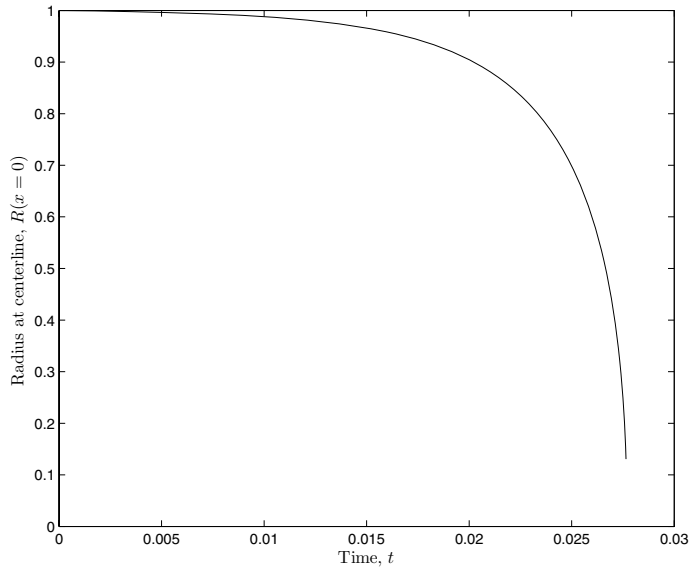


FIG. 3.5. *Minimum radius of the glass tube at the origin as a function of time.*

In Figure 3.6, we plot the dimensionless stress in the glass tube and the dimensionless breaking stress at the centerline as a function of time. As the glass tube thins, the stress increases dramatically, but breaking also is aided by heating, which acts to decrease the breaking stress. Nevertheless, during the time near breaking, the dynamics is dominated by increases in the stress.

3.2. Approximation of tip cross-sectional area. A distinct feature of the glass electrode formation process is the existence of two different regimes if the heating rate $\mathcal{H}$ is large. At early times, the glass is heated, but the viscosity is sufficiently large

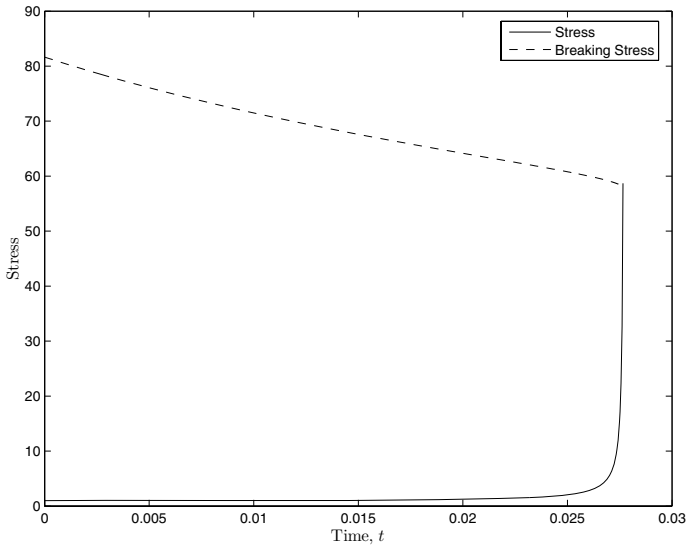


FIG. 3.6. *Dimensionless stress in the glass tube and the dimensionless breaking stress at the centerline as functions of time.*

that little stretching occurs. When the glass has absorbed sufficient thermal energy, the viscosity drops dramatically, and the glass stretches rapidly before breaking. This is largely due to two facts: the relatively large heater strength $\mathcal{H}$ and the exponential dependence of the viscosity on temperature. In the following, we will explain briefly how to use a local asymptotic analysis to obtain approximate solutions for these two regimes.

We start by examining (3.7) for the cross-sectional area s. For a more general case with nonuniform heater strength, we could apply the same local analysis to the set of governing equation (3.2). For simplicity, we will discuss only the case of constant heater strength. When $\mathcal{H} \gg 1$, we can distinguish two cases, i.e., $s \approx 1$ and $s \ll 1$.

3.2.1. Case 1: $s \approx 1$. When $s \approx 1$, or more precisely, $\sqrt{s} = 1 - o(\mathcal{H}^{-1})$, the first two terms in (3.7) essentially cancel, and the remaining two terms are in balance

$$\ln[1 + 2\mathcal{H}(1 - \sqrt{s})] = \mathcal{H}\tau,$$

which yields

$$s \approx 1 - \frac{e^{\mathcal{H}\tau} - 1}{\mathcal{H}}. \tag{3.24}$$

This approximation is valid from $\mathcal{H}\tau = 0$ up to $\mathcal{H}\tau = O(\ln 2\mathcal{H})$. Using (3.6), we find that during this regime the temperature rises from zero to $\theta \approx \ln(2\mathcal{H})$.

3.2.2. Case 2: $s \ll 1$. Near breaking, the glass tube has stretched significantly at the center and the cross-sectional area s tends to zero. Use of (3.7) yields the following approximation:

$$\sqrt{s} - 1 + \ln(1 + 2\mathcal{H}) + \ln\left(1 - \frac{2\mathcal{H}\sqrt{s}}{1 + 2\mathcal{H}}\right) = \mathcal{H}\tau.$$

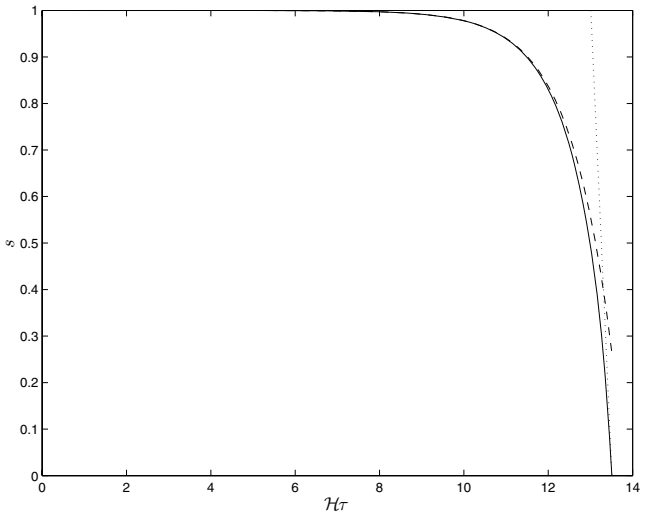


FIG. 3.7. *The cross-sectional area s is plotted as a function of the scaled time $\mathcal{H}\tau$ for a large value of $\mathcal{H} = 10^6$. The solid curve represents the exact solution, while the dashed curve represents the local asymptotic solution for $1 - s \ll 1$,* (3.24)*, and the dotted curve represents the local asymptotic solution for $s \ll 1$,* (3.25)*. The thread heats up with little thinning for a period of time $\tau = O(\mathcal{H}^{-1} \ln(2\mathcal{H}))$ and then thins over a period of time $\tau = O(\mathcal{H}^{-1})$.*

Expanding the logarithmic function in $\sqrt{s}$ yields

$$s \approx 2\ln(1 + 2\mathcal{H}) - 2(\mathcal{H}\tau + 1). \tag{3.25}$$

If the tube were not to break, it would pinch off at time

$$\tau = \frac{\ln(1 + 2\mathcal{H}) - 1}{\mathcal{H}}.$$

Hence, rapid stretching occurs within a relatively short period of time of $O(\mathcal{H}^{-1})$. This is much shorter than the initial phase, which lasted for a time on the order of $O(\mathcal{H}^{-1} \ln(2\mathcal{H}))$. Based on the approximate pinching time, temperature at pinch-off can be approximated by $\theta \approx \ln(2\mathcal{H} + 1)$. Thus the temperature variation during rapid stretching is much smaller than the $\ln(2\mathcal{H})$ variation that occurred in the initial phase.

In Figure 3.7, we have plotted the approximate solutions (3.24) and (3.25) and the exact solution (3.7). The local asymptotic solutions approximate the exact solution well in each regime.

3.2.3. Cross-sectional area at breaking. We now can find approximations that allow us to easily control the shape of the tip. The minimum area of the electrode, s_b, will be significantly smaller than the initial area of the tube; therefore, the breaking time occurs quite close to the pinch-off time. As discussed earlier, at the pinching time, the temperature is given by $\theta \approx \ln(2\mathcal{H} + 1)$. This temperature is also a good approximation for the temperature near the breaking time, because in the time between breaking and pinch-off the change in temperature is small. Hence, the breaking criterion is well approximated by

$$s_b \approx \frac{\sqrt{\ln(2\mathcal{H} + 1) + \theta_0/\theta_a}}{C_b}.$$

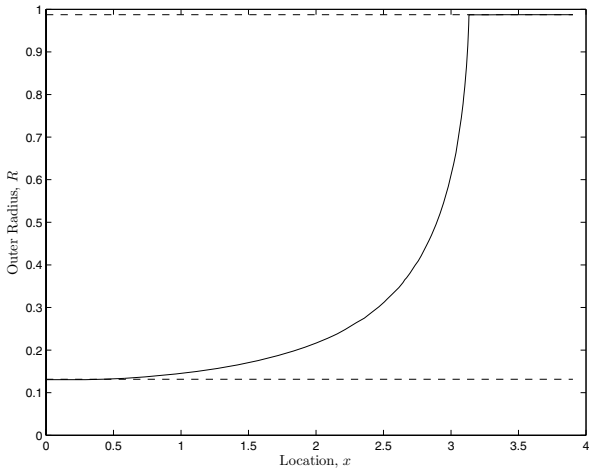


FIG. 3.8. *Microelectrode shape at the breaking time. Also included are the approximations for the tip width and the width of the glass tube far from the tip.*

Also, the area far from the tip, $1 - \tau_b$, can be approximated by $1 - \tau_{pinch}$ as the difference is $O(\mathcal{H}^{-1})$.

In Figure 3.8, we plot the final shape of the electrode along with the approximations to the tip width and the width far from the tip. They can be seen to be in excellent agreement, and this is the case over a wide range of parameters, especially for the tip width. We also have computed the difference between the two solutions, which is less than 2%.

3.3. Symmetric pulling with constant force and nonuniform heating. In this case, we also consider a tube that is initially located in the region $-1/2 \le x \le 1/2$ and a puller that is the same as in the above section except that the heating is spatially nonuniform. Rather than use a constant heating profile, we use the profile

$$H(x) = \exp(-4\pi x^2/l_h^2).$$

This has the property that the maximum heater intensity and the integrated heat intensity, $\int H(x)\mathrm{d}x$, are the same as for the piecewise constant heating used in the previous section.

In Figure 3.9, a numerical solution of the final shape of the electrode at the breaking time (solid line) is plotted along with the approximate theory for the tip radius and the radius far from the tip (horizontal dotted lines). In Figures 3.10 and 3.11, we plot the evolution of the outer radius and temperature. We see that the evolution of the glass tube profile is somewhat similar to that for the constant heating case. Of particular interest is the fact that the approximate theory for the uniform heating case still gives an extremely accurate approximation to the final tip radius. This can be understood by again dividing the dynamics into two stages: the first stage in which the glass heats up with little deformation and the second stage in which significant deformation takes place. We consider material elements near the location of the maximum in the heater intensity. In the first stage these points heat up with very little motion of the glass tube. Thus, the temperatures in the uniform and nonuniform calculations are almost identical. Then, since there is very little heating in the second stage, the breaking criteria will be achieved at approximately

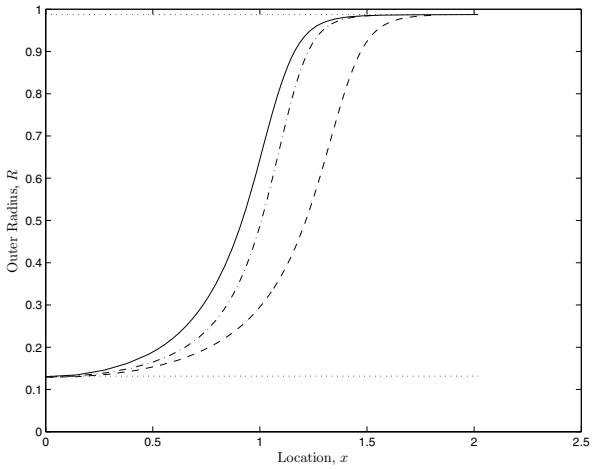


FIG. 3.9. *Final shapes at the breaking times of the microelectrodes for nonuniform heating. Profiles are given for the symmetrical pulling case (solid curve) and for the asymmetrical pulling case (dash-dot curve for the microelectrode formed from the part of the glass tube that remains attached to the fixed wall, dashed curve for the microelectrode formed by the part of the tube which is being pulled). Also, the tip radius and the radius far from the tip (horizontal dotted lines) are given by the approximate theory.*

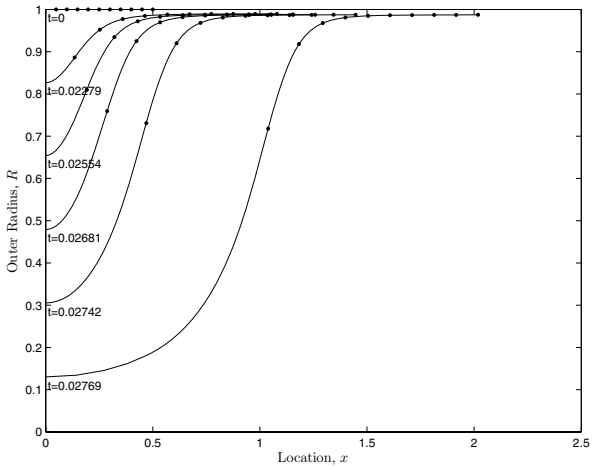


FIG. 3.10. *Outer radius of a glass tube during the evolution of the microelectrode formation under symmetric pulling with a constant force and nonuniform heating.*

the same radius in both the uniform and nonuniform cases. The only significant difference between the uniform and nonuniform cases is that uniform heating means that a larger section of the glass will be heated enough that it will significantly deform. Hence, the total extension of the glass tube before breaking is significantly longer.

3.4. Asymmetric pulling with constant force and nonuniform heating. We now consider the case of a tube that is initially located in the region $0 \leq x \leq 1$, is fixed at one end ($x = 0$), and is pulled at the other end with constant force. The tube

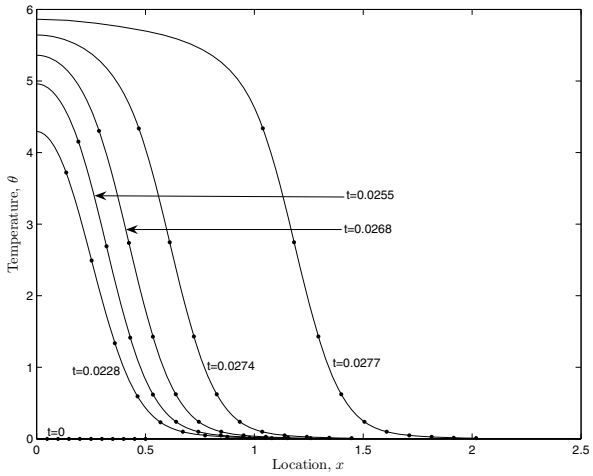


FIG. 3.11. *Temperature of a glass tube during the evolution of the microelectrode formation under symmetric pulling with a constant force and nonuniform heating.*

is heated in the same way as in the previous section, but with the maximum intensity centered on the midpoint of the tube,

$$H(x) = \exp\left(-\frac{4\pi(x-1/2)^2}{l_h^2}\right).$$

In Figure 3.9, we plotted the final shape of the electrodes at the breaking time. The dash-dot curve represents microelectrode shape of the section of the tube that remains attached to the fixed wall (reversed in x for comparison purposes), and the dashed curve represents the shape of the section to which the force was being applied. The horizontal dotted lines represent the approximate theory for the tip radius and radius far from the tip. In Figures 3.12 and 3.13, we plot the evolution of the outer radius and temperature. Again, the approximate theory gives an excellent approximation for the breaking tip radius, as explained in the above section. The difference between the electrode shapes to the left and right of the breaking point can be explained by the small differences that occur during the second stage of the evolution. Both microelectrode tips in the asymmetrical pulling case extend further than the tip for the symmetric case because the hottest part of the tube moves relative to the heater maximum. Therefore, more of the glass tube near the breaking point is significantly heated and can stretch more easily.

We note that, for spatially uniform heating, an analytical solution may be obtained using a procedure similar to that used in section 3.1. There are a number of cases to be considered, and the solutions become slightly complicated. Therefore, we present the results in Appendix B.

3.5. Variable pulling force. We now consider the case of a variable pulling force. We take a tube that is initially located in the region $0 \le x \le 1$, is fixed at one end ($x = 0$), and whose other end is attached to a mass that falls under gravity. In this case, the situation is slightly more complicated. For the vertical puller, the dimensionless stress is given by $s^{-1}(1 - F_r\ \mathrm{d}^2\ell/\mathrm{d}t^2)$. If one ignores breaking, it is easy to show that this solution will never pinch off and that $\ell \to t^2/(2F_r)$ as $t \to \infty$. This corresponds to the weight simply falling due to gravity, and the glass thread exerts a

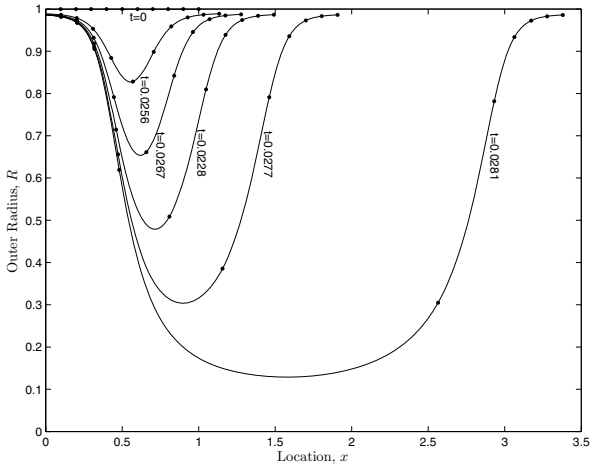


FIG. 3.12. *Outer radius of a glass tube during the evolution of the microelectrode formation under asymmetric pulling with a constant force and nonuniform heating.*

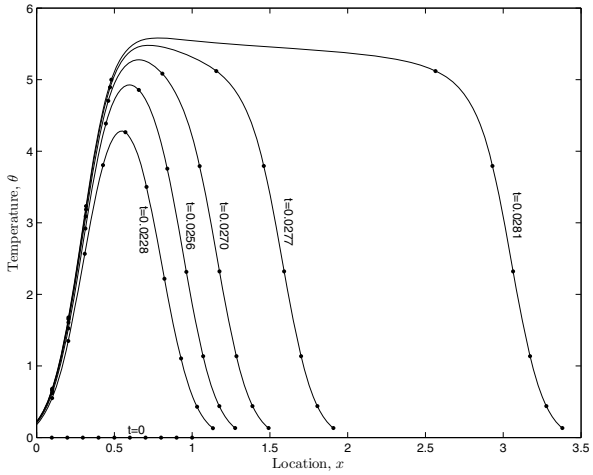


FIG. 3.13. *Temperature of a glass tube during the evolution of the microelectrode formation under asymmetric pulling with a constant force and nonuniform heating.*

negligible force on the weight. It is instructive to consider the case with no heating, i.e., the glass tube has a spatially uniform radius. In this case, it is easy to show that the stress is given by $\mathrm{d}\ln(\ell)/\mathrm{d}t$. Therefore, in the limit when $t \to \infty$, the stress tends to $2/t$, which is a decreasing function of time.

At early times, the behavior is similar to the constant force cases. This is because the initial viscosity is high and so the deformation, and hence the acceleration of the weight, are negligible. However, when the material becomes hot, it deforms rapidly. Thus, the stress increases due to thinning in the tube thickness. Therefore, the acceleration terms reduce the effective force experienced by the glass. Even though the glass is thinning, the overall stress decreases, because the force experienced by the glass is reduced by the acceleration of the weight in the same way as in the isothermal case. Therefore, we expect that the stress will attain a maximum value at finite time.

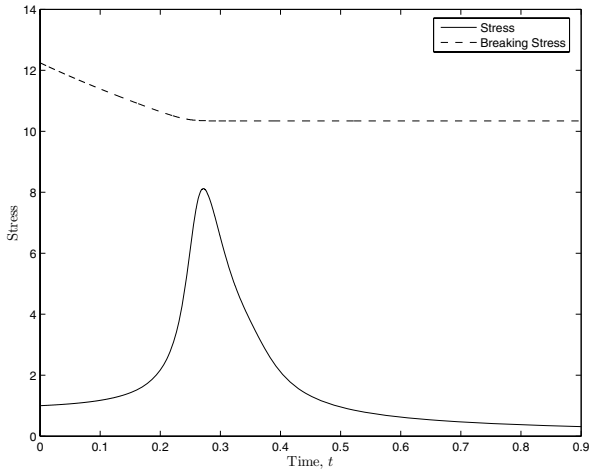


FIG. 3.14. *Dimensionless stress and dimensionless breaking stress in the glass tube closest to the location of breaking as a function of time. No breaking case.*

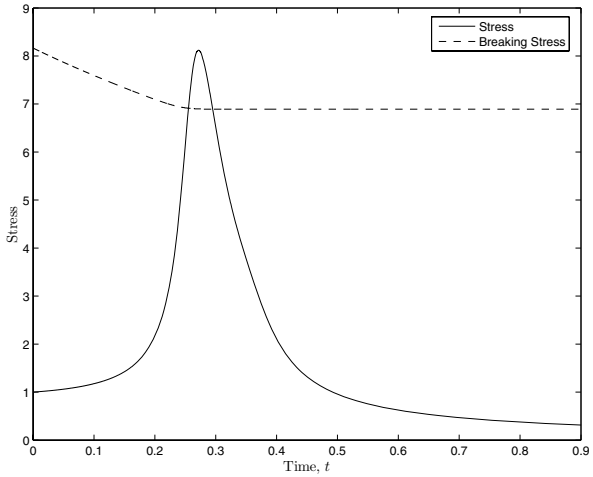


FIG. 3.15. *Dimensionless stress and dimensionless breaking stress in the glass tube closest to the location of breaking as a function of time. The tube breaks when the two curves first intersect.*

In Figures 3.14 and 3.15, we plot the dimensionless stress and the dimensionless breaking stress at the location that is closest to breaking as a function of time. For large values of $\mathcal{H}$ and C_b, as in Figure 3.14, the maximum stress is always less than the breaking stress. However, for smaller values of $\mathcal{H}$ and C_b, as in Figure 3.15, the stress may reach the breaking stress. This explains why vertical pullers typically require two pulls to break the glass tube. The first pull decreases the radius, which means that the new values of $\mathcal{H}$ and C_b for the second pull will be reduced. This implies that breaking will be more easily achieved in the second pull.

We caution that in the case when breaking does not occur, the tube becomes very thin, and inertia will ultimately become important, as shown by Stokes and Tuck [19]. Nevertheless, the inclusion of glass inertia will further reduce the viscous stress which makes it even more difficult to reach the breaking criterion. Therefore, the general

conclusion reached based on the simplified model remains useful. As we noted earlier, in this paper we focus only on successful pullings of microelectrodes, and the free-fall case will not be pursued here.

4. Discussion. We now discuss the effects of the parameters and puller designs on the shape of the final electrode. In order to do this, we will focus on the analytical solutions and results from the numerical method discussed in the previous section.

4.1. Shape control. There are two parameters that are relatively easy to vary continuously in an experimental context. These are the dimensional force, F_0, and the dimensional temperature of the heater, θ_h. In practice, one can use a graph of τ_{pinch} against $\mathcal{H}$ to choose the appropriate value of the heater strength $\mathcal{H}$ to achieve the required maximum area, $1 - \tau_{pinch}$. One then can choose C_b to give the required minimum area, s_b. Once the desired values of C_b and $\mathcal{H}$ are known, one simply chooses the dimensional force, F_0, to achieve the C_b value and then chooses θ_h to obtain the $\mathcal{H}$ value. This means that the appropriate operating conditions can be well approximated by simply using the universal graph of τ_{pinch} against $\mathcal{H}$.

While it is relatively easy to set the values of applied force and heater temperature, we also need to make sure that the glass breaks while the extension of the tube is within the physical length of the puller. This can be achieved by choosing the correct heater length, after the tip radius and other parameter values are determined. In Figure 4.1, we show how we can control the final tip radius by varying the applied force. We vary the applied force and use values of the other parameters from Tables 2.1 and 2.2. The simulation results (circles) and the approximate theory (solid line) show excellent agreement over many orders of magnitude in the applied force. By reducing the applied force, effectively, we can reduce the stress in the tube, and this allows the tube to reach a smaller radius before exceeding the breaking stress.

4.2. Sensitivity analysis. In order to determine the relative robustness of the symmetric and asymmetric methods, we performed a sensitivity analysis of the final shape with respect to changes in the parameters $\mathcal{H}$ and C_b. We found that the

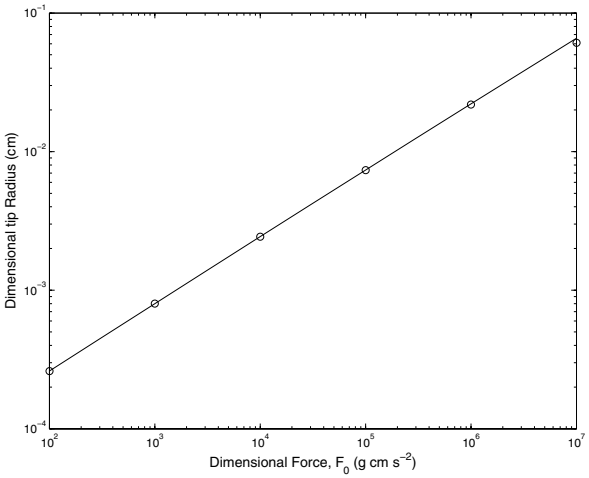


FIG. 4.1. *Dimensional tip radius as a function of the applied force.*

sensitivities are almost identical over a wide range of parameter values; therefore neither parameter has a significant advantage with regard to robustness. The symmetric puller has the advantage in that it can create two identical electrodes in a single pull. However, the asymmetric puller may be cheaper to build and easier to operate since it requires only a force be exerted at one end.

The sensitivity of vertical pullers to changes in parameters can be quite strong. This is particularly true if the glass breaks while the extension of the tube occurs near the maximum stress level. Then small increases in $\mathcal{H}$ or C_b might prevent the stress from reaching the breaking value, and no electrode would be formed. Even if this does not occur, the sensitivity of the variable force method is still substantially larger than for the constant force methods. The fact that a double pull is required also means that the procedure is much more difficult to implement and therefore less robust.

4.3. Concluding remarks. In this paper, we have simplified a model proposed in an earlier study (Huang et al., [10]) for glass microelectrode formation. Using a dimensional analysis argument, we have shown that the conductive heat transfer is small compared to the radiative and convective transfer and therefore can be neglected in the temperature equation. We have developed a Lagrangian-based method to solve the model equations and compute explicit solutions to the time-dependent equations for some simple cases. By investigating the effects of the parameters on the final shape of the microelectrodes, we have shown that vertical pullers are much less robust than horizontal pullers.

By considering the simplified models, we have been able to understand a number of important features of the resulting electrode. First, for typical parameter values, the surface tension is much smaller than the applied force, and we have shown that the ratio of the inner to outer radius will remain constant. Therefore, if a specified ratio is required in the final electrode, the only way to achieve this is to start with a tube that has that required ratio. Second, the length of the electrode tip is of the same order of magnitude as the length of the heater. This means that the tip length can be controlled by controlling the portion of the glass that is heated significantly. Third, using our approximate theory, we have shown that an excellent approximation to the tip width can be obtained from a very simple formula. This gives an extremely practical and straightforward method of determining the parameter values required to achieve a given tip radius.

In some cases, controlling the tip radius may not be sufficient, and the user may wish to control the entire tip profile. This can be achieved either by using a spatially dependent heater profile or by allowing a time-dependent pulling force. By choosing different heater profiles or pulling forces, one can produce a tip shape close to a desirable one. Since our method is simple, robust, and extremely efficient to implement numerically, it can be used to estimate the tip profile when the heating profile and/or pulling force are given. By using standard optimization techniques on this function, we then can achieve an approximation to the required profile.

Finally, the analysis and solution methodology presented in this paper are not restricted to glass microelectrode formation and may have a number of important applications in other glass formation processes. For example, the pulling of optical fibers uses a similar setup, even though the objective for optical fiber pulling is rather different. Instead of seeking the conditions to break the glass tube under stretching, a good optical fiber pulling device must ensure that the glass stretches without being broken. However, our model and numerical method can be readily applied to fiber pulling, with minor adjustments on the handling of boundary conditions.

Appendix A. Derivation of the temperature equation (2.18). In the long-wavelength limit of small radius to length aspect ratio, axial conduction can be neglected, and the heat equation is given by

$$\theta_t + u\theta_x = \frac{k}{\rho c_p}\frac{1}{y}\frac{\partial}{\partial y}\left(y\frac{\partial\theta}{\partial y}\right), \tag{A.1}$$

where y is the axisymmetric radial position and c_p and k are the specific heat capacity and the thermal conductivity of the fluid, respectively. The radiative boundary condition is given by

$$k\left.\frac{\partial\theta}{\partial y}\right|_{y=r} = 0, \tag{A.2}$$

$$-k\left.\frac{\partial\theta}{\partial y}\right|_{y=R} = \frac{k_B E_h \varepsilon_h \alpha\left(\theta_h^4 - \theta^4\right)}{1-(1-\alpha)(1-\varepsilon_h)} + \frac{k_B E_b \varepsilon_b \alpha\left(\theta_b^4 - \theta^4\right)}{1-(1-\alpha)(1-\varepsilon_b)}, \tag{A.3}$$

where k_B is the Boltzmann constant, α is the absorptivity, and θ_h is the heater temperature. We have assumed that the axial conduction is negligible and that the radiation heat exchange occurs only on the outer surface of the glass tube, as in [9].

We now nondimensionalize the heat equation (A.1) and the boundary condition (A.3) by using the scalings

$$u = u_0 u',\ \ s = s_0 s',\ \ y = R_0 y',\ \ x = \ell_0 x',\ \ t = \ell_0 u_0^{-1} t',\ \ \theta = \theta_0 + \theta_a \theta',\ \ u_0 = \frac{\ell_0 F_0}{3\mu_0 s_0}.$$

Dropping the primes, the heat equation becomes

$$Pe\left(\frac{\partial\theta}{\partial t} + u\frac{\partial\theta}{\partial x}\right) = \frac{1}{y}\frac{\partial}{\partial y}\left(y\frac{\partial\theta}{\partial y}\right), \tag{A.4}$$

where

$$Pe = \frac{\rho c_p u_0 R_0^2}{\ell_0 k} = 4.2\times 10^{-3} \tag{A.5}$$

is the transverse Peclet number, which represents the ratio of heat advected along the thread to heat conducted across the thread. The radiative boundary condition becomes

$$\left.\frac{\partial\theta}{\partial y}\right|_{y=\frac{r}{R_0}} = 0, \qquad \left.\frac{\partial\theta}{\partial y}\right|_{y=\frac{R}{R_0}} = BiH(x,\theta), \tag{A.6}$$

where

$$Bi = \frac{\alpha k_b \theta_h^4 R_0 \epsilon_h}{k\theta_a[1-(1-\alpha)(1-\epsilon_h)]} = 3.5\times 10^{-1} \tag{A.7}$$

is the Biot number. The dimensionless function

$$H(x,\theta) = E_h(x)\left(1-\left(\frac{\theta_0+\theta\theta_a}{\theta_h}\right)^4\right) + \frac{\epsilon_b[1-(1-\alpha)(1-\epsilon_h)]E_b(x)(\theta_b^4-(\theta_0+\theta_a\theta)^4)}{\epsilon_h[1-(1-\alpha)(1-\epsilon_b)]\theta_h^4}$$

represents the magnitude of the net heat flux that is absorbed when the temperature is θ. We now assume that both Bi and Pe are small, and from (2.2) we note that

$$\mathcal{H} \equiv \frac{2Bi}{(1-\beta_0^2)Pe}. \tag{A.8}$$

We assume that the temperature has an asymptotic expansion of the form

$$\theta = \Theta_0 + Bi\Theta_1 + \cdots; \tag{A.9}$$

then substituting this into (A.4) and (A.6) and collecting the terms to zeroth and first order in Bi yields

$$\frac{1}{y}\frac{\partial}{\partial y}\left(y\frac{\partial \Theta_0}{\partial y}\right) = 0 \quad \text{with} \quad \left.\frac{\partial \Theta_0}{\partial y}\right|_{y=\frac{r}{R_0}} = 0, \quad \left.\frac{\partial \Theta_0}{\partial y}\right|_{y=\frac{R}{R_0}} = 0 \tag{A.10}$$

and

$$\frac{1}{y}\frac{\partial}{\partial y}\left(y\frac{\partial \Theta_1}{\partial y}\right) = \frac{2}{(1-\beta_0^2)\mathcal{H}}\left(\frac{\partial \Theta_0}{\partial t} + u\frac{\partial \Theta_0}{\partial x}\right) \tag{A.11}$$

with

$$\left.\frac{\partial \Theta_1}{\partial y}\right|_{y=r} = 0, \quad \left.\frac{\partial \Theta_1}{\partial y}\right|_{y=R} = H\left(x, \Theta_0\right). \tag{A.12}$$

Equation (A.10) implies that at leading order Θ_0 is independent of y. Therefore, using (A.11)–(A.12), we see that Θ_0 satisfies

$$\frac{\partial \Theta_0}{\partial t} + u\frac{\partial \Theta_0}{\partial x} = \frac{\mathcal{H}H\left(x, \Theta_0\right)}{\sqrt{s}}\sqrt{\frac{1-\beta_0^2}{1-\beta^2}}, \tag{A.13}$$

where $\beta = r/R$. For notational brevity, we use θ to denote the leading order term, Θ_0, and obtain the equation in the final form as (2.18).

At the leading order, the viscosity is independent of the radial coordinate because the viscosity is a function of the temperature.

Appendix B. Exact solution for asymmetrical pulling. In this appendix, we obtain the exact solutions for the microelectrode shape and the temperature distribution for a glass tube undergoing asymmetrical pulling with uniform heating from a finite length heater. The initial velocity at each point in the glass tube is zero, and the other initial and boundary conditions are

$$\theta(\xi,0) = 0,\ s(\xi,0) = 1,\ x(\xi,0) = \xi,\ \theta(0,\tau) = 0,\ s(0,\tau) = 1,\ x(0,\tau) = 0. \tag{B.1}$$

The heater is located between $\xi = \ell_1$ and $\xi = \ell_2$. Let τ_b be the breaking time and τ_* be the time when the material point $\xi = \ell_1$ has passed location ℓ_2.

We need to consider two different cases: $\tau_b \le \tau_*$ and $\tau_b > \tau_*$.

B.1. Case 1: $\tau_b \le \tau_*$. Let $\xi_1(\tau_b)$ and $\xi_2(\tau_b)$ be the initial locations of the material points that are at ℓ_1 and ℓ_2 at τ_b. We have $\xi_1(\tau_b) < \ell_1 < \xi_2(\tau_b) < \ell_2$. Thus, there exist five regions: (1) $0 \le \xi \le \xi_1(\tau_b)$, (2) $\xi_1(\tau_b) \le \xi \le \ell_1$, (3) $\ell_1 \le \xi \le \xi_2(\tau_b)$, (4) $\xi_2(\tau_b) \le \xi \le \ell_2$, and (5) $\ell_2 \le \xi \le 1$.

B.1.1. $0 \le \xi \le \xi_1(\tau_b)$. In this region, for $0 < \tau < \tau_b$, we have

$$\theta_\tau = 0, \quad s_\tau = -e^\theta, \quad s x_\xi = 1,$$

which, combined with (B.1), gives

$$\theta_1 = 0, \quad s_1 = 1 - \tau, \quad x_1 = \frac{\xi}{1-\tau}.$$

Thus, at τ_b, the solution is

$$\theta_1 = 0, \quad s_1 = 1 - \tau_b, \quad x_1 = \frac{\xi}{1-\tau_b}, \tag{B.2}$$

from which we obtain $\xi_1(\tau_b) = (1-\tau_b)\ell_1$.

B.1.2. $\xi_1(\tau_b) \le \xi \le \ell_1$. (I). $0 \le \tau \le \tau_1(\xi)$, where $\tau_1(\xi)$ is the time when the material point that is initially at ξ crosses the point ℓ_1. The solution is

$$\theta_2^- = 0, \quad s_2^- = 1 - \tau, \quad x_2^- = \frac{\xi - \xi_1(\tau_b)}{1-\tau} + x_1(\xi_1(\tau_b), \tau) = \frac{\xi}{1-\tau},$$

from which we obtain

$$\ell_1 = \frac{\xi}{1 - \tau_1(\xi)}$$

or

$$\tau_1(\xi) = 1 - \frac{\xi}{\ell_1}.$$

(II). $\tau_1(\xi) \le \tau \le \tau_b$. In this region, we have

$$\theta_\tau = \frac{\mathcal{H}}{\sqrt{s}}, \quad s_\tau = -e^\theta, \quad s x_\xi = 1,$$

from which we obtain

$$\begin{aligned}
&\frac{2\mathcal{H}\sqrt{1-\tau_1(\xi)}+1}{2\mathcal{H}} \ln\left[1 + 2\mathcal{H}\left(\sqrt{1-\tau_1(\xi)} - \sqrt{s_2^+}\right)\right] \\
&\quad = \sqrt{1-\tau_1(\xi)} - \sqrt{s_2^+} + \mathcal{H}[\tau - \tau_1(\xi)], \\
&\quad \left(2\mathcal{H}\sqrt{1-\tau_1(\xi)}+1\right)\theta_2^+ + 1 - e^{\theta_2^+} = 2\mathcal{H}^2[\tau - \tau_1(\xi)], \\
&x_2^+ = \ell_1 + \int_{\xi_1(\tau)}^{\xi} \frac{1}{s_2^+(\tau_1(\eta))} d\eta.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{2\mathcal{H}\sqrt{1-\tau_1(\xi)}+1}{2\mathcal{H}} \ln\left[1 + 2\mathcal{H}\left(\sqrt{1-\tau_1(\xi)} - \sqrt{s_2^+}\right)\right] \\
&\quad = \sqrt{1-\tau_1(\xi)} - \sqrt{s_2^+} + \mathcal{H}[\tau_b - \tau_1(\xi)], \qquad \text{(B.3)} \\
&\quad \left(2\mathcal{H}\sqrt{1-\tau_1(\xi)}+1\right)\theta_2^+ + 1 - e^{\theta_2^+} = 2\mathcal{H}^2[\tau_b - \tau_1(\xi)], \qquad \text{(B.4)} \\
&x_2^+ = \ell_1 + \int_{\xi_1(\tau_b)}^{\xi} \frac{1}{s_2^+(\tau_1(\eta))} d\eta. \qquad \text{(B.5)}
\end{aligned}$$

B.1.3. $\ell_1 \le \xi \le \xi_2(\tau_b)$. For $0 \le \tau \le \tau_b$, we have

$$\theta_\tau = \frac{\mathcal{H}}{\sqrt{s}}, \quad s_\tau = -e^\theta, \quad s x_\xi = 1,$$

from which we have

$$\begin{aligned}
&\sqrt{s_3} - 1 + \frac{2\mathcal{H}+1}{2\mathcal{H}} \ln\left(1 + 2\mathcal{H} - 2\mathcal{H}\sqrt{s_3}\right) = \mathcal{H}\tau,\\
&(2\mathcal{H}+1)\,\theta_3 + 1 - e^{\theta_3} = 2\mathcal{H}^2\tau,\\
&x_3 = \frac{\xi - \ell_1}{s_3(\tau)} + x_2^+(\ell_1, \tau).
\end{aligned}$$

At τ_b, we have

$$\sqrt{s_3} - 1 + \frac{2\mathcal{H}+1}{2\mathcal{H}} \ln\left(1 + 2\mathcal{H} - 2\mathcal{H}\sqrt{s_3}\right) = \mathcal{H}\tau_b, \tag{B.6}$$

$$(2\mathcal{H}+1)\,\theta_3 + 1 - e^{\theta_3} = 2\mathcal{H}^2\tau_b, \tag{B.7}$$

$$x_3 = \frac{\xi - \ell_1}{s_3(\tau_b)} + x_2^+(\ell_1, \tau_b). \tag{B.8}$$

From the above equation, we determine $\xi_2(\tau_b)$ to be

$$\xi_2(\tau_b) = \ell_1 + [\ell_2 - x_2^+(\ell_1, \tau_b)] s_3(\tau_b). \tag{B.9}$$

B.1.4. $\xi_2(\tau_b) \le \xi \le \ell_2$. (I). $0 \le \tau \le \tau_2(\xi)$, where $\tau_2(\xi)$ is the time at which the material point ξ crosses the point ℓ_2. In this case, we have

$$\theta_\tau = \frac{\mathcal{H}}{\sqrt{s}}, \quad s_\tau = -e^\theta, \quad s x_\xi = 1,$$

from which we have

$$\begin{aligned}
&\sqrt{s_4^-} - 1 + \frac{2\mathcal{H}+1}{2\mathcal{H}} \ln\left(1 + 2\mathcal{H} - 2\mathcal{H}\sqrt{s_4^-}\right) = \mathcal{H}\tau,\\
&(2\mathcal{H}+1)\,\theta_4^- + 1 - e^{\theta_4^-} = 2\mathcal{H}^2\tau,\\
&x_4^- = \frac{\xi - \xi_2(\tau_b)}{s_4^-(\tau)} + x_3(\xi_2(\tau_b), \tau).
\end{aligned}$$

From the last equation, we obtain

$$\ell_2 == \frac{\xi - \xi_2(\tau_b)}{s_4^-(\tau_2(\xi))} + x_3(\xi_2(\tau_b), \tau_2(\xi)). \tag{B.10}$$

From this equation, we can find the value of $\tau_2(\xi)$.

(II). $\tau_2(\xi) \le \tau \le \tau_b$. In this case,

$$\theta_\tau = 0, \quad s_\tau = -e^\theta, \quad s x_\xi = 1,$$

from which we have

$$\begin{aligned}
&\theta_4^+(\xi, \tau) = \theta_4^+(\tau_2(\xi)) = \theta_4^-(\tau_2(\xi)),\\
&s_4^+(\xi, \tau) = \tau_2(\xi) - \tau + s_4^+(\tau_2(\xi)) = \tau_2(\xi) - \tau + s_4^-(\tau_2(\xi)),\\
&x_4^+(\xi, \tau) = \ell_2 + \int_{\xi_2(\tau)}^{\xi} \frac{1}{s_4^+(\eta, \tau)} d\eta.
\end{aligned}$$

At τ_b, we have

$$\theta_4^+(\xi,\tau_b)=\theta_4^+(\tau_2(\xi))=\theta_4^-(\tau_2(\xi)), \tag{B.11}$$

$$s_4^+(\xi,\tau_b)=\tau_2(\xi)-\tau_b+s_4^+(\tau_2(\xi))=\tau_2(\xi)-\tau_b+s_4^-(\tau_2(\xi)), \tag{B.12}$$

$$x_4^+(\xi,\tau_b)=\ell_2+\int_{\xi_2(\tau_b)}^{\xi}\frac{1}{s_4^+(\eta,\tau_b)}d\eta. \tag{B.13}$$

B.1.5. $\ell_2 \le \xi \le 1$. In this region,

$$\theta_\tau=0,\quad s_\tau=-e^\theta,\quad sx_\xi=1$$

is valid for $0\le\tau\le\tau_b$, from which we have

$$\theta_5=0,\quad s_5=1-\tau,\quad x_5=\frac{\xi-\ell_2}{1-\tau}+x_4^+(\ell_2,\tau).$$

At τ_b, we have

$$\theta_5=0,\quad s_5=1-\tau_b,\quad x_5=\frac{\xi-\ell_2}{1-\tau_b}+x_4^+(\ell_2,\tau_b). \tag{B.14}$$

B.2. Case 2: $\tau_b > \tau_*$. In this case, there are also five regions: (1) $0\le\xi\le\xi_1(\tau_b)$, (2) $\xi_1(\tau_b)\le\xi\le\xi_2(\tau_b)$, (3) $\xi_2(\tau_b)\le\xi\le\ell_1$, (4) $\ell_1\le\xi\le\ell_2$, and (5) $\ell_2\le\xi\le1$.

B.2.1. $0\le\xi\le\xi_1(\tau_b)$. In this region, for $0<\tau<\tau_b$, the solution is the same as that in Case 1,

$$\theta_1=0,\quad s_1=1-\tau,\quad x_1=\frac{\xi}{1-\tau},$$

and at τ_b the solution is

$$\theta_1=0,\quad s_1=1-\tau_b,\quad x_1=\frac{\xi}{1-\tau_b}, \tag{B.15}$$

from which we obtain $\xi_1(\tau_b)=(1-\tau_b)\ell_1$.

B.2.2. $\xi_1(\tau_b)\le\xi\le\xi_2(\tau_b)$. (I). When $0\le\tau\le\tau_1(\xi)$, the solution is

$$\theta_2^-=0,\quad s_2^-=1-\tau,\quad x_2^-=\frac{\xi-\xi_1(\tau_b)}{1-\tau}+x_1(\xi_1(\tau_b),\tau)=\frac{\xi}{1-\tau},$$

from which we obtain

$$\ell_1=\frac{\xi}{1-\tau_1(\xi)}$$

or

$$\tau_1(\xi)=1-\frac{\xi}{\ell_1}.$$

(II). When $\tau_1(\xi)\le\tau\le\tau_b$ in this region, we have

$$\theta_\tau=\frac{\mathcal{H}}{\sqrt{s}},\quad s_\tau=-e^\theta,\quad sx_\xi=1,$$

from which we obtain

$$\frac{2\mathcal{H}\sqrt{1-\tau_1(\xi)}+1}{2\mathcal{H}}\ln\left[1+2\mathcal{H}\left(\sqrt{1-\tau_1(\xi)}-\sqrt{s_2^+}\right)\right]$$
$$=\sqrt{1-\tau_1(\xi)}-\sqrt{s_2^+}+\mathcal{H}[\tau-\tau_1(\xi)],$$

$$\left(2\mathcal{H}\sqrt{1-\tau_1(\xi)}+1\right)\theta_2^+ +1-e^{\theta_2^+}=2\mathcal{H}^2[\tau-\tau_1(\xi)],$$

$$x_2^+=\ell_1+\int_{\xi_1(\tau)}^{\xi}\frac{1}{s_2^+(\tau_1(\eta))}d\eta.$$

Therefore,

$$\frac{2\mathcal{H}\sqrt{1-\tau_1(\xi)}+1}{2\mathcal{H}}\ln\left[1+2\mathcal{H}\left(\sqrt{1-\tau_1(\xi)}-\sqrt{s_2^+}\right)\right]$$
$$=\sqrt{1-\tau_1(\xi)}-\sqrt{s_2^+}+\mathcal{H}[\tau_b-\tau_1(\xi)], \tag{B.16}$$

$$\left(2\mathcal{H}\sqrt{1-\tau_1(\xi)}+1\right)\theta_2^+ +1-e^{\theta_2^+}=2\mathcal{H}^2[\tau_b-\tau_1(\xi)], \tag{B.17}$$

$$x_2^+=\ell_1+\int_{\xi_1(\tau_b)}^{\xi}\frac{1}{s_2^+(\tau_1(\eta))}d\eta. \tag{B.18}$$

The value of $\xi_2(\tau_b)$ can be obtained from the following equation:

$$\ell_2=\ell_1+\int_{\xi_1(\tau_b)}^{\xi_2(\tau_b)}\frac{1}{s_2^+(\tau_1(\eta))}d\eta. \tag{B.19}$$

This is a nonlinear equation for $\xi_2(\tau_b)$, which can be obtained using an iterative method, after replacing the integral by a numerical quadrature.

B.2.3. $\boldsymbol{\xi_2(\tau_b)\le\xi\le\ell_1}$. (I). When $0\le\tau\le\tau_1(\xi)$, the solution is

$$\theta_3^-=0,\quad s_3^-=1-\tau,\quad x_3^-=\frac{\xi}{1-\tau}.$$

(II). When $\tau_1(\xi)\le\tau\le\tau_b$ in this region, we have

$$\theta_\tau=\frac{\mathcal{H}}{\sqrt{s}},\quad s_\tau=-e^\theta,\quad sx_\xi=1,$$

from which we obtain

$$\frac{2\mathcal{H}\sqrt{1-\tau_1(\xi)}+1}{2\mathcal{H}}\ln\left[1+2\mathcal{H}\left(\sqrt{1-\tau_1(\xi)}-\sqrt{s_3^*}\right)\right]$$
$$=\sqrt{1-\tau_1(\xi)}-\sqrt{s_3^*}+\mathcal{H}[\tau-\tau_1(\xi)], \tag{B.20}$$

$$\left(2\mathcal{H}\sqrt{1-\tau_1(\xi)}+1\right)\theta_3^* +1-e^{\theta_3^*}=2\mathcal{H}^2[\tau-\tau_1(\xi)], \tag{B.21}$$

$$x_3^*=x_2^+(\xi_2(\tau_b),\tau)+\int_{\xi_2(\tau_b)}^{\xi}\frac{1}{s_3^*(\eta,\tau)}d\eta. \tag{B.22}$$

(III). When $\tau_* \le \tau \le \tau_b$ in this region, the solution is

$$\theta_3^+(\xi,\tau) = \theta_3^+(\xi,\tau_2(\xi)) = \theta_3^*(\xi,\tau_2(\xi)),$$
$$s_3^+(\xi,\tau) = s_3^+(\xi,\tau_2(\xi)) + \tau_2(\xi) - \tau = s_3^*(\xi,\tau_2(\xi)) + \tau_2(\xi) - \tau,$$
$$x_3^+(\xi,\tau) = \ell_2 + \int_{\xi_2(\tau)}^{\xi} \frac{1}{s_3^+(\eta,\tau)} d\eta,$$

where $\theta_3^*(\xi,\tau_2(\xi))$ and $s_3^*(\xi,\tau_2(\xi))$ are from (B.21) and (B.20), with τ replaced by $\tau_2(\xi)$, and $\tau_2(\xi)$ is obtained by applying (B.22) at ℓ_2:

$$\ell_2 = \ell_1 + \int_{\xi_1(\tau)}^{\xi} \frac{1}{s_3^*(\eta,\tau_2(\xi))} d\eta. \tag{B.23}$$

This is a nonlinear equation for $\tau_2(\xi)$, which can be solved using an iterative method.

At τ_b, we have

$$\theta_3^+(\xi,\tau_b) = \theta_3^+(\xi,\tau_2(\xi)) = \theta_2^*(\xi,\tau_2(\xi)), \tag{B.24}$$
$$s_3^+(\xi,\tau_b) = s_3^+(\xi,\tau_2(\xi)) + \tau_2(\xi) - \tau_b = s_2^*(\xi,\tau_2(\xi)) + \tau_2(\xi) - \tau_b, \tag{B.25}$$
$$x_3^+(\xi,\tau_b) = \ell_2 + \int_{\xi_2(\tau_b)}^{\xi} \frac{1}{s_3^+(\eta,\tau_b)} d\eta. \tag{B.26}$$

B.2.4. $\boldsymbol{\ell_1 \le \xi \le \ell_2}$. (I). When $0 \le \tau \le \tau_2'(\xi)$ in this region, we have

$$\sqrt{s_4^-} - 1 + \frac{2\mathcal{H}+1}{2\mathcal{H}} \ln\left(1 + 2\mathcal{H} - 2\mathcal{H}\sqrt{s_4^-}\right) = \mathcal{H}\tau,$$
$$(2\mathcal{H}+1)\,\theta_4^- + 1 - e^{\theta_4^-} = 2\mathcal{H}^2\tau,$$
$$x_4^- = \frac{\xi - \ell_1}{s_4^-(\tau)} + x_3^*(\ell_1,\tau).$$

From the last equation, we obtain

$$\ell_2 = \frac{\xi - \ell_1}{s_4^-(\tau_2'(\xi))} + x_3^*(\ell_1,\tau_2'(\xi)). \tag{B.27}$$

From this equation, we can find the value of $\tau_2'(\xi)$, which is the time that the material point which was initially at ξ crosses ℓ_2. Note that τ_2' is different from τ_2 in region (3) since the solutions in the two regions are different.

(II). When $\tau_2'(\xi) \le \tau \le \tau_*$ in this region, we have

$$\theta_4^*(\xi,\tau) = \theta_4^*(\tau_2'(\xi)) = \theta_4^-(\tau_2'(\xi)),$$
$$s_4^*(\xi,\tau) = \tau_2'(\xi) - \tau + s_4^*(\tau_2'(\xi)) = \tau_2'(\xi) - \tau + s_4^-(\tau_2'(\xi)),$$
$$x_4^*(\xi,\tau) = \ell_2 + \int_{\xi_2(\tau)}^{\xi} \frac{1}{s_4^*(\eta,\tau)} d\eta.$$

(III). When $\tau_* \le \tau \le \tau_b$ in this region we have

$$\theta_4^+(\xi,\tau) = \theta_4^+(\tau_2'(\xi),\tau_*),$$
$$s_4^+(\xi,\tau) = \tau_* - \tau + s_4^*(\tau_2'(\xi)),$$
$$x_4^+(\xi,\tau) = x_3^+(\ell_1,\tau) + \int_{\ell_1}^{\xi} \frac{1}{s_4^+(\eta,\tau)} d\eta.$$

At τ_b, we have

$$\theta_4^+(\xi,\tau_b) = \theta_4^+(\tau_2'(\xi),\tau_*), \tag{B.28}$$

$$s_4^+(\xi,\tau_b) = \tau_* - \tau_b + s_4^*(\tau_2'(\xi)), \tag{B.29}$$

$$x_4^+(\xi,\tau_b) = x_3^+(\ell_1,\tau_b) + \int_{\ell_1}^{\xi} \frac{1}{s_4^+(\eta,\tau_b)}\,d\eta. \tag{B.30}$$

B.2.5. $\ell_2 \le \xi \le 1$. In this region we have

$$\theta_5 = 0, \quad s_5 = 1-\tau, \quad x_5 = \frac{\xi-\ell_2}{1-\tau} + x_4^*(\ell_2,\tau)$$

for $0 \le \tau \le \tau_*$ and

$$\theta_5 = 0, \quad s_5 = 1-\tau, \quad x_5 = \frac{\xi-\ell_2}{1-\tau} + x_4^+(\ell_2,\tau)$$

for $\tau_* \le \tau \le \tau_b$.

At τ_b, we have

$$\theta_5 = 0, \quad s_5 = 1-\tau_b, \quad x_5 = \frac{\xi-\ell_2}{1-\tau_b} + x_4^+(\ell_2,\tau_b). \tag{B.31}$$

Acknowledgments. We wish to thank Drs. Demetrius Papageorgiou, Michael Siegel, Yuan-Nan Young, and Wendy Zhang for useful discussions at the Focused Research Group (FRG), in Banff. Also, we thank the Banff International Research Station (BIRS) for funding the FRG, and the staff at BIRS for their wonderful efforts to make the FRG such a productive event. Finally we wish to express our gratitude to the anonymous referees who helped us to improve the paper by providing constructive comments and suggestions.

REFERENCES

[1] M. Coenen, *Festigkeit von Glasschmelzen*, Glastech. Ber., 51 (1978), pp. 17–20.
[2] Corning Glass Company, *Pyrex Glass Code* 7740*, Material Properties*, Brochure Pyrex B-87, 1987.
[3] M. M. Denn, *Continuous drawing of liquids to form fibers*, in Annu. Rev. Fluid Mech. 12, Annual Reviews, Palo Alto, CA, 1980, pp. 365–387.
[4] J. Dewynne, J. R. Ockendon, and P. Wilmott, *On a mathematical model for fiber tapering*, SIAM J. Appl. Math., 49 (1989), pp. 983–990.
[5] J. N. Dewynne, J. R. Ockendon, and P. Wilmott, *A systematic derivation of the leading-order equations for extensional flows in slender geometries*, J. Fluid Mech., 244 (1992), pp. 323–338.
[6] A. D. Fitt, K. Furusawa, T. M. Monro, and C. P. Please, *Modeling the fabrication of hollow fibers: Capillary drawing*, J. Lightwave Technol., 19 (2001), pp. 1924–1931.
[7] A. D. Fitt, K. Furusawa, T. M. Monro, and C. P. Please, *The mathematical modelling of capillary drawing for holey fiber manufacture*, J. Engrg. Math., 43 (2002), pp. 201–227.
[8] D. G. Flaming and K. T. Brown, *Micropipette puller design, Form of the heating filament and effects of filament width on tip length and diameter*, J. Neurosci. Methods, 6 (1982), pp. 91–102.
[9] P. Gospodinov and A. L. Yarin, *Drawing resonance of optical microcapillaries in non-isothermal drawing*, Int. J. Multiphase Flow, 23 (1997), pp. 967–976.
[10] G. Gupta and W. W. Schultz, *Non-isothermal flows of Newtonian slender glass fibers*, Int. J. Nonlinear Mech., 33 (1998), pp. 151–163.
[11] G. Gupta, W. W. Schultz, E. M. Arruda, and X. Lu, *Nonisothermal model of glass fiber drawing stability*, Rhoel. Acta, 35 (1996), pp. 584–596.

[12] H. Huang, R. M. Miura, W. P. Ireland, and E. Puil, *Heat-induced stretching of a glass tube under tension: Application to glass microelectrodes*, SIAM J. Appl. Math., 63 (2003), pp. 1499–1519.
[13] A. Kaye, *Convected coordinates and elongational flow*, J. Non-Newtonian Fluid Mech., 40 (1991), pp. 55–77.
[14] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser, Boston, Cambridge, MA, 1992.
[15] S. Martinoia, P. Massobrio, M. Bove, and G. Massobrio, *Cultured neurons coupled to microelectrode arrays: Circuit models, simulations and experimental data*, IEEE Trans. Biomed. Eng., 51 (2004), pp. 859–964.
[16] A. Peskoff and R. S. Eisenberg, *Interpretation of some microelectrode measurements of electrical properties of cells*, Ann. Rev. Biomed. Eng., 2 (1973), pp. 65–80.
[17] H. Scholze, *Glass, Nature, Structure, and Properties*, (translated by M. J. Lakin), Springer-Verlag, New York, 1990, pp. 255–272.
[18] E. M. Snell, *Some electrical properties of fine tipped pipette microelectrodes*, in Glass Microelectrodes, M. Lavallee, A. F. Shanne, and N. C. Hubert, eds., Wiley, New York, 1969, pp. 111–123.
[19] Y. M. Stokes and E. O. Tuck, *The role of inertia in extensional fall of a viscous drop*, J. Fluid Mech., 498 (2004), pp. 205–225.
[20] S. D. R. Wilson, *The slow dripping of a viscous fluid*, J. Fluid Mech., 190 (1988), pp. 561–570.

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 667–686

PASSIVITY OF MAGNETOSTRICTIVE MATERIALS*

SINA VALADKHAN†, KIRSTEN MORRIS‡, AND AMIR KHAJEPOUR†

Abstract. Magnetostrictive materials display large force and displacement in response to an applied field, as well as short response time. However, their nonlinear and hysteretic behavior has hindered their use. We prove, using the physics of the material, that these materials are passive. The corresponding energy storage function is shown to be the Helmholtz energy. This result is independent of the model used. The effect of varying load is included. Passivity is important because it can be used to obtain control systems that maintain stability despite uncertainties and disturbances. The minima of the storage function are also obtained. The storage function is written explicitly in the case of a common model for these materials, the Preisach model.

Key words. passivity, smart materials, Preisach model, magnetostrictive materials, stability

AMS subject classifications. 93D09, 93D25, 93D15, 93A30, 82D40

DOI. 10.1137/060651264

1. Introduction.

There has been a growing demand by industry in recent years for micropositioning devices. Micropositioning actuators are now frequently seen in scanning microscopes, chip manufacturing machines, biological cell micromanipulation and optical fiber alignment devices. Currently, many of these micropositioning tasks are done with piezoceramic actuators. Piezoceramic actuators exhibit almost linear behavior and have a reasonably fast response time.

Still, there is a demand for actuators with a larger stroke and faster response time. For this reason, the possibility of using other active materials for actuation is being examined. Terfenol-D, an alloy of iron, terbium, and dysprosium, has many advantages. Terfenol-D is a magnetostrictive material. Compared to other active materials, it has very large force and displacement with a short response time that makes it an attractive choice for actuation.

The use of magnetostrictive materials has been hindered by the fact that their response is highly nonlinear and hysteretic. Because of this nonlinearity, Terfenol-D actuators are difficult to control. In many micropositioning tasks, submicron accuracy is required. To achieve the required performance, actuators need to be used in a closed-loop feedback system. The controller in the feedback system must be able to stabilize the system under all conditions.

Dependence of the hysteresis on many physical conditions together with the nonlinear nature of the system make it difficult to establish stability for the closed loop. External physical conditions such as mechanical loading and temperature affect the behavior of magnetostrictive materials. Stability and performance of the control system must be maintained despite these system uncertainties and also despite disturbances. One of the most useful methods for showing stability of nonlinear systems is *passivity*. There are many passive physical systems [1]. Passive systems are important because the stability of closed-loop passive systems can be easily established. For

*Received by the editors January 30, 2006; accepted for publication (in revised form) October 20, 2006; published electronically March 2, 2007.
http://www.siam.org/journals/siap/67-3/65126.html
†Department of Mechanical Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada (svaladkh@uwaterloo.ca, akhajepour@uwaterloo.ca).
‡Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada (kmorris@uwaterloo.ca).

many nonlinear systems, this approach is the only way to show stability. Passivity has been used to obtain closed-loop stability for nonlinear systems in over 300 papers published in the last 10 years.

The Preisach model [2] is among the oldest models for magnetic materials. This model has been successfully applied to many hysteretic systems [3, 4]. In [5], the Preisach model is used with a set of ordinary differential equations to develop a rate-dependent hysteresis model. Open-loop stability and other properties of the model are discussed. These results are used to develop a model inverse-based controller for a magnetostrictive actuator [6].

In [7], an energy-based version of the Preisach model is introduced. Unlike the classical Preisach model, this model is based on a physical model for the material. In [3], it is shown that the Preisach operator is passive if the system output is the time-derivative of the output. The associated storage function is also computed. The result is applied to the control of a shape memory alloy actuator. In [8], this approach is extended to position control. The passivity results [3] are used in [9] to establish asymptotic stability of closed-loop systems containing hysteresis.

In the next section we give a brief review of standard material on passivity. It is subsequently shown, using physics, that magnetostrictive materials are passive. The storage function is identified to be the Helmholtz energy. No assumption on the model is used. The effects of varying load are included. The Preisach model is then introduced and the energy storage function is written explicitly using this model. The system equilibrium points are identified and discussed.

2. Passivity. In this section, passivity is defined in a dynamical systems framework. This framework will be used later for magnetostrictive materials. Consider a system with input $u \in U$, output $y \in U$, and state $x \in X$. The following is a standard definition for dynamical systems [1].

DEFINITION 1. *A dynamical system is defined through input, output and state spaces U and X, a readout operator r, and a state transition operator ϕ. The readout operator is a map from $U \times X$ to U. The state transition operator is a map from $\mathbb{R}^2 \times X \times U$ to X. The state transition operator must have the following properties for all $x_0 \in X$, t_0, t_1, $t_2 \in \mathbb{R}$, u, u_1, $u_2 \in U$:*

Consistency: $\phi(t_0, t_0, x_0, u) = x_0$.

Determinism: $\phi(t_1, t_0, x_0, u_1) = \phi(t_1, t_0, x_0, u_2)$ *for all* $t_1 \geq t_0$ *when* $u_1(t) = u_2(t)$ *for all* $t_0 \leq t \leq t_1$.

Semigroup: $\phi(t_2, t_0, x_0, u) = \phi(t_2, t_1, \phi(t_1, t_0, x_0, u), u)$ *when* $t_0 \leq t_1 \leq t_2$.

Stationarity: $\phi(t_1 + T, t_0 + T, x_0, u_T) = \phi(t_1, t_0, x_0, u)$ *for all* $t_1 \geq t_0$, $T \in \mathbb{R}$ *when* $u_T(t) = u(t + T)$ *for all* $t \in \mathbb{R}$.

DEFINITION 2 (see [1]). *Consider a dynamical system with state variables x, an input u, and output y. If there is a real-valued function $S(x)$ satisfying the following relation for any $t_i \leq t_f$ and if $S(x)$ is bounded from below, the dynamical system is called passive:*

$$S(x(t_i)) + \int_{t_i}^{t_f} \langle u, y \rangle \, dt \geq S(x(t_f)). \tag{1}$$

In this definition, $\langle ., . \rangle$ is the inner product on U. The variables u and y are vectors of the same dimension, so that $\langle u, y \rangle$ is defined. The scalar function $S(x)$ is called the storage function. Passive systems are frequently seen in engineering. The storage function is often the energy.

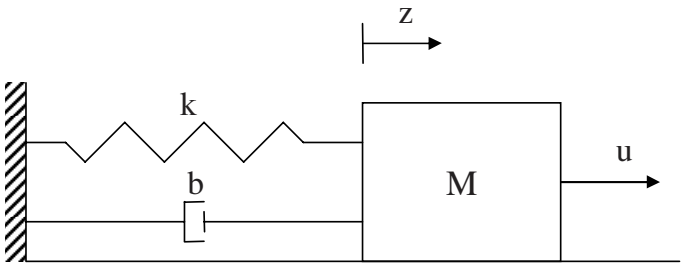


FIG. 1. *A spring-mass-dashpot system.*

Example. Consider a spring-mass-dashpot system (Figure 1). The following equation describes this system:

$$M\frac{d^2z}{dt^2} + b\frac{dz}{dt} + kz = u, \tag{2}$$

where u is the external force applied. The velocity of the mass $\dot{z}$ is considered to be the system output: $y = \dot{z}$. The state variables are z and $\dot{z}$. If both sides of (2) are multiplied by $\dot{z}$ and integrated from t_i to t_f, it becomes

$$\frac{M}{2}\left(\dot{z}^2(t_f) - \dot{z}^2(t_i)\right) + \int_{t_i}^{t_f} b\dot{z}^2 dt + \frac{k}{2}\left(z^2(t_f) - z^2(t_i)\right) = \int_{t_i}^{t_f} \langle u, y\rangle \, dt. \tag{3}$$

In this example, total energy is

$$E(z, \dot{z}) = \frac{1}{2}kz^2 + \frac{1}{2}M\dot{z}^2. \tag{4}$$

Using this definition, (3) can be rewritten as

$$E(z(t_i), \dot{z}(t_i)) + \int_{t_i}^{t_f} \langle u, y\rangle \, dt \geq E(z(t_f), \dot{z}(t_f)). \tag{5}$$

The storage function $E(z, \dot{z})$ is always nonnegative and, hence, bounded from below. As a result, this system is passive. When $u = 0$, the system goes to a state which minimizes E. The energy E is minimized when $z = 0$, $\dot{z} = 0$. This is the global system equilibrium point.

When the force applied to the system includes a constant force, such as gravity, its effect can be included in the system storage function. If the force applied to the mass is $F_{const} + u$, the following storage function is minimized at the equilibrium point:

$$\bar{E} = E - F_{const}z. \tag{6}$$

In this case, the equilibrium point is $z = \frac{F_{const}}{k}$, $\dot{z} = 0$.

Define the operator $\|.\|$ to be the Euclidean norm; that is, for any vector v, $\|v\|^2 = \langle v, v\rangle$. The following definitions are used to establish stability for the system [10, 11].

DEFINITION 3. *The set L_2 is the set of functions $x : \mathbb{R} \to \mathbb{R}^n$ for which the following expression is bounded:*

$$\int_0^\infty \|x(t)\|^2 \, dt < \infty. \tag{7}$$

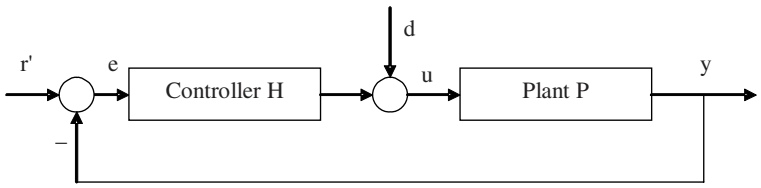


FIG. 2. *The standard feedback configuration.*

DEFINITION 4. *The set L_{2e} is the set of functions $x : \mathbb{R} \to \mathbb{R}^n$ for which the following expression is bounded for all $T \in \mathbb{R}$:*

$$\int_0^T \|x(t)\|^2 \, dt < \infty. \tag{8}$$

DEFINITION 5. *A mapping $R : L_{2e} \to L_{2e}$ is said to be L_2-stable if $x \in L_2$ implies that $Rx \in L_2$.*

Suppose that a given system P is passive. Consider the general feedback control configuration shown in Figure 2. If the controller H satisfies certain conditions, the following result can be used to show the stability of the controlled system.

THEOREM 6 (see [11, Theorem 10, p. 182]). *Consider the feedback system shown in Figure 2, where H and P map U to U. The set U is a subset of L_{2e}. Assume that for any r' and d in L_2 there are solutions e and u in L_{2e} and there are constants α_1, α_2, α_3, β_1, β_2, and β_3 such that for every real T and $x \in L_{2e}$ the following conditions hold:*

$$\begin{aligned} &\text{I} \quad \sqrt{\int_0^T \|Hx\|^2 \, dt} \le \alpha_1 \sqrt{\int_0^T \|x\|^2 \, dt} + \beta_1, \\ &\text{II} \quad \int_0^T \langle x, Hx \rangle \, dt \ge \alpha_2 \int_0^T \|x\|^2 \, dt + \beta_2, \\ &\text{III} \quad \int_0^T \langle Px, x \rangle \, dt \ge \alpha_3 \int_0^T \|Px\|^2 \, dt + \beta_3. \end{aligned} \tag{9}$$

If $\alpha_2 + \alpha_3 > 0$, then r', $d \in L_2$ implies that e, u, He, $y \in L_2$.

A passive system satisfies the third condition with $\alpha_3 = 0$ and $\beta_3 = \inf S(x) - S(x(0))$. The second and third conditions are similar to requiring that plant and controller be passive, but slightly stronger since $\alpha_2 + \alpha_3$ has to be strictly positive. The last line of the theorem states that the closed loop is L_2-stable.

This theorem can be used to establish stability for a large class of nonlinear systems. For many systems this theorem is the only way to establish stability. The passivity results which will be shown later can be used with this theorem to show stability for the magnetostrictive system.

3. Passivity for magnetostrictive materials. Since magnetostrictive materials dissipate energy, we expect them to be passive with some energy function as the storage function. In this section, the physical parameters of magnetostrictive materials are introduced. Three different energy functions for magnetostrictive materials and their suitability as a storage function are discussed. Finally, a proof of passivity is given.

Magnetostrictive materials react to a magnetic field. Suppose that a magnetostrictive sample is excited in a magnetic field produced by a coil. If there is an electrical

current in the coil, a nonzero magnetic field H is seen around the coil. Magnetic field H is a vector field, and it depends on the electrical current and the geometry. Magnetic field H acts on the magnetostrictive sample, and it is usually considered to be the input for the hysteretic system. As a result of this magnetic field, a magnetization M is seen in the material. The magnetization M is also a vector field, and it is considered to be the response or output of the hysteretic system. The relation between H and M depends on the material.

The magnetization M is not the only parameter affected by an external magnetic field H. The mechanical variables are also affected. For a material where the magnetic and mechanical responses are decoupled, the stress σ is usually considered to be the input for the mechanical part, and the strain ε, the response. For magnetostrictive materials, a magnetic field affects both magnetization and strain, and similarly for the stress. For magnetostrictive materials, generalized force and displacement are defined as follows:

$$F = \begin{pmatrix} \mu_0 H \\ \sigma \end{pmatrix}, \tag{10}$$

$$X = \begin{pmatrix} M \\ \varepsilon \end{pmatrix}. \tag{11}$$

Generalized force F is the system input and time-derivative of generalized displacement $\dot{X}$, the output. The constant μ_0 is a physical constant to ensure that $\mu_0 \langle H, M \rangle$ has the unit of energy per unit volume.

Various energy functions can be associated with magnetostrictive materials. Here these energy functions are introduced and their suitability as a storage function is discussed.

3.1. The internal energy. The internal energy U is the total potential energy stored in the material. The first law of thermodynamics holds for this energy function:

$$\frac{dU}{dt} = \frac{dQ}{dt} + \frac{dW}{dt}, \tag{12}$$

where $\frac{dQ}{dt}$ is the rate of thermal energy supplied to the material and $\frac{dW}{dt}$ is the rate of magnetic/mechanical work done on the system. The inequality of Clausius [12, p. 205] states that for any process $\frac{dS}{dt} \geq \frac{1}{T}\frac{dQ}{dt}$, where T is the temperature and S is the entropy. Using this inequality, the first law can be written as

$$\frac{dU}{dt} \leq T\frac{dS}{dt} + \frac{dW}{dt}. \tag{13}$$

A relation similar to the passivity inequality can be obtained by integrating both sides of (13) from t_i to t_f:

$$U_i + \int_{t_i}^{t_f} \left(T\frac{dS}{dt} + \frac{dW}{dt} \right) dt \geq U_f. \tag{14}$$

It is seen that thermal terms should appear in the system input/output; i.e., u should be $\begin{pmatrix} \mu_0 H \\ \sigma \\ T \end{pmatrix}$ and y should be $\begin{pmatrix} \dot{M} \\ \dot{\varepsilon} \\ \dot{S} \end{pmatrix}$. Since the energy stored in the material is limited, the amount of energy which can be pulled out of the material is also limited. This means that the energy function U has a lower bound. As a result, the internal energy U can be used as a storage function.

Thermal variables are usually difficult to work with, and for magnetostrictive materials they are difficult to measure. Extra thermal input and output are disadvantages to using internal energy as a storage function. For this reason, the internal energy is not chosen as the storage function.

3.2. The Gibbs energy. The following relation defines the Gibbs energy:

$$G = U - TS - \langle F, X \rangle . \tag{15}$$

Using the relation $\frac{dW}{dt} = \left\langle F, \frac{dX}{dt} \right\rangle$ and (10), (11) and (13), we obtain

$$\frac{dG}{dt} \leq -S\frac{dT}{dt} - \mu_0 \left\langle M, \frac{dH}{dt} \right\rangle - \varepsilon \frac{d\sigma}{dt}. \tag{16}$$

The Gibbs energy is a function of H. This means that H has to be included in the system states. This is awkward for several reasons. First, in this application H is an input. Second, consider a situation in which $\varepsilon = 0$ and H has a large value. The Gibbs energy can be made arbitrarily small by increasing H. This means that the Gibbs energy does not have a lower bound, and hence it is not a suitable storage function.

3.3. The Helmholtz energy. The Helmholtz free energy ψ is defined as

$$\psi = U - TS, \tag{17}$$

where T and S are the temperature and total entropy, respectively, of the system. Using the inequality of Clausius, the first law of thermodynamics can be written as

$$\frac{d\psi}{dt} \leq -S\frac{dT}{dt} + \frac{dW}{dt}. \tag{18}$$

Under constant temperature, this equation simplifies to

$$\frac{d\psi}{dt} \leq \frac{dW}{dt}. \tag{19}$$

This relation states that the work provided is more than the rate at which Helmholtz free energy is increased. It can be said that part of the work energy provided is absorbed by the system and added to the stored energy, while the rest is wasted in energy dissipation. It seems that the Helmholtz free energy is the energy actually stored in the system. In this respect, the Helmholtz energy is comparable to the energy storage function E in the mechanical example. Since the energy E is the storage function for the mechanical example, this comparison suggests the Helmholtz energy as the storage function. In the next subsection, a detailed proof of passivity, with the Helmholtz free energy as the storage function, is given.

3.4. Proof of passivity. It is assumed that, during any process discussed here, no phase transition occurs; for example, the material is not melting. This guarantees the existence of partial derivatives. All of the processes are under constant air pressure. Work done by the air pressure is neglected. For simplicity, from now on, it is also assumed that the thermal connection between the material and the surrounding environment is so good that the temperature of the material is always close to the room temperature T_0 and constant.

In a magnetic material, the ratio between the dipole magnetic energy and the energy of thermal fluctuations plays an important role. If the dipole magnetic energy

is small compared to thermal fluctuations, the material is called *paramagnetic*. In this case, the dipoles are mostly affected by thermal fluctuations and the external magnetic field H. Dipole-dipole interaction is weak. Because of thermal fluctuations, paramagnetic materials are memoryless and have no hysteresis. On the other hand, if the dipole magnetic energy is large compared to thermal fluctuations, the material is called *ferromagnetic*. Dipoles in a ferromagnetic sample retain their state, and the material has memory. These materials are hysteretic. Because of strong dipole-dipole interactions in ferromagnetic materials, the models available for these materials are complex and difficult to use. The energy of thermal fluctuations depends linearly on temperature. For this reason if a ferromagnetic material is heated, in a certain temperature it becomes paramagnetic. This transition temperature is called the Curie temperature T_c. Curie temperature is fairly high for most of the ferromagnetic materials. For iron $T_c = 1043\text{K}$.

When a ferromagnetic material is heated beyond T_c, it becomes paramagnetic, and during this heating process, the entropy of the materials is increased. In the following lemmas, this fact is used together with entropy relations for a paramagnetic material to show an upper bound for the entropy in a ferromagnetic material. The first lemma is used to show that the Helmholtz free energy has a lower bound.

LEMMA 7. *For a paramagnetic material at a constant temperature, the entropy S has an upper bound.*

Proof. The strength of a magnetic dipole is denoted by a constant positive half-integer J. This constant depends on the material under discussion. The following equations define entropy for a single dipole in a paramagnetic sample [13, pp. 213, 215, and 259]:

$$\begin{aligned} \beta &= \frac{1}{kT}, \\ \eta &= c\beta \left\| H \right\|, \\ Z &= \frac{\sinh\left[(J+\frac{1}{2})\eta\right]}{\sinh\left[\frac{1}{2}\eta\right]}, \\ S &= k\left(\ln Z - \beta \frac{\partial \ln Z}{\partial \beta}\right), \end{aligned} \tag{20}$$

where c is a positive constant and k is the Boltzmann constant $k = 1.38e-23\frac{\text{J}}{\text{K}}$.

In a paramagnetic sample with N dipoles, total magnetic entropy is simply N times the entropy of a single dipole. Total magnetic entropy is maximized when $H = 0$. (See the appendix.) This result is consistent with physics since in the presence of an external magnetic field, dipoles become oriented and the overall system disorder is reduced. Thus,

$$S_{\max} = S_{H=0} = kN \ln(2J+1). \tag{21}$$

Thus, at a constant temperature, the magnetic portion of entropy has an upper bound, $S_{\max}$.

The nonmagnetic portion of the entropy is a function of temperature and external load. At any temperature, this entropy is maximized for the highest possible (tensile) external load. This means that at any temperature, the nonmagnetic portion of the entropy has an upper bound. Thus at any temperature, the total entropy has an upper bound. □

The paramagnetic state is usually obtained at a high temperature. In order to have an upper bound for entropy in normal working conditions of the material, the lemma above should be extended to ferromagnetic materials.

LEMMA 8. *For any magnetic material at a constant temperature, the entropy S has an upper bound.*

Proof. Lemma 7 states that the entropy has an upper bound for the paramagnetic state. Here we are interested in the ferromagnetic state.

To obtain a relation for entropy in the ferromagnetic state, consider a process in which the ferromagnetic material is heated from an arbitrary initial state to a state in which the material is paramagnetic. The entropy and temperature for the initial state are S_i and T_i, respectively. For the paramagnetic state, the entropy and temperature are S_p and T_p, respectively. From Lemma 7, it is known that S_p has an upper bound.

The entropy is a function of the system states [12, p. 217]. The difference between any two arbitrary states is only a function of the states. This difference is independent of the process which connects the two states. This fact holds for the process mentioned above. The difference $S_p - S_i$ does not depend on the process as long as the initial and final conditions remain the same. For simplicity, consider a process in which the temperature is increased monotonically.

Since the temperature is always increasing during this process, there should be a nonnegative heat flow to the material during the process:

$$\frac{dQ}{dt} \geq 0. \tag{22}$$

The inequality of Clausius states that for any process $\frac{dS}{dt} \geq \frac{1}{T}\frac{dQ}{dt}$. As a result, in this process $\frac{dS}{dt} \geq 0$ or $S_p - S_i \geq 0$. Since S_p has an upper bound, S_i is bounded from above. This concludes the proof. ∎

The following is an immediate result of the lemma above.

THEOREM 9. *For a constant temperature, the Helmholtz free energy $\psi = U - TS$ is bounded from below.*

Proof. Lemma 8 states that the entropy has an upper bound. This means that $-TS$ has a lower bound. The internal energy U has a lower bound. This results in ψ being bounded from below. ∎

THEOREM 10. *The following passivity condition is satisfied when the storage function is the Helmholtz free energy ψ:*

$$\psi_i + \int_{t_i}^{t_f} \left\langle F, \frac{dX}{dt} \right\rangle dt \geq \psi_f. \tag{23}$$

Here, subscripts i and f denote initial and final conditions, respectively; F and X are the generalized force applied to the system and the generalized system output, respectively, as defined in (10) *and* (11)*; and σ and ε are stress and strain, respectively.*

Proof. If the temperature is constant, (18) can be written as

$$\frac{d\psi}{dt} \leq \frac{dW}{dt}, \tag{24}$$

where $\frac{dW}{dt} = \left\langle F, \frac{dX}{dt} \right\rangle$ is the rate of magnetic/mechanical work done on the system.

If both sides are integrated from t_i to t_f, we obtain

$$\psi_f - \psi_i \leq \int_{t_i}^{t_f} \left\langle F, \frac{dX}{dt} \right\rangle dt \tag{25}$$

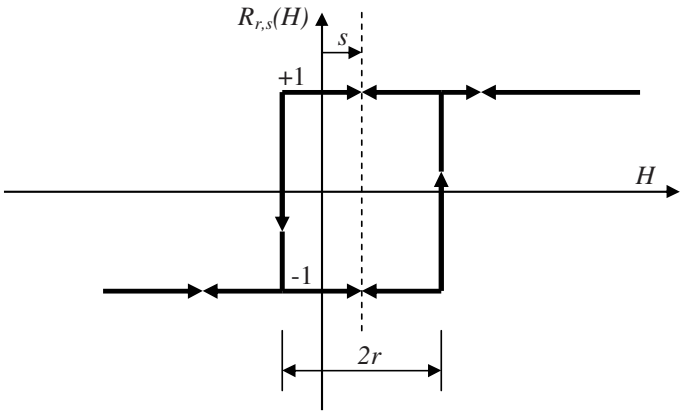


FIG. 3. *The Preisach relay.*

or

$$\psi_i + \int_{t_i}^{t_f} \left\langle F, \frac{dX}{dt} \right\rangle dt \geq \psi_f. \tag{26}$$

Theorem 9 shows that the Helmholtz free energy is bounded from below, which means that it is a valid storage function. This concludes the proof. □

The proof above shows the passivity of a magnetostrictive system with a three-dimensional magnetic field and a one-dimensional stress-strain. In this proof, no model for the magnetostrictive material is assumed. Passivity is shown with fundamental laws of physics only. In fact, the theorem above can be applied to any model for magnetostrictive materials.

4. The Preisach model. The Preisach model [2] is a very common model in the smart materials literature; for examples, see [3, 4, 14, 15]. In [15], it is used to model magnetostrictive materials. It has been shown that this model can represent magnetostrictive materials accurately [16]. This model is briefly explained here; for a detailed description, see [2]. In this model, a one-dimensional magnetic field is assumed, which results in the magnetic field H and magnetization M being scalars. It is assumed that the output is the weighted sum of the output of a continuum of hysteresis relays. The output of each relay can be either $+1$ or -1, determined by the previous relay value and the input, magnetic field H. In Figure 3 a typical hysteresis relay is shown.

The model output is

$$M(t) = \int_0^\infty \int_{-\infty}^\infty R_{r,s}[H(\cdot)](t)\mu(r,s)dsdr. \tag{27}$$

Here, $R_{r,s}$ is the output of the relay defined by r and s, and $\mu(r,s)$ is a weight function determined by experimental data.

Consider a two-dimensional coordinate system with variables r and s as shown in Figure 4. Each point r, s in this coordinate system is in a one-to-one relation with a Preisach relay $R_{r,s}$ and its corresponding weight $\mu(r,s)$. The plane defined by variables r and s is called the Preisach plane. Because the system input is limited, the relays with a large r or s do not change and cannot contribute to a change in the

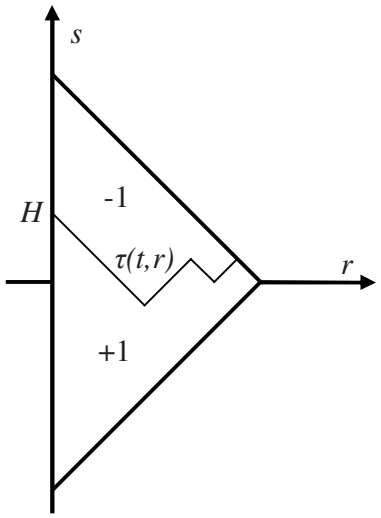


FIG. 4. *A typical Preisach plane boundary.*

model output. For simplicity, it is assumed that the weight function $\mu(r,s)$ is zero for these relays. In Figure 4, these relays are outside of the bold triangle. Since the output is not affected by these relays, they are not considered.

If, in the Preisach plane, the relays equal to -1 are separated with a line from the relays at $+1$, a boundary $s = \tau(t,r)$ will be produced, as shown in Figure 4. This boundary is important since if $\tau(t,r)$ is available, the output of all relays are known. Thus, knowledge of $\tau(t,r)$ and the input $H(t)$ determines future values of $\tau(t,r)$ and hence $M(t)$. In other words, $\tau(t,r)$ contains the *memory* of the system. The Preisach model is a dynamical system with $\tau(t,r)$ as the state [17]. The model output can be rewritten in terms of the boundary:

$$M(t) = 2\int_0^\infty \int_{-\infty}^{\tau(t,r)} \mu(r,s)dsdr - \int_0^\infty \int_{-\infty}^{\infty} \mu(r,s)dsdr. \tag{28}$$

Note that the Preisach boundary $\tau(t,r)$ and the vertical axis $r = 0$ in Figure 4 intersect at the current input value; that is,

$$\tau(t,0) = H. \tag{29}$$

4.1. Energy-based Preisach model. In this model, a physical model for magnetostrictive materials is used to develop a special type of Preisach model that is based on energy considerations [7, 16]. Here, the material is assumed to be composed of a large number of weakly interacting dipoles. The Helmholtz free energy for a single dipole can be modeled by three parabolas [7], [15, p. 188] (Figure 5):

$$\psi(M,\varepsilon) = \frac{1}{2}Y\varepsilon^2 - Y\gamma\varepsilon M^2 + \begin{cases} \frac{\mu_0\eta'}{2}(M+M_R)^2, & M \le -M_I, \\ \frac{\mu_0\eta'}{2}(M-M_R)^2, & M \ge M_I, \\ \frac{\mu_0\eta'}{2}(M_R-M_I)(M_R-\frac{M^2}{M_I}), & |M| < M_I, \end{cases} \tag{30}$$

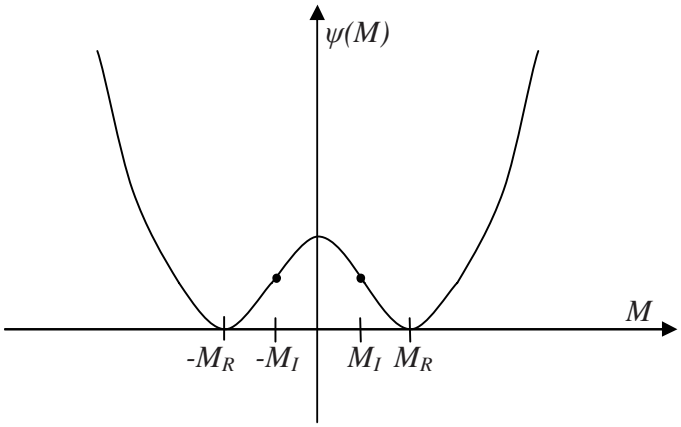


FIG. 5. *The Helmholtz free energy.*

where the variable M is the magnetization for the dipole, the parameter η' is a constant, γ is the magnetomechanical coupling constant, and Y is Young's modulus. The parameter M_R is the remanence magnetization. In the absence of strain ε, $\pm M_R$ are the minima of ψ. The parameter M_R is assumed to be the same for all dipoles. The parameter M_I is the inflection point where the second derivative of ψ changes sign. Unlike M_R, because of the nonhomogeneities and imperfections in the material, M_I is different for each dipole. For a valid Helmholtz free energy $M_R > M_I$. This ensures that the Helmholtz free energy has two distinct minima, as shown in Figure 5.

Define H_0 to be the local magnetic field at a dipole. Because of the imperfections and nonhomogeneities in the material, the local magnetic field H_0 might not be equal to the external magnetic field H. It is assumed that the difference $s = H - H_0$ is constant over time for each dipole.

The parameters s and M_I describe each dipole. Define

$$r = \eta'(M_R - M_I) + \frac{2}{\mu_0} Y\gamma\varepsilon M_I. \tag{31}$$

It will be shown later that it is easier to use r as defined in (31) to describe each dipole instead of M_I. This definition of r is consistent with r for a Preisach relay, as shown in Figure 3.

For a dipole, the Gibbs energy is

$$G_{r,s}(H_0, M_{r,s}, \sigma, \varepsilon) = \psi_{r,s}(M_{r,s}, \varepsilon) - \mu_0 H_0 M_{r,s} - \sigma\varepsilon, \tag{32}$$

as shown in Figure 6.

Consider a single dipole in a process in which the temperature, magnetic field H, and stress are constant. In this case, (16) simplifies to

$$\frac{dG_{r,s}}{dt} \leq 0. \tag{33}$$

This relation states that during this process, G has to either stay constant or decrease. At a stable equilibrium point, the Gibbs energy is minimized [15, pp. 65 and 184]. In this case, the derivative of Gibbs energy has to be zero with respect to

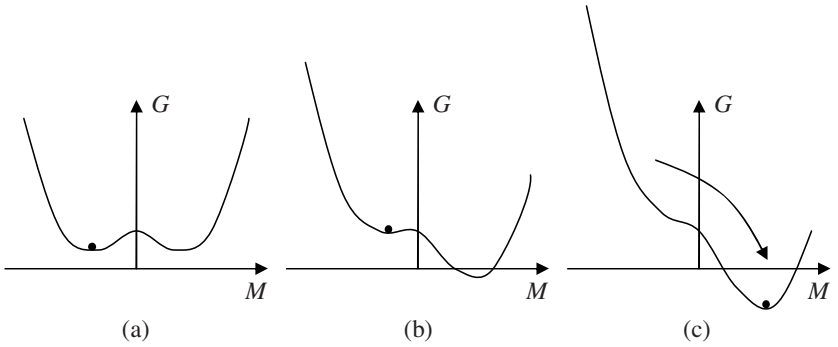


FIG. 6. (a) *Gibbs energy when* $H_0 = 0$, (b) *Gibbs energy for a positive* H_0, (c) *if* H_0 *is further increased, at some point, only one minimum exists.*

unconstrained variables:

$$\left(\frac{\partial G_{r,s}(H_0, M_{r,s}, \sigma, \varepsilon)}{\partial M_{r,s}}\right)_{T,H_0,\sigma,\varepsilon} = 0, \tag{34}$$

$$\left(\frac{\partial G_{r,s}(H_0, M_{r,s}, \sigma, \varepsilon)}{\partial \varepsilon}\right)_{T,H_0,\sigma,M_{r,s}} = 0. \tag{35}$$

By combining (32) and (34), the following relation is obtained:

$$\mu_0 H_0 = \left(\frac{\partial \psi}{\partial M_{r,s}}\right)_{T,\varepsilon}. \tag{36}$$

In a magnetic system with many dipoles, the dipole dynamics are very fast. If the magnetic field is not very rapidly changing, the magnetic field appears to be almost constant for each dipole over the time constant of the dipole. The magnetization for a dipole is a minimum of the Gibbs energy.

By combining (30), (32), and (36), the equilibrium magnetization for a dipole is obtained:

$$M^*_{r,s} = \frac{H - s + R_{r,s}\eta' M_R}{\eta' - \frac{2Y\gamma\varepsilon}{\mu_0}}. \tag{37}$$

If the dipole is in the left minimum in Figure 6(a), $R_{r,s} = -1$, and if the dipole is in the right minimum, $R_{r,s} = +1$.

As seen in Figure 6, if $H_0 = 0$, two minima exist. For a small positive H_0 as shown in Figure 6(b), still two minima exist, but if H_0 is further increased, at some point, one disappears, as shown in Figure 6(c). At this time, dipole magnetization moves to the new minimum. This transition is shown with an arrow in Figure 6(c).

Using (31), it can be shown that if $H \geq s + r$, the $R = -1$ minimum does not exist. Similarly, for $H \leq s - r$, the $R = +1$ minimum vanishes. For $s - r < H < s + r$, two minimums exist, which means that both $R = -1$ and $R = +1$ are possible. It is seen that for the Preisach relay introduced in Figure 3, the output -1 is nonexistent if $H \geq s + r$, and $+1$ vanishes if $H \leq s - r$. For the values between $s - r$ and $s + r$, both outputs are possible. This similarity between the dipole and a Preisach relay shows that the definition of r and s are consistent with r and s of a Preisach relay.

For a large magnetic field, the dipole magnetization is the right minimum. At this minimum, the Gibbs energy is

$$G_{r,s} = \frac{1}{2}Y\varepsilon^2 + \frac{\mu_0}{2}\eta M_R^2 - \frac{\mu_0(H - s + \eta M_R)^2}{2(\eta - \frac{2Y\gamma\varepsilon}{\mu_0})} - \sigma\varepsilon. \tag{38}$$

It is seen that, if $\varepsilon = 0$, the Gibbs energy can be made arbitrarily small by increasing H. This means that the Gibbs energy is unbounded from below.

Assuming a distribution $\mu(r, s)$ for the dipoles, the overall magnetization can be obtained:

$$M_{Tot} = C \int_0^\infty \int_{-\infty}^\infty M_{r,s}^* \mu(r, s) ds dr. \tag{39}$$

Define $\mathcal{I}_n$ to be

$$\mathcal{I}_n = \int_0^\infty \int_{-\infty}^\infty s^n \mu(r, s) ds dr, \tag{40}$$

where $n = 0, 1$, or 2. Using (37), M_{Tot} can be written as follows:

$$M_{Tot} = \frac{C}{\eta' - \frac{2Y\gamma\varepsilon}{\mu_0}} \left[\mathcal{I}_0(H - M_R\eta') - \mathcal{I}_1 + 2M_R\eta' \int_0^\infty \int_{-\infty}^{\tau(t,r)} \mu(r, s) ds dr \right], \tag{41}$$

where C is a constant and $\tau(t, r)$ is the Preisach boundary for the relay configuration $R_{r,s}$. The experimental data can be used to find the optimum weight function $\mu(r, s)$. A few common choices for $\mu(r, s)$ can be found in [16, 18].

Unlike the Preisach model, magnetization in this model depends on ε. In this model, σ and H are the inputs. The Preisach plane boundary $\tau(t, r)$ and strain ε are the system states. The outputs are ε and M. The magnetization is determined by (41). Combining (30), (32), and (35), we obtain

$$\varepsilon = \frac{\sigma}{Y} + \gamma M^2, \tag{42}$$

which determines strain ε.

4.2. Helmholtz free energy using the Preisach model. In this section, the total Helmholtz free energy for a magnetostrictive material is calculated using the physical Preisach model. Since this function is the system storage function, it is written as a function of system states $\tau(t, r)$ and ε.

As stated before, the local magnetic field H_0 might not be equal to the external magnetic field H. This difference between H and H_0 should have some effect on the energy functions. For example, consider a dipole with a negative s when the dipole magnetization is increased by dM and the external magnetic field H is constant: Work done by the external magnetic source is HdM, and work done on the dipole is $H_0 dM = HdM - sdM$. It is seen that the work done on the dipole is more that the work done by the external magnetic field. This extra work is not done by the external field. The imperfections and nonhomogeneities which are the source of the difference between H and H_0 should have done this work on the dipole. As a result, they need to be considered when the overall system Helmholtz free energy is computed.

From (32), we have $G_{r,s}(H_0, M_{r,s}) = \psi_{r,s}(M_{r,s}) - \mu_0 H_0 M_{r,s} - \sigma\varepsilon$. Define $\bar{\psi}_{r,s}(M_{r,s})$ and $\bar{G}_{r,s}(H, M_{r,s})$ to be the Helmholtz free energy and Gibbs energy, respectively, written in terms of external variables. When the system is viewed from an

external point of view, the combined effect of the dipole and the imperfections is seen. To find $\bar{\psi}_{r,s}(M_{r,s})$ and $\bar{G}_{r,s}(H, M_{r,s})$, an assumption for the imperfections and nonhomogeneities must be made and, based on that, the contribution to the Helmholtz free energy computed. Another approach is to construct $\bar{\psi}_{r,s}(M_{r,s})$ by studying the equilibrium points of the system for a constant magnetic field.

The equilibrium points for a constant magnetic field in terms of the external variables (H,$M_{r,s}$) can be obtained via two methods:

1. The equilibrium condition can be written for $\bar{G}_{r,s}(H, M_{r,s})$.
2. The system parameters can be transformed to the local variables (H_0, $M_{r,s}$). The equilibrium condition is written for $G_{r,s}(H_0, M_{r,s})$, and the results are transformed back to the external variables.

These two methods must be equivalent.

The equilibrium conditions for $\bar{G}_{r,s}(H, M_{r,s})$ and $G_{r,s}(H_0, M_{r,s})$ are

$$\left(\frac{\partial \bar{G}_{r,s}(H, M_{r,s})}{\partial M_{r,s}}\right)_{T,H} = 0, \qquad \left(\frac{\partial G_{r,s}(H_0, M_{r,s})}{\partial M_{r,s}}\right)_{T,H_0} = 0, \tag{43}$$

where $H = H_0 + s$ and s is assumed constant. Further,

$$\begin{aligned}\left(\frac{\partial G(H_0, M_{r,s})}{\partial M_{r,s}}\right)_{T,H_0} &= \left(\frac{\partial}{\partial M_{r,s}}\right)_{T,H} (\psi(M_{r,s}) - \mu_0 H_0 M_{r,s} - \sigma\varepsilon) \\ &= \left(\frac{\partial}{\partial M_{r,s}}\right)_{T,H} (\psi(M_{r,s}) - \mu_0 H M_{r,s} + \mu_0 s M_{r,s} - \sigma\varepsilon) \\ &= 0.\end{aligned} \tag{44}$$

Now, $\bar{G}_{r,s}(H, M_{r,s})$ equals $G_{r,s}(H - s, M_{r,s})$ or

$$\bar{G}_{r,s}(H, M_{r,s}) = \psi(M_{r,s}) - \mu_0 H M_{r,s} + \mu_0 s M_{r,s} - \sigma\varepsilon. \tag{45}$$

It can be shown that the equilibrium conditions (43) are identical. Defining $\bar{\psi}_{r,s}(M_{r,s})$ so that $\bar{G}_{r,s}(H, M_{r,s}) = \bar{\psi}_{r,s} - \mu_0 H M_{r,s} - \sigma\varepsilon$, analogously with (32), we have

$$\bar{\psi}_{r,s}(M_{r,s}) = \psi(M_{r,s}) + \mu_0 s M_{r,s}. \tag{46}$$

Equation (37) gives the equilibrium magnetization for a dipole. By combining (30), (37), and (46), the equilibrium value of $\bar{\psi}_{r,s}$ for each dipole is obtained:

$$\bar{\psi}^*_{r,s} = \frac{1}{2} Y\varepsilon^2 + \frac{\frac{\mu_0}{2}(H^2 - s^2) - \eta' M_R (Y\gamma\varepsilon M_R - \mu_0 s R_{r,s})}{\eta' - \frac{2Y\gamma\varepsilon}{\mu_0}}. \tag{47}$$

Similar to (39), by assuming a distribution for r and s, the Helmholtz free energy for the entire system can be found using the superposition principle:

$$\psi_{Tot}(\tau(t,r), \varepsilon) = C \int_0^\infty \int_{-\infty}^\infty \bar{\psi}^*_{r,s} \mu(r,s) ds dr. \tag{48}$$

By combining (29), (47), and (48), the following equation is obtained:

$$\begin{aligned}\psi_{Tot}(\tau(t,r), \varepsilon) = \frac{C\mathcal{I}_0}{2} Y\varepsilon^2 + \frac{C}{\eta' - \frac{2Y\gamma\varepsilon}{\mu_0}} \Bigg(&\frac{\mu_0 \mathcal{I}_0 \tau^2(t,0)}{2} - \eta' Y\gamma\varepsilon M_R^2 \mathcal{I}_0 \\ &+ \eta' M_R \mu_0 A - \frac{\mu_0}{2} \mathcal{I}_2 \Bigg),\end{aligned} \tag{49}$$

where $A = \int_0^\infty \int_{-\infty}^\infty R_{r,s} s \mu(r,s) ds dr = 2 \int_0^\infty \int_{-\infty}^{\tau(t,r)} s \mu(r,s) ds dr - \mathcal{I}_1$.

This is the value of the Helmholtz free energy, the storage function for the magnetostrictive system, for any ε and Preisach boundary $\tau(t,r)$. The only nontrivial aspect of calculating $\psi_{Tot}(\tau(t,r),\varepsilon)$ is efficient computation of A. It is seen that the double integral of A is very similar to the double integral used for computing M (27). In fact, any efficient algorithm used for the computation of M can be used here, for example that on [2, p. 37]; only the weight function is slightly different.

4.3. Minimum of the storage function. In this section, the Preisach boundary that globally minimizes the storage function is obtained.

Suppose that when $\tau(t,r)=\tau^*(t,r)$ and $\varepsilon=\varepsilon^*$, $\psi_{Tot}(\tau(t,r),\varepsilon)$ is globally minimized. If ε is held fixed at $\varepsilon=\varepsilon^*$ and $\tau(t,r)$ is changed, $\psi_{Tot}(\tau(t,r),\varepsilon^*)$ is minimized when $\tau(t,r)=\tau^*(t,r)$. This means that $\tau^*(t,r)$ globally minimizes the following function:

$$\psi_{Tot}(\tau(t,r),\varepsilon^*)=\frac{C\mathcal{I}_0}{2}Y\varepsilon^{*2}+\frac{C}{\eta'-\frac{2Y\gamma\varepsilon^*}{\mu_0}}\left(\frac{\mu_0\mathcal{I}_0\tau^2(t,0)}{2}-\eta' Y\gamma\varepsilon^* M_R^2\mathcal{I}_0 + \eta' M_R\mu_0 A-\frac{\mu_0}{2}\mathcal{I}_2\right). \tag{50}$$

The following terms are the only variable parts of the storage function:

$$\begin{aligned} F_1(\tau(t,r))&=\frac{\mu_0\mathcal{I}_0\tau^2(t,0)}{2},\\ F_2(\tau(t,r))&=A. \end{aligned} \tag{51}$$

Assume that the weight function $\mu(r,s)$ is nonnegative for all r and s. Since $\eta'-\frac{2Y\gamma\varepsilon^*}{\mu_0}$ is a positive quantity, if F_1 and F_2 are minimized at the same time, the storage function is minimized. Function F_1 is minimized when $\tau(t,0)=0$. Function F_2 is minimized when A is minimized:

$$A=2\int_0^\infty\int_{-\infty}^{\tau(t,r)} s\mu(r,s)dsdr-\mathcal{I}_1. \tag{52}$$

The sign of the integrand equals the sign of s. This integration is minimized when the region of integration is the subset of the Preisach plane on which the integrand is negative. This is the lower half of the Preisach plane. Thus, the integration is minimized when the boundary $\tau(t,r)=0$. This Preisach plane boundary is shown in Figure 7.

Function F_2 is globally minimized with the boundary $\tau(t,r)=0$. Since for this boundary $\tau(t,0)=0$, this boundary also globally minimizes F_1. This results in global minimization of the storage function.

It is commonly seen that the weight function $\mu(r,s)$ is an even function of s; that is, $\mu(r,s)=\mu(r,-s)$ for all r and s [16, 18]. If this condition holds, by substituting the Preisach boundary $\tau(t,r)=0$ into (41), it is seen that the resulting magnetization is zero. In this case there is no magnetic field H, magnetization M, or flux density B. This state is called the demagnetized state and is the state of lowest "energy" for the system.

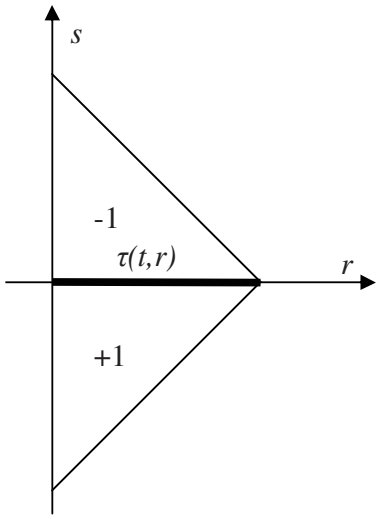


FIG. 7. *The global minimum Preisach boundary.*

5. Storage function in the presence of a constant input. When the stress and magnetic field applied to the system include a constant portion, the system can be simplified by redefining the input as $\bar{u} = \begin{pmatrix} \mu_0(H - H_{const}) \\ \sigma - \sigma_{const} \end{pmatrix}$, while the output is not changed. In this case, the system is passive with the following storage function:

$$\psi^F = \psi_{Tot} - \mu_0 \langle H_{const}, M_{Tot} \rangle - \sigma_{const}\varepsilon, \tag{53}$$

where ψ_{Tot} is the system Helmholtz free energy and M_{Tot} is the system magnetization. This situation is analogous to the example of a spring with a constant imposed force, such as gravity, discussed in section 2.

THEOREM 11. *In the presence of a constant input, the following passivity condition is satisfied when the storage function is* ψ^F*:*

$$\psi_i^F + \int_{t_i}^{t_f} \left\langle \bar{u}, \frac{dX}{dt} \right\rangle dt \geq \psi_f^F. \tag{54}$$

Subscripts i *and* f *denote initial and final conditions, respectively, and* $X = \begin{pmatrix} M \\ \varepsilon \end{pmatrix}$ *is the generalized displacement.*

Proof. If the definition of $\bar{u}$ and ψ^F is substituted into the result of Theorem 10, the result is

$$\begin{aligned} \psi_i^F + \mu_0 \langle H_{const}, M_{Tot,i} \rangle + \sigma_{const}\varepsilon_i + \int_{t_i}^{t_f} \left\langle \bar{u} + \begin{pmatrix} \mu_0 H_{const} \\ \sigma_{const} \end{pmatrix}, \frac{dX}{dt} \right\rangle dt \\ \geq \psi_f^F + \mu_0 \langle H_{const}, M_{Tot,f} \rangle + \sigma_{const}\varepsilon_f. \end{aligned} \tag{55}$$

This simplifies to

$$\psi_i^F + \int_{t_i}^{t_f} \left\langle \bar{u}, \frac{dX}{dt} \right\rangle dt \geq \psi_f^F. \tag{56}$$

Since both M_{Tot} and ε have a lower bound and an upper bound, existence of a lower bound for ψ_{Tot} implies that ψ^F has a lower bound. The proof is complete. ∎

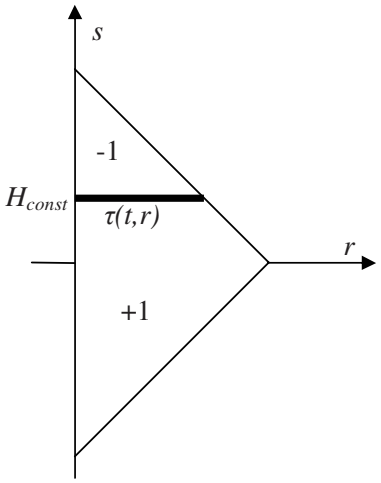


FIG. 8. *The global minimum Preisach boundary in the presence of a constant input.*

The storage function ψ^F can be written as a function of the Preisach boundary $\tau(t,r)$ and ε by combining (49) and (53):

$$\psi^F(\tau(t,r),\varepsilon) = \frac{C\mathcal{I}_0}{2}Y\varepsilon^2 + \frac{C}{\eta' - \frac{2Y\gamma\varepsilon}{\mu_0}}\left(\frac{\mu_0\mathcal{I}_0(\tau(t,0)-H_{const})^2}{2} - \frac{\mu_0\mathcal{I}_0H_{const}^2}{2}\right.$$
$$- \eta' Y\gamma\varepsilon M_R^2\mathcal{I}_0 + \mu_0 H_{const}M_R\eta'\mathcal{I}_0 + \mu_0 H_{const}\mathcal{I}_1$$
$$\left. - \eta' M_R\mu_0\mathcal{I}_1 + \mu_0\eta' M_R\bar{A} - \frac{\mu_0}{2}\mathcal{I}_2\right) - \sigma_{const}\varepsilon, \tag{57}$$

where $\bar{A} = 2\int_0^\infty \int_{-\infty}^{\tau(t,r)}(s - H_{const})\mu(r,s)dsdr$. Using an argument similar to that of the previous section, it can be shown that the following boundary minimizes the storage function:

$$\tau(t,r) = H_{const}. \tag{58}$$

This boundary is shown in Figure 8. For a constant input, this is the state of minimum energy. The magnetization in this state is the anhysteretic magnetization.

6. Conclusions. In this article, magnetostrictive transducers were introduced in a dynamical system framework. Passivity of this system was shown using fundamental physical relations. For the energy-based Preisach model, the system states were defined, and the storage function was computed. System equilibrium points were also identified and discussed.

The passivity results discussed in this paper can be used to show the stability of a closed-loop system. Future work includes the design and optimization of a robustly stabilizing controller for magnetostrictive transducers.

Appendix. The maximization of entropy. In this appendix, it is shown that the entropy function (20) is maximized when $H = 0$.

From subsection 3.4, the following relations define entropy for a paramagnetic sample with N dipoles:

$$\begin{aligned} Z &= \frac{\sinh\left[(J+\frac{1}{2})\eta\right]}{\sinh\left[\frac{1}{2}\eta\right]}, \\ S &= kN\left(\ln Z - \beta\frac{\partial \ln Z}{\partial\beta}\right), \\ \eta &= c\beta\left\|H\right\|, \\ \beta &= \frac{1}{kT}, \end{aligned} \tag{59}$$

where c is a positive constant.

Define $D = \frac{\eta}{2} = \frac{c\beta}{2}\left\|H\right\|$ and $q = 2J+1$. Since J is a positive half-integer, q is an integer greater than one. We can write

$$S = kN\left(\ln\frac{\sinh qD}{\sinh D} - qD\coth qD + D\coth D\right). \tag{60}$$

This function is not defined at $D=0$, but $\lim_{D\to 0} S(D)$ exists:

$$\begin{aligned} \lim_{D\to 0} S(D) &= \lim_{D\longrightarrow 0} kN\left(\ln\frac{\sinh qD}{\sinh D} + \frac{D\cosh D\sinh qD - qD\cosh qD\sinh D}{\sinh D\sinh qD}\right) \\ &= \lim_{D\longrightarrow 0} kN\left(\ln\frac{qD+h.o.t.}{D+h.o.t.} + \frac{\frac{D^4}{6}(2q-2q^3)+h.o.t.}{qD^2+h.o.t.}\right) \\ &= kN\ln q. \end{aligned} \tag{61}$$

For $D\neq 0$, $S(D) = S(-D)$; i.e., this is an even function. We do not need to analyze this function for both positive and negative values of D. For simplicity $D>0$ is studied.

If $D>0$,

$$\frac{dS}{dD} = kN\left(\frac{q^2 D}{\sinh^2 qD} - \frac{D}{\sinh^2 D}\right). \tag{62}$$

It will be shown that for $D>0$, $\frac{dS}{dD} < 0$. Consider the Taylor series of the following expression:

$$\begin{aligned} \sinh qD - q\sinh D &= qD + \frac{q^3D^3}{3!} + \frac{q^5D^5}{5!} + \cdots - qD - \frac{qD^3}{3!} - \frac{qD^5}{5!} - \cdots \\ &= q\left((q^2-1)\frac{D^3}{3!} + (q^4-1)\frac{D^5}{5!} + \cdots\right). \end{aligned} \tag{63}$$

Since $q>1$ and $D>0$ all of the terms in the Taylor series are positive. It follows that

$$\sinh qD - q\sinh D > 0. \tag{64}$$

This inequality can be written as

$$1 < \frac{\sinh qD}{q\sinh D} \tag{65}$$

or

$$1 < \frac{\sinh^2 qD}{q^2 \sinh^2 D}. \tag{66}$$

This inequality can be further written as

$$\frac{q^2}{\sinh^2 qD} - \frac{1}{\sinh^2 D} < 0. \tag{67}$$

By rewriting (62), the terms in (67) appear in the derivative of S with respect to D:

$$\frac{dS}{dD} = kND\left(\frac{q^2}{\sinh^2 qD} - \frac{1}{\sinh^2 D}\right). \tag{68}$$

Since $D > 0$, this implies that $\frac{dS}{dD} < 0$.

Since $\frac{dS}{dD} < 0$, S can be increased by lowering D, or

$$\sup_{D>0} S(D) = \lim_{D\to 0} S(D) = kN \ln q. \tag{69}$$

Since $S(D)$ is an even function, this result can be extended to all values of $D \neq 0$: $kN \ln q$ is an upper bound for $S(D)$.

REFERENCES

[1] J. C. Willems, *Dissipative dynamical systems, Part* I: *General theory*, Arch. Ration. Mech. Anal., 45 (1972), pp. 321–351.

[2] I. Mayergoyz, *Mathematical Models of Hysteresis and Their Applications*, Elsevier Academic Press, Amsterdam, Boston, 2003.

[3] R. B. Gorbet, K. A. Morris, and D. W. L. Wang, *Passivity-based stability and control of hysteresis in smart actuators*, IEEE Trans. Control Systems Tech., 9 (2001), pp. 5–16.

[4] D. Hughes and J. T. Wen, *Preisach modeling of piezoceramic and shape memory alloy hysteresis*, Smart Materials and Structures, 6 (1997), pp. 287–300.

[5] X. Tan, J. S. Baras, and P. S. Krishnaprasad, *A dynamic model for magnetostrictive hysteresis*, Proceedings of the American Control Conference, Vol. 2, 2003, pp. 1074–1079.

[6] X. Tan and J. S. Baras, *Modeling and control of hysteresis in magnetostrictive actuators*, Automatica, 40 (2004), pp. 1469–1480.

[7] R. C. Smith, M. J. Dapino, and S. Seelecke, *Free energy model for hysteresis in magnetostrictive transducers*, J. Appl. Phys., 93 (2003), pp. 458–466.

[8] R. B. Gorbet and K. A. Morris, *Closed-loop position control of preisach hystereses*, J. Intell. Material Systems Struct., 14 (2003), pp. 483–495.

[9] W. M. Haddad, V. Chellaboina, and J. Oh, *Linear controller analysis and design for systems with input hystereses nonlinearities*, J. Franklin Inst., 340 (2003), pp. 371–390.

[10] M. Vidyasagar, *Nonlinear Systems Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1993.

[11] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-output Properties*, Academic Press, New York, 1975.

[12] M. W. Zemansky and R. H. Dittman, *Heat and Thermodynamics*, McGraw–Hill, New York, Montreal, 1981.

[13] F. Reif, *Fundamentals of Statistical and Thermal Physics*, McGraw–Hill, New York, 1965.

[14] R. B. Gorbet, D. W. L. Wang, and K. A. Morris, *Preisach model identification of a two-wire SMA actuator*, Proceedings of the IEEE International Conference on Robotics and Automation, Vol. 3, 1998, pp. 2161–2167.

[15] R. C. Smith, *Smart Material Systems: Model Development*, Frontiers Appl. Math. 32, SIAM, Philadelphia, 2005.

[16] S. Valadkhan, K. A. Morris, and A. Khajepour, *A review and comparison of hysteresis models for magnetostrictive materials*, J. Intell. Material Systems Struct., to appear.
[17] R. B. Gorbet, K. A. Morris, and D. W. L. Wang, *Control of hysteretic systems: A state-space approach*, in Learning, Control, and Hybrid Systems, Y. Yamamoto, S. Hara, B. A. Francis, and M. Vidyasagar, eds., Springer-Verlag, London, 1999, pp. 432–451.
[18] R. C. Smith and M. J. Dapino, *A homogenized energy model for the direct magneto-mechanical effect*, IEEE Trans. Magnetics, 42 (2006), pp. 1944–1957.

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 687–713

GUIDED MODES IN PERIODIC SLABS: EXISTENCE AND NONEXISTENCE*

STEPHEN SHIPMAN† AND DARKO VOLKOV‡

Abstract. For homogeneous lossless three-dimensional periodic slabs of fixed arbitrary geometry, we characterize guided modes by means of the eigenvalues associated with a variational formulation. We treat robust modes, which exist for frequencies and wavevectors that admit no propagating Bragg harmonics and therefore persist under perturbations, as well as nonrobust modes, which can disappear under perturbations due to radiation loss. We prove the nonexistence of guided modes, both robust and nonrobust, in "inverse" structures, for which the celerity inside the slab is less than the celerity of the surrounding medium. The result is contingent upon a restriction on the width of the slab but is otherwise independent of its geometry.

Key words. guided mode, periodic slab, photonic crystal, nonexistence in inverse crystals

AMS subject classifications. 78A50, 78M30, 35P15

DOI. 10.1137/050647189

1. Introduction.

The subject of our investigation is the existence and nonexistence of linear scalar waves guided by periodically structured lossless material slabs (Figure 1). These *guided modes* occur in linear acoustic theory and, in the two-dimensional reduction, in which the structure is invariant in one of the two directions of periodicity, they describe guided polarized electromagnetic fields. Guided modes are characterized by their frequency and Bloch wavevector in the plane of periodicity, and they decay exponentially with distance away from the slab.

We distinguish between two types of guided mode. Those of the first type cannot be destroyed by radiation losses under perturbation because they possess a frequency and wavevector for which no Bragg, or Fourier, harmonics propagate away from the slab (they are all evanescent); we call these *robust guided modes*. Those possessing frequency and wavevector for which some propagating Bragg harmonics exist can be destroyed by radiation loss by "coupling" to these harmonics under perturbation of the structure, frequency, or wavevector. These *nonrobust guided modes* are known to be connected with anomalous scattering behavior in the vicinity of the frequency and wavenumber of the mode.

It is recognized that guided modes as well as the transmission anomalies associated with them will be useful in the design of photonic devices. These phenomena appear in many different photonic structures, and there is a large body of literature devoted to them. We mention just a few references. Anomalous transmission is typically characterized by sharp dips and peaks in the transmission coefficient. An in-depth computational analysis of their relation to leaky modes for slabs that are invariant in the transverse direction is given in Tikhodeev et al. [1]. Explicit asymptotic formulas for very general geometry for some types of perturbations have been calculated by Shipman and Venakides [2]. The connection between transmission enhancement and

*Received by the editors December 9, 2005; accepted for publication (in revised form) November 3, 2006; published electronically March 2, 2007.
http://www.siam.org/journals/siap/67-3/64718.html
†Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803 (shipman@math.lsu.edu).
‡Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609 (darko@wpi.edu).

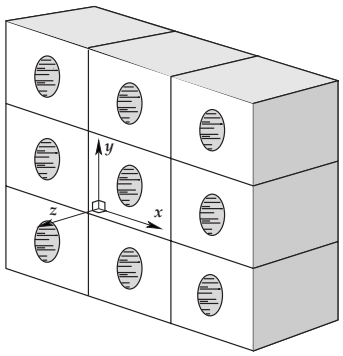


FIG. 1. *A slab structure periodic in the x and y directions and finite in the z direction.*

particular types of guided mode on metal films called "surface plasmons" has been studied in a series of papers by several authors; see [3] and [4], for example. An important class of guided modes that we do not treat here consists of those in optical fibers or periodic pillars (see [5], for example). Our present study focuses on the existence and nonexistence of guided modes in lossless dielectric slabs.

The existence of guided modes can be proved using variational principles. Bonnet-Bendhia and Starling [6] treat two-dimensional electromagnetic structures consisting of lossless penetrable and conducting components in which the dielectric coefficient is essentially an arbitrary function. The treatment of nonrobust modes is more delicate because establishing their existence requires proving the vanishing of the propagating Bragg harmonics. The frequencies of these modes, for a given wavevector, are called (as in [6]) *singular frequencies* of the problem of scattering, or diffraction, of plane waves by the slab.

In our study, we consider three-dimensional homogeneous dielectric structures. In this case we are able to prove a nonexistence theorem for *inverse structures*. An inverse structure is one for which the speed of waves, or the celerity, is higher inside the structure than in the surrounding medium. This result is easily understood through the following example. It is simple to calculate fields that are totally internally reflected within an infinite pane of glass surrounded by air. However, if the roles of the air and the glass are switched, such fields no longer exist. A similar result is expected for slabs with more general geometry. This is the content of Theorem 4.1. The result is subject to a restriction on the width of the slab, which depends on the frequency and wavevector; we do not know if this restriction is necessary or if it is only a artifact of our method of proof.

For the existence theory, we include complete proofs in the appendix (section 6) to make the work coherent and self-contained and in order to set the context and notation for the proof of the nonexistence result.

The governing equation of dynamics is the linear wave equation arising in small-amplitude acoustic theory:

$$\varepsilon \frac{\partial^2}{\partial t^2} w(x,y,z,t) = \nabla \cdot \frac{1}{\mu} \nabla w(x,y,z,t). \tag{1}$$

The positive material parameters ε and μ depend in general on the position within the slab but are constant outside of the slab. We will restrict our analysis to slabs in which these parameters are constant inside. The spatial factor $\tilde{u}(x,y,z)$ of a time-harmonic

solution

$$w(x,y,z,t) = \tilde{u}(x,y,z)\, e^{-i\omega t}$$

is described by the Helmholtz equation

$$\nabla\cdot \frac{1}{\mu}\nabla\tilde{u}(x,y,z) + \varepsilon\omega^2\, \tilde{u}(x,y,z) \;=\; 0. \tag{2}$$

We are interested in solutions of the Helmholtz equation that are of the pseudoperiodic form

$$\tilde{u}(x,y,z) = u(x,y,z)\, e^{i(\kappa_1 x+\kappa_2 y)}, \quad u \text{ periodic in } x \text{ and } y,$$

in which $u(x,y,z)$ has the same periods as the guiding slab structure. The vector $\boldsymbol{\kappa} = \langle \kappa_1, \kappa_2, 0\rangle$ is known as the *Bloch wavevector*, and the field $\tilde{u}$ is called a Bloch wave. Such a solution to the Helmholtz equation gives rise to a solution of the linear wave equation when multiplied by a harmonic factor in t:

$$w(x,y,z,t) = u(x,y,z)e^{i(\kappa_1 x+\kappa_2 y-\omega t)}.$$

This solution is a plane wave traveling in the direction of the vector $\boldsymbol{\kappa}$ with wave number $|\boldsymbol{\kappa}| = \sqrt{\kappa_1^2+\kappa_2^2}$, frequency ω, and speed $\omega/|\boldsymbol{\kappa}|$, modulated periodically through multiplication by the factor $u(x,y,z)$.

Fundamental to the structure of Bloch waves is their decomposition in the x and y variables into Fourier harmonics, often called Bragg harmonics, in the regions away from the slab ($|z|$ sufficiently large):

$$u(x,y,z) = \sum_{m,n=-\infty}^{\infty} (c_{mn}^{+}e^{\nu_{mn}z} + c_{mn}^{-}e^{-\nu_{mn}z})e^{i(mx+ny)}.$$

Each element of the sum is a separable solution to the Helmholtz equation, and the coefficients c_{mn}^{+} and c_{mn}^{-} differ from one side of the slab to the other. The exponents ν_{mn}, as explained in more detail below, are purely imaginary for a finite number of pairs (m,n), corresponding to the *propagating Fourier harmonics*. For all other pairs, assuming $\nu_{mn} \neq 0$ for all (m,n), this exponent is real, and boundedness of u requires that the coefficients of the exponentially growing components vanish. Thus these pairs correspond to the evanescent harmonics. Assuming then that u is bounded, we can say that *a guided mode is supported by the slab structure if the coefficients of all propagating Fourier harmonics vanish.*

The paper is organized as follows. In section 2, we formulate a precise definition of a guided mode in its strong and weak forms. We explain the relation between vanishing propagating Fourier modes and absence of energy loss by radiation. In section 3, we discuss the existence of sequences of material constants, depending on the geometry of the structure, the wavevector, and the frequency, that admit guided modes in the regime of no radiating Fourier harmonics. We also prove the existence of sequences of material constants for structures symmetric about a plane, depending on the frequency and wave number along the plane of symmetry, for which guided modes that travel parallel to the plane of symmetry exist. All of the proofs are deferred to the appendix. In section 4, we prove that guided modes cannot exist in "inverse" slab structures; specifically, we show that, under a suitable restriction on their width, slabs with constant μ and ϵ whose value interior to the slab is less than its value in the exterior, never admit guided modes. In section 5, we show a few numerical computations of nonrobust guided modes.

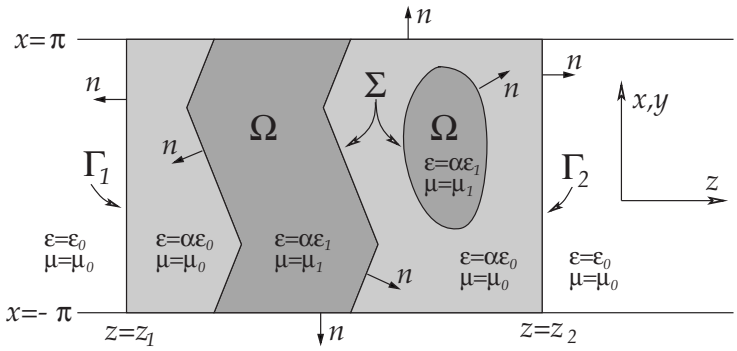


FIG. 2. *A two-dimensional depiction of one period of a possible augmented three-dimensional slab structure.*

2. Mathematical formulation of guided modes. A variational description of guided modes requires truncation of the domain in the z direction (directed away from the slab) and the introduction of an auxiliary parameter α, serving as the eigenvalue, by which the coefficient ϵ is multiplied within the truncated domain (Figure 2). The monotonicity of certain eigenvalue sequences $\alpha_j(\epsilon_1)$ with respect to the interior constant ϵ_1 is utilized in the proof of nonexistence of modes in section 4.

Let $\tilde{\Omega}$ denote a domain in $\mathbb{R}^3$ with C^2 boundary $\partial\tilde{\Omega}$ that is bounded in the z direction and 2π-periodic in the x and y directions (Figure 2). This means that

i. there are numbers $z_1 < z_2$ such that

$$z_1 < \inf\{z : (x, y, z) \in \tilde{\Omega}\} < \sup\{z : (x, y, z) \in \tilde{\Omega}\} < z_2;$$

ii. if $(x, y, z) \in \tilde{\Omega}$, then $(x + 2\pi j, y, z) \in \tilde{\Omega}$ and $(x, y + 2\pi j, z) \in \tilde{\Omega}$ for each integer j.

Let $\mathcal{S}$ denote the infinite square cylinder containing one period of $\tilde{\Omega}$,

$$\mathcal{S} = \{(x, y, z) : -\pi < x < \pi,\ -\pi < y < \pi\},$$

and denote by Ω the part of $\tilde{\Omega}$ contained in $\mathcal{S}$, constituting one period of $\tilde{\Omega}$,

$$\Omega = \tilde{\Omega} \cap \mathcal{S} = \{(x, y, z) \in \tilde{\Omega} : -\pi < x < \pi,\ -\pi < y < \pi\},$$

and by Σ the part of $\partial\tilde{\Omega}$ contained in $\mathcal{S}$,

$$\Sigma = \partial\tilde{\Omega} \cap \mathcal{S}.$$

The boundary $\partial\Omega$ of Ω includes Σ and possibly parts of the boundary $\partial\mathcal{S}$ of $\mathcal{S}$. Denote by $\mathcal{R}$ the part of $\mathcal{S}$ between $z = z_1$ and $z = z_2$,

$$\mathcal{R} = \{-\pi < x < \pi,\ -\pi < y < \pi,\ z_1 < z < z_2\},$$

and by $\Gamma = \Gamma_1 \cup \Gamma_2$ the square parts of the boundary of $\mathcal{R}$ parallel to the xy-plane:

$$\Gamma_1 = \{(x, y, z) : -\pi < x < \pi,\ -\pi < y < \pi,\ z = z_1\},$$
$$\Gamma_2 = \{(x, y, z) : -\pi < x < \pi,\ -\pi < y < \pi,\ z = z_2\}.$$

We fix outward-pointing normal vectors n to all of the surfaces, as shown in Figure 2. Let the piecewise constant functions ε and μ be defined by

$$\varepsilon(r) = \begin{cases} \alpha\epsilon_1, & r \in \Omega, \\ \alpha\epsilon_0, & r \in \mathcal{R} \setminus \Omega, \\ \epsilon_0, & r \in \mathcal{S} \setminus \mathcal{R}, \end{cases} \qquad \mu(r) = \begin{cases} \mu_1, & r \in \Omega, \\ \mu_0, & r \in \mathcal{S} \setminus \Omega, \end{cases} \tag{3}$$

in which ϵ_0, ϵ_1, α, μ_0, and μ_1 are fixed positive numbers.

We will be considering guided modes in the augmented structure consisting of the flat slab in $\mathbb{R}^3$ filling the space between the planes $z = z_1$ and $z = z_2$ with the periodic structure $\tilde{\Omega}$ embedded in it. For $\alpha = 1$, this augmented structure reduces to $\tilde{\Omega}$ itself. We denote the augmented structure by $\tilde{\Omega}_{\text{aug}}$:

$$\tilde{\Omega}_{\text{aug}} = \text{the augmented structure in Figure 2 extended periodically to } \mathbb{R}^3.$$

When referring to $\tilde{\Omega}_{\text{aug}}$, the material constants (3), repeated periodically, are tacitly assumed.

Given a frequency ω and Bloch wavevector $\boldsymbol{\kappa} = \langle \kappa_1, \kappa_2, 0 \rangle$, we seek solutions $\tilde{u} = \tilde{u}(x, y, z)$ of the Helmholtz equation (2) in $\mathcal{S}$ with $\boldsymbol{\kappa}$-pseudoperiodic boundary conditions in x and y, that is,

$$\tilde{u}(\pi, y, z) = e^{2\pi i \kappa_1} \tilde{u}(-\pi, y, z), \quad \partial_n \tilde{u}(\pi, y, z) = -e^{2\pi i \kappa_1} \partial_n \tilde{u}(-\pi, y, z), \tag{4}$$

$$\tilde{u}(x, \pi, z) = e^{2\pi i \kappa_2} \tilde{u}(x, -\pi, z), \quad \partial_n \tilde{u}(x, \pi, z) = -e^{2\pi i \kappa_2} \partial_n \tilde{u}(x, -\pi, z). \tag{5}$$

The minus signs arise because the normal vector n to $\partial\mathcal{S}$ is always taken to point out from $\mathcal{S}$. Such a solution can be extended to a $\boldsymbol{\kappa}$-pseudoperiodic solution in $\mathbb{R}^3$, that is, retaining $\tilde{u}$ to denote this extension,

$$\tilde{u}(x, y, z) = e^{i(\kappa_1 x + \kappa_2 y)} u(x, y, z),$$

in which $u(x, y, z)$ is 2π-periodic in x and y. It is convenient to work with the function u, which has periodic boundary conditions in $\mathcal{S}$:

$$u(\pi, y, z) = u(-\pi, y, z), \quad \partial_n u(\pi, y, z) = -\partial_n u(-\pi, y, z), \tag{6}$$

$$u(x, \pi, z) = u(x, -\pi, z), \quad \partial_n u(x, \pi, z) = -\partial_n u(x, -\pi, z). \tag{7}$$

The Helmholtz equation (2) for $\tilde{u}$ is equivalent to the following modified equation for the periodic factor u:

$$(\nabla + i\boldsymbol{\kappa}) \cdot \frac{1}{\mu} (\nabla + i\boldsymbol{\kappa}) u + \varepsilon\omega^2 u = 0. \tag{8}$$

In the precise formulation of a guided mode below, Condition 2.2, we make clear the implied behavior of a solution to this equation at the surfaces of discontinuity of ε and μ, namely, Σ and Γ.

We take $\boldsymbol{\kappa} = \langle \kappa_1, \kappa_2, 0 \rangle$ to lie in the first symmetric Brillouin zone pertaining to our structure, which is 2π-periodic in x and y, that is,

$$-1/2 \le \kappa_1 < 1/2 \quad \text{and} \quad -1/2 \le \kappa_2 < 1/2.$$

In the intervals $(-\infty, z_1)$ and (z_2, ∞), every periodic solution u of (8) is equal to a superposition of Fourier harmonics:

$$\begin{aligned} u(x, y, z) &= \sum_{m,n=-\infty}^{\infty} (a_{mn}^{+} e^{\nu_{mn} z} + a_{mn}^{-} e^{-\nu_{mn} z}) e^{i(mx+ny)}, \quad z < z_1, \\ u(x, y, z) &= \sum_{m,n=-\infty}^{\infty} (b_{mn}^{+} e^{\nu_{mn} z} + b_{mn}^{-} e^{-\nu_{mn} z}) e^{i(mx+ny)}, \quad z > z_2, \end{aligned} \tag{9}$$

in which

$$\nu_{mn}^2 = -\epsilon_0\mu_0\omega^2 + (m+\kappa_1)^2 + (n+\kappa_2)^2, \tag{10}$$

provided that $\nu_{mn}^2 \neq 0$ for all integer pairs (m,n). If $\nu_{mn}^2 = 0$ for some pair (m,n), then its contribution to the sum (9) must be replaced by

$$\begin{array}{ll} (a_{mn}^+ + a_{mn}^- z)e^{i(mx+ny)}, & z < z_1, \\ (b_{mn}^+ z + b_{mn}^-)e^{i(mx+ny)}, & z > z_2. \end{array} \tag{11}$$

For a finite number of pairs (m,n) we have $\nu_{mn}^2 < 0$, and we take $\operatorname{Im}\nu_{mn} > 0$; these correspond to the harmonics in (9) whose two terms have constant modulus and oscillate as functions of z. We denote this set of *propagating harmonics* by $\mathcal{P}$:

$$\mathcal{P} = \left\{(m,n) \in \mathbb{Z}^2 : \nu_{mn}^2 < 0\right\} \qquad \text{(propagating Fourier harmonics)}. \tag{12}$$

We call the harmonics of the form (11), for which $\nu_{mn}^2 = 0$, the *linear harmonics*. We denote the union of the linear and propagating harmonics by $\tilde{\mathcal{P}}$:

$$\tilde{\mathcal{P}} = \left\{(m,n) \in \mathbb{Z}^2 : \nu_{mn}^2 \leq 0\right\} \quad \text{(linear and propagating Fourier harmonics)}. \tag{13}$$

For a generic set of parameters ϵ_0, μ_0, α, $\boldsymbol{\kappa}$, and ω, there are no linear harmonics; that is, $\tilde{\mathcal{P}} = \mathcal{P}$. For all pairs such that $\nu_{mn}^2 > 0$ we take $\operatorname{Re}\nu_{mn} > 0$; these correspond to the exponential harmonics. We require the solution u to be bounded, so that

$$a_{mn}^- = 0 \quad \text{and} \quad b_{mn}^+ = 0 \quad \text{for all linear and exponential harmonics}, \tag{14}$$

to exclude unbounded growth as $|z| \to \infty$. The harmonics that are exponentially decaying as $|z| \to \infty$ are called the decaying harmonics, or *evanescent harmonics*.

The energy conservation law holds for solutions of the Helmholtz equation. This means that the the time-averaged energy flux through in $\mathcal{S}$ through planes parallel to the xy-plane is independent of z. Only the propagating harmonics contribute to this energy, and equating its values through Γ_1 and Γ_2 gives

$$\sum_{(m,n)\in\mathcal{P}} \nu_{mn}\left(|a_{mn}^+|^2 - |a_{mn}^-|^2\right) = \sum_{(m,n)\in\mathcal{P}} \nu_{mn}\left(|b_{mn}^+|^2 - |b_{mn}^-|^2\right). \tag{15}$$

A *guided mode* u, which we will define precisely in Definition 2.3, is a nonzero solution of the Helmholtz equation with exponential decay as $|z| \to \infty$. If u satisfies the condition (14) of boundedness as well as the vanishing of the linear and propagating harmonics, that is, $a_{mn}^+ = a_{mn}^- = b_{mn}^+ = b_{mn}^- = 0$ for all $(m,n) \in \tilde{\mathcal{P}}$, then u has exponential decay as $|z| \to \infty$, and its periodic extension to $\mathbb{R}^3$ is a guided mode. In the generic case that there are no linear harmonics, that is, $\nu_{mn}^2 \neq 0$ for each (m,n), we may characterize guided modes by the condition that

$$a_{mn}^- = 0 \quad \text{and} \quad b_{mn}^+ = 0 \qquad \text{for all } (m,n) \qquad (\text{if } \tilde{\mathcal{P}} = \mathcal{P}). \tag{16}$$

Indeed, (16) and (15) together imply in this case the *vanishing of all propagating harmonics as well as all exponentially growing harmonics.* In the general case, in which linear harmonics may exist, we must augment this condition to exclude these harmonics explicitly:

$$\begin{array}{lll} a_{mn}^- = 0 \quad \text{and} \quad b_{mn}^+ = 0 & \text{for all} & (m,n), \\ a_{mn}^+ = 0 \quad \text{and} \quad b_{mn}^- = 0 & \text{if} & \nu_{mn}^2 = 0. \end{array} \tag{17}$$

The reason for treating the case of no linear harmonics specially is that the condition (16) is simple. In fact, it is equivalent to the condition that u obey the following Dirichlet-to-Neumann map on Γ defined in terms of the Fourier coefficients of u restricted to Γ_1 and Γ_2:

(18)
$$\begin{array}{lllll} \text{If} & u|_{\Gamma_1} = \sum a_{mn} e^{i(mx+ny)} & \text{and} & u|_{\Gamma_2} = \sum b_{mn} e^{i(mx+ny)}, \\ \text{then} & \partial_n u|_{\Gamma_1} = -\sum \nu_{mn} a_{mn} e^{i(mx+ny)} & \text{and} & \partial_n u|_{\Gamma_2} = -\sum \nu_{mn} b_{mn} e^{i(mx+ny)}, \end{array}$$

in which the sum is over all integer pairs (m, n). This motivates the definition of an operator B on functions defined on Γ, which will enable us to give a precise formulation of a guided mode. For the differential, or strong, formulation, this Dirichlet-to-Neumann map needs to be defined only for twice differentiable functions in $\mathcal{R} \setminus \Sigma$ whose first derivative is continuous up to Γ. We give a refined definition that will accommodate the variational, or weak, formulation, which includes functions in $H^1(\mathcal{R})$, that is, functions belonging to $L^2(\mathcal{R})$ that possess weak first derivatives also belonging to $L^2(\mathcal{R})$. For such functions, a restriction to Γ is well defined as a function in the fractional Sobolev space $H^{1/2}(\Gamma)$, but a normal derivative is not well defined. The Dirichlet-to-Neumann map is replaced by a bounded operator B from $H^{1/2}(\Gamma)$ to its dual space $H^{-1/2}(\Gamma)$, which coincides with the Dirichlet-to-Neumann map (18) when restricted to twice differentiable functions with continuous derivatives up to Γ.

DEFINITION 2.1 (Dirichlet-to-Neumann map B). *Let $f \in H^{1/2}(\Gamma)$ be given, and represent f as $f = (f^1, f^2)$ according to the decomposition $H^{1/2}(\Gamma) = H^{1/2}(\Gamma_1) \oplus H^{1/2}(\Gamma_2)$. Set $\hat{f}_{mn} = (\hat{f}^1_{mn}, \hat{f}^2_{mn})$, where $\hat{f}^{1,2}_{mn}$ are the Fourier coefficients of $f^{1,2}$. Note that $\nu_{mn}\hat{f}_{mn} = (\nu_{mn}\hat{f}^1_{mn}, \nu_{mn}\hat{f}^2_{mn}) \in H^{-1/2}(\Gamma_1) \oplus H^{-1/2}(\Gamma_2) = H^{-1/2}(\Gamma)$, and define Bf through its Fourier coefficients by setting*

$$(\widehat{Bf})_{mn} = \nu_{mn}\hat{f}_{mn} \qquad \textit{(definition of } B\textit{).} \tag{19}$$

We use integral notation to denote the action of the function $Bf \in H^{-1/2}(\Gamma)$ on $g \in H^{1/2}(\Gamma)$, and this action is concretely expressed through the Fourier coefficients of f and g:

$$\int_\Gamma (Bf)g = \sum_{m,n=-\infty}^{\infty} \nu_{mn}(\hat{f}^1_{mn}\hat{g}^1_{mn} + \hat{f}^2_{mn}\hat{g}^2_{mn}). \tag{20}$$

The action of Bf restricted to Γ_j, for $j = 1, 2$, is given by

$$\int_{\Gamma_j} (Bf)g = \sum_{m,n=-\infty}^{\infty} \nu_{mn}\hat{f}^j_{mn}\hat{g}^j_{mn}.$$

If all the ν_{mn} are positive, that is, if $\tilde{\mathcal{P}} = \emptyset$, then B is a positive operator; that is, for each $f \in H^{1/2}(\mathcal{R})$, $\int_\Gamma (Bf)\bar{f} > 0$.

We are now ready to state the condition that allows a precise definition of a guided mode. Condition 2.2 makes precise the behavior of a solution to the Helmholtz equation (8) at the surfaces of discontinuity of the functions ε and μ (see (3)) and enforces the exponential decay through the Dirichlet-to-Neumann operator B. The condition (17) necessary when linear harmonics are present is stated separately for that case (see (21)).

CONDITION 2.2 (strong condition for a guided mode). *Let u be a twice differentiable function in $\mathcal{S} \setminus (\Sigma \cup \Gamma)$ with continuous value and first derivative up to $\partial\mathcal{S}$, Σ,*

and Γ. *Denote by* $\partial_n u_\pm$ *the values of the normal derivative of* u *on* Σ *and* Γ, *where the* $+$*-sign refers to the side toward the direction of the normal vector* n. *If* $\nu_{mn}^2 \neq 0$ *for all* (m,n), *then* u *satisfies the strong condition for a guided mode, provided that*

i. $(\nabla + i\boldsymbol{\kappa})^2 u + \mu_0\epsilon_0\omega^2 u = 0$ *in* $\mathcal{S}\setminus\mathcal{R}$,
ii. $(\nabla + i\boldsymbol{\kappa})^2 u + \alpha\mu_0\epsilon_0\omega^2 u = 0$ *in* $\mathcal{S}\setminus\Omega$,
iii. $(\nabla + i\boldsymbol{\kappa})^2 u + \alpha\mu_1\epsilon_1\omega^2 u = 0$ *in* Ω,
iv. u *is continuous in* $\mathcal{S}$,
v. $\partial_n u_+ = \partial_n u_- = -Bu$ *on* Γ,
vi. $\mu_1(\partial_n u_+ + (i\boldsymbol{\kappa}\cdot n)u) = \mu_0(\partial_n u_- + (i\boldsymbol{\kappa}\cdot n)u)$ *on* Σ,
vii. $u(-\pi,y,z) = u(\pi,y,z)$ *and* $\partial_n u(-\pi,y,z) = -\partial_n u(\pi,y,z)$,
viii. $u(x,-\pi,z) = u(x,\pi,z)$ *and* $\partial_n u(x,-\pi,z) = -\partial_n u(x,\pi,z)$.

If $\nu_{mn} = 0$ *for some* (m,n), *then for each such pair we require, in addition, that the corresponding Fourier coefficient of* u *be zero on* Γ*:*

$$(2\pi)^2(\widehat{u|_{\Gamma_j}})_{mn} = \int_{\Gamma_j} u(x,y,z)e^{-i(mx+ny)} = 0, \quad j = 1,2 \qquad ((m,n)\in\tilde{\mathcal{P}}\setminus\mathcal{P}). \tag{21}$$

Definition 2.3 (guided mode). *A* guided mode *in the augmented periodic slab structure* $\tilde{\Omega}_{aug}$ *is the pseudoperiodic extension to* $\mathbb{R}^3$ *of a function of the form*

$$u(x,y,z)e^{i(\kappa_1 x+\kappa_2 y-i\omega t)},$$

in which u *satisfies the strong Condition* 2.2.

It is possible to restrict analysis to the region $\mathcal{R}$, for if we omit the condition (i) and the first equality in (v), then a function in $\overline{\mathcal{R}}$ (the closure of $\mathcal{R}$) satisfying the remaining conditions can be extended in a unique way to $\mathcal{S}$ such that (i) is satisfied simply by declaring

$$\begin{aligned} u(x,y,z) &= \sum_{m,n=-\infty}^{\infty} a_{mn}e^{\nu_{mn}(z-z_1)}e^{i(mx+ny)}, \quad z\le z_1, \\ u(x,y,z) &= \sum_{m,n=-\infty}^{\infty} b_{mn}e^{-\nu_{mn}(z-z_2)}e^{i(mx+ny)}, \quad z\ge z_2, \end{aligned} \tag{22}$$

in which a_{mn} are the Fourier coefficients of $u|_{\Gamma_1}$ and b_{mn} are the Fourier coefficients of $u|_{\Gamma_2}$, for this function satisfies the condition $Bu = \partial_n u_+ = \partial_n u_-$ on Γ.

We will need a variational formulation for guided modes. The appropriate function space is the periodic subspace $H^1_{\rm per}(\mathcal{R})$ of the Sobolev space $H^1(\mathcal{R})$ of functions in $L^2(\mathcal{R})$ with weak gradients in $L^2(\mathcal{R})$: $H^1_{\rm per}(\mathcal{R})$ is the subspace of functions $f\in H^1(\mathcal{R})$ satisfying $f(-\pi,y,z) = f(\pi,y,z)$ and $f(x,-\pi,z) = f(x,\pi,z)$, where the boundary values of f are well defined by a bounded trace operator to $H^{1/2}(\partial\mathcal{S})$. $H^1_{\rm per}(\mathcal{R})$ is a Hilbert space, retaining the same inner product as $H^1(\mathcal{R})$:

$$(u,v)_{H^1(\mathcal{R})} = \int_{\mathcal{R}} (u\bar{v} + \nabla u\nabla\bar{v}).$$

In referring to the trace of f on Γ, we will be more precise and denote the trace operator by $T: H^1(\mathcal{R})\to H^{1/2}(\Gamma)$, so that the restriction of f to Γ is denoted by Tf.

Condition 2.4 (weak condition for a guided mode, first form). *A function* $u\in H^1_{per}(\mathcal{R})$ *satisfies the weak condition for a guided mode, provided*

$$\int_{\mathcal{R}} \frac{1}{\mu}(\nabla + i\boldsymbol{\kappa})u\cdot(\nabla - i\boldsymbol{\kappa})\bar{v} + \frac{1}{\mu_0}\int_\Gamma (BTu)(T\bar{v}) - \int_{\mathcal{R}} \varepsilon\omega^2 u\bar{v} = 0 \quad \textit{for all} \quad v\in H^1_{per}(\mathcal{R}). \tag{23}$$

In case $\nu_{mn}^2 = 0$ *for any pair* (m,n), *it is required additionally that* $(\widehat{Tf})_{mn} = 0$.

The sesquilinear form in Condition 2.4 is conjugate-symmetric in $H^1_{\text{per}}(\mathcal{R})$ if and only if $\mathcal{P} = \emptyset$. We introduce a subspace X in which it is always conjugate-symmetric, and, in fact, positive, namely, the subspace of functions whose traces on Γ have vanishing Fourier coefficients for $(m,n) \in \tilde{\mathcal{P}}$.

$$(\widehat{Tf})_{mn} = 0 \quad \text{for} \quad (m,n) \in \tilde{\mathcal{P}} \qquad \text{(defining condition for } f \in X\text{)},$$

or, equivalently,

$$\int_{\Gamma_1} (Tf) e^{-i(mx+ny)} = \int_{\Gamma_2} (Tf) e^{-i(mx+ny)} = 0 \quad \text{for} \quad (m,n) \in \tilde{\mathcal{P}} \quad \text{(condition for } f \in X\text{)}.$$

X is closed under the norm of $H^1(\mathcal{R})$. In order to exclude linear and propagating Fourier harmonics from the extension to all of $\mathcal{S}$ of a function f in X, which has well defined normal derivatives on Γ, *it must also be demanded that the normal derivative have vanishing Fourier coefficients for* $(m,n) \in \tilde{\mathcal{P}}$. We give an alternate variational formulation of guided modes in the weak Condition 2.5.

CONDITION 2.5 (weak condition for a guided mode, second form). *A function* $u \in H^1_{per}(\mathcal{R})$ *that possesses a normal derivative on* Γ *satisfies the weak condition for a guided mode, provided that* $u \in X$ *and*

i. $\int_{\mathcal{R}} \frac{1}{\mu} (\nabla + i\boldsymbol{\kappa}) u \cdot (\nabla - i\boldsymbol{\kappa}) \bar{v} + \frac{1}{\mu_0} \int_\Gamma (BTu)(T\bar{v}) - \int_{\mathcal{R}} \varepsilon\omega^2 u\bar{v} = 0$ *for all* $v \in X$,

ii. $(\widehat{\partial_n u|_\Gamma})_{mn} = 0$ *for all* $(m,n) \in \tilde{\mathcal{P}}$.

We prove in Theorem 2.7 that Conditions 2.2, 2.4, and 2.5 are all equivalent. In particular, a function in $H^1_{\text{per}}(\mathcal{R})$ that satisfies Condition 2.4 is in fact regular and satisfies the other two conditions, and Condition 2.5.i actually implies the existence of a normal derivative on Γ.

In section 3, we will show the existence of a sequence of relations between α and ϵ_1, for each choice of ω and $\boldsymbol{\kappa}$, that describe all of the pairs (α, ϵ_1) that support a solution of Condition 2.5.i. Because these solutions are in X, the coefficients in their Fourier expansion for all $(m,n) \in \mathcal{P}$ (see (9)) satisfy

$$|a^+_{mn}| - |a^-_{mn}| = 0 \quad \text{and} \quad |b^+_{mn}| - |b^-_{mn}| = 0,$$

implying the vanishing of energy flux in the z direction.

Condition 2.5.ii indicates that guided modes typically do not exist in the $(\boldsymbol{\kappa}, \omega)$-regime of propagating or linear harmonics ($\tilde{\mathcal{P}} \neq \emptyset$) due to this extra condition that each of these harmonics must satisfy. The vanishing of this finite number of harmonics must be accomplished through the tuning of other parameters of the structure. In particular, if the structure is symmetric about the yz-plane and $\boldsymbol{\kappa} = (0, \kappa_2, 0)$, or it is symmetric about the xz-plane and $\boldsymbol{\kappa} = (\kappa_1, 0, 0)$, then the functions satisfying Condition 2.5.i are symmetric or antisymmetric. We focus on structures with symmetry about the yz-plane. Ω is symmetric about the yz-plane if

$$(x,y,z) \in \Omega \implies (-x,y,z) \in \Omega.$$

In this case, the antisymmetric solutions to (i) also satisfy (ii) for all $(m,n) \in \tilde{\mathcal{P}}$ with m even. Thus, if there is only one propagating mode $(0,0)$ and the rest are evanescent, then Condition 2.5 is satisfied in full, and the solutions therefore represent *nonrobust guided modes* traveling parallel to the plane of symmetry. These modes are nonrobust because, under a general perturbation of κ_1 or the structure itself, the $(0,0)$ harmonic, which is not evanescent, is no longer guaranteed to vanish.

We make the formulation for an antisymmetric nonrobust mode in a symmetric structure precise in Condition 2.6 and Theorem 2.7. For this, we introduce the orthogonally complementary subspaces X^{sym} and X^{ant} of X to treat the case that Ω is symmetric about the yz-plane, $\kappa_1 = 0$, and $\tilde{\mathcal{P}} \neq \emptyset$:

$$X^{\mathrm{sym}} = \{v \in X : v(x,y,z) = v(x,y,z) \text{ a.e. in } \mathcal{R}\},$$
$$X^{\mathrm{ant}} = \{v \in X : v(x,y,z) = v(-x,y,z) \text{ a.e. in } \mathcal{R}\}.$$

It is straightforward to verify that X^{sym} and X^{ant} are orthogonal in the usual H^1 and L^2 inner products on X and with respect to the sesquilinear form on the left-hand side of Condition 2.6.i:

$$X = X^{\mathrm{sym}} \oplus X^{\mathrm{ant}}.$$

CONDITION 2.6 (weak condition for a nonrobust guided mode). *Suppose that Ω is symmetric about the yz-plane and that $\kappa_1 = 0$. A function $u \in H^1_{per}(\mathcal{R})$ satisfies the weak condition for an antisymmetric nonrobust mode, provided $u \in X^{ant}$ and*

i. $\int_{\mathcal{R}} \frac{1}{\mu}(\nabla + i\boldsymbol{\kappa})u \cdot (\nabla - i\boldsymbol{\kappa})\bar{v} + \frac{1}{\mu_0}\int_\Gamma (BTu)(T\bar{v}) - \int_{\mathcal{R}} \varepsilon\omega^2 u\bar{v} = 0$ *for all* $v \in X^{ant}$,

ii. $\tilde{\mathcal{P}} \neq \emptyset$ *and* $(\partial_n \widehat{u|_\Gamma})_{mn} = 0$ *for all* $(m,n) \in \tilde{\mathcal{P}}$ *with* m *odd.*

THEOREM 2.7 (equivalence of strong and weak conditions).

i. *Let u satisfy Condition* 2.2. *Then the restriction of u to $\mathcal{R}$ is in $H^1_{per}(\mathcal{R})$ and satisfies Condition* 2.4.

ii. *Let u satisfy Condition* 2.4. *Then u can be extended to a twice differentiable function in $\mathcal{S} \setminus (\Sigma \cup \Gamma)$ with continuous value and first derivative up to $\partial\mathcal{S}$, Σ, and Γ. This extension satisfies Condition* 2.2, *and u is in X and satisfies Condition* 2.5.

iii. *If u satisfies Condition* 2.5, *then u satisfies Condition* 2.4.

iv. *If u satisfies Condition* 2.6, *then u satisfies Condition* 2.5 *(for $\kappa_1 = 0$).*

3. Existence of guided modes. The theoretical development presented in this section is in essence that followed in [6]. Nevertheless, we feel that complete proofs are necessary to ensure that consistency and mathematical rigor is observed. The proofs are given in the appendix (section 6).

Define the following sesquilinear forms in $H^1_{\mathrm{per}}(\mathcal{R})$:

$$A(u,v) = \int_{\mathcal{R}} \frac{1}{\mu}(\nabla + i\boldsymbol{\kappa})u \cdot (\nabla - i\boldsymbol{\kappa})\bar{v} + \frac{1}{\mu_0}\int_\Gamma (BTu)(T\bar{v}), \tag{24}$$

$$\ell(u,v) = \int_{\mathcal{R}\setminus\Omega} \epsilon_0\omega^2 u\bar{v} + \int_\Omega \epsilon_1\omega^2 u\bar{v}. \tag{25}$$

Notice that A depends on $\boldsymbol{\kappa}$, ω, ϵ_0, μ_0, and μ_1 (the dependence on ω and ϵ_0 is through B—see (19) and (10)), and ℓ depends on ω, ϵ_0, and ϵ_1; neither of them depends on α.

A function $u \in H^1_{\mathrm{per}}(\mathcal{R})$ satisfies the weak condition for a guided mode if and only if $A(u,v) - \alpha\ell(u,v) = 0$ for each $v \in H^1_{\mathrm{per}}(\mathcal{R})$. This is equivalent to the condition that u is an eigenfunction of the map $H^1_{\mathrm{per}}(\mathcal{R}) \to H^1_{\mathrm{per}}(\mathcal{R})^* :: u \mapsto \overline{A(u,\cdot)}$ (the asterisk denotes the dual space) with eigenvalue α in the sense that

$$A(u,\cdot) = \ell(\alpha u, \cdot).$$

If the set of propagating harmonics is empty ($\mathcal{P} = \emptyset$), then A is conjugate-symmetric; that is, $A(u,v) = \overline{A(v,u)}$ for all $u, v \in H^1_{\mathrm{per}}(\mathcal{R})$. Otherwise, it is not due

to the purely imaginary values of ν_{mn} in the definition (2.1) of B for $(m,n) \in \mathcal{P}$. In X, both A and ℓ are conjugate-symmetric, and therefore the eigenvalues α are real.

Define the Rayleigh quotient by

$$J(u) = \frac{A(u,u)}{\ell(u,u)} = \frac{\int_{\mathcal{R}} \frac{1}{\mu} |(\nabla + i\boldsymbol{\kappa})u|^2 + \frac{1}{\mu_0}\int_\Gamma (BTu)(T\bar{u})}{\epsilon_0\omega^2 \int_{\mathcal{R}\setminus\Omega} |u|^2 + \epsilon_1\omega^2 \int_\Omega |u|^2}. \tag{26}$$

Recall that the operator B depends on κ, ω, ϵ_0, and μ_0, but not on ϵ_1. The dependence of $J(u)$ on ϵ_1 comes only in the second term of ℓ. For an exposition of the role of the Rayleigh quotient in the theory of eigenvalues of elliptic operators, the reader may refer to Jost [7, section 8.5] or Gould [8, Chapter II]; a more brief discussion is found in Gilbarg and Trudinger [9, section 8.12].

THEOREM 3.1 (eigenvalue sequences). *There exists a sequence of real numbers (eigenvalues) $\{\alpha_j\}_{j=0}^\infty$ and functions (eigenfunctions) $\{\psi_j\}_{j=0}^\infty$ such that*

i. $0 < \alpha_0 \le \alpha_1 \le \cdots \le \alpha_j \le \cdots$,
ii. $\alpha_j \to \infty$ *as* $j \to \infty$,
iii. $A(\psi_j, v) = \alpha_j \ell(\psi_j, v)$ *for all* $v \in X$,
iv. *if* $A(\psi, v) = \alpha\ell(\psi, v)$ *for all* $v \in X$, *then there is an integer* j *such that* $\alpha = \alpha_j$ *and* $\psi \in span\{\psi_k : \alpha = \alpha_k\}$,
v. *the sequence* $\{\psi_j\}_{j=0}^\infty$ *is an orthonormal Hilbert-space basis for* $L^2(\mathcal{R}, \ell)$.

The eigenvalues and eigenfunctions arise from the process of successive minimization of the Rayleigh quotient:

$$\alpha_j = \inf_{u \in X_j, u \neq 0} J(u) = J(\psi_j), \quad \psi_j \in X_j,$$

in which

$$X_j = \{v \in X : A(\psi_k, v) = 0 \text{ for } k = 0, \ldots, j-1\}.$$

If Ω is symmetric about the yz-plane and $\kappa_1 = 0$, then $\{\psi_j\}_{j=0}^\infty$ is the union of two nondecreasing sequences $\{\psi_j^{sym}\}_{j=0}^\infty$ and $\{\psi_j^{ant}\}_{j=0}^\infty$ from X^{sym} and X^{ant}, respectively. We denote the associated sequences of eigenvalues by $\{\alpha_j^{sym}\}_{j=0}^\infty$ and $\{\alpha_j^{ant}\}_{j=0}^\infty$. The symmetric and antisymmetric eigenfunctions and associated eigenvalues arise from minimization of the Rayleigh quotient over X^{sym} and X^{ant}:

$$\alpha_j^{ant} = \inf_{u \in X_j^{ant}, u \neq 0} J(u) = J(\psi_j^{ant}), \quad \psi_j \in X_j^{ant},$$

in which

$$X_j^{ant} = \{v \in X^{ant} : A(\psi_k^{ant}, v) = 0 \text{ for } k = 0, \ldots, j-1\}.$$

LEMMA 3.2. *The eigenvalues α_j are continuous strictly decreasing functions of ϵ_1, and $\alpha_j \to 0$ as $\epsilon_1 \to \infty$. Similarly, the α_j are continuous strictly decreasing functions of μ_1, and $\alpha_j \to 0$ as $\mu_1 \to \infty$.*

Using this lemma, let us parse Theorem 3.1 in a form in which α is fixed, so that the functions ε and μ given by (3) outside of the domain Ω are fixed. For the sake of argument, let us also fix μ and denote the dependence of α_j on ϵ_1 by $\alpha_j(\epsilon_1)$. Since $\alpha_j(\epsilon_1) \nearrow \infty$ as $j \to \infty$ and because of Lemma 3.2, we may set

$$N_\alpha = \min\left\{j : \lim_{\epsilon \to 0} \alpha_j(\epsilon) > \alpha\right\}$$

and implicitly define a nondecreasing sequence of material parameters $\epsilon_1 = \{E_j(\alpha)\}_{j=N_\alpha}^{\infty}$ and a sequence of corresponding functions $\{\phi_j(\alpha)\}_{j=N_\alpha}^{\infty}$ that satisfy

$$\alpha_j(E_j(\alpha)) = \alpha, \qquad \phi_j(\alpha) = \psi_j(E_j(\alpha)).$$

Because the functions $\alpha_j(\epsilon_1)$ are continuous and strictly decreasing in ϵ, the values $E_j(\alpha)$ are defined uniquely, and, as functions of α, are continuous and strictly decreasing. By Theorem 3.1, these sequences satisfy

$$A(\phi_j(\alpha), v) = \alpha \ell_{E_j(\alpha)}(\phi_j(\alpha), v) \qquad \text{for all } v \in X.$$

If ω and $\boldsymbol{\kappa}$ are such that the medium exterior to $\mathcal{R}$ admits no propagating or linear Fourier harmonics, that is, if $\tilde{\mathcal{P}} = \emptyset$, then each $\phi_j(\alpha)$ can be extended to all of $\mathcal{S}$ as a guided mode, for the second part of Condition 2.5 requiring vanishing of all of these harmonics is vacuously satisfied.

In the case that Ω is symmetric about the yz-plane, $\boldsymbol{\kappa} = \langle 0, \kappa_2, 0\rangle$, and there is only one propagating Fourier harmonic, namely $(0,0)$, the antisymmetric eigenfunctions $\phi_j^{\text{ant}}(\epsilon_1)$ satisfy Condition 2.6 and therefore give rise to nonrobust guided modes.

We summarize these results in the following theorem.

THEOREM 3.3 (existence of guided modes). *For each $\alpha > 0$, there exists a sequence $\{E_j(\alpha)\}_{j=N_\alpha}^{\infty}$ of real numbers and a sequence $\{\phi_j(\alpha)\}_{j=N_\alpha}^{\infty}$ of functions from X such that*

i. *for each $\alpha > 0$, $0 < E_0(\alpha) \le E_1(\alpha) \le \cdots \le E_j(\alpha) \le \cdots$,*
ii. *for each $\alpha > 0$, $E_j(\alpha) \to \infty$ as $j \to \infty$,*
iii. *for each integer $n \ge 0$, $E_j(\alpha)$ is a strictly decreasing function of α, and $E_j(\alpha) \to \infty$ as $\alpha \to 0$,*
iv. *$\phi_j(\alpha)$ satisfies Condition* 2.5.i *for guided modes.*

If ω and $\boldsymbol{\kappa}$ are such that the medium exterior to $\mathcal{R}$ admits no propagating or linear Fourier harmonics, that is, if $\tilde{\mathcal{P}} = \emptyset$, then for each α and each j, the function $\phi_j(\alpha)$ satisfies both conditions of Condition 2.5. *In particular, it can be extended into $\mathcal{S}$ to a function that satisfies Condition* 2.2 *and gives rise to a guided mode*

$$\psi_j(\alpha)(x,y,z)e^{i(\kappa_1 x + \kappa_2 y - \omega t)}$$

in the augmented slab structure defined by the functions (3).

If Ω is symmetric about the yz-plane, $\boldsymbol{\kappa} = (0, \kappa_2, 0)$, and there is only one propagating Fourier harmonic $(0,0)$, the rest being evanescent, then for each α and j the function $\psi_j^{ant}(\alpha)$ satisfies Condition 2.5. *In particular, it can be extended into $\mathcal{S}$ to a function that satisfies Condition* 2.2, *giving rise to a nonrobust guided mode traveling parallel to the yz-plane:*

$$\psi_j^{ant}(\alpha)(x,y,z)e^{(\kappa_2 y - i\omega t)}.$$

An analogous statement holds if Ω is symmetric about the xz-plane.

There are two special cases of interest:

A. Taking $\alpha = 1$ gives the unaugmented structure $\tilde{\Omega}$, as the material properties ε and μ in $\mathcal{S} \setminus \mathcal{R}$ and $\mathcal{R} \setminus \Omega$ coincide:

$$\varepsilon(r) = \begin{cases} \epsilon_0, & r \in \mathcal{S} \setminus \Omega, \\ \epsilon_1, & r \in \Omega, \end{cases} \qquad \text{and} \quad \mu(r) = \begin{cases} \mu_0, & r \in \mathcal{S} \setminus \Omega, \\ \mu_1, & r \in \Omega. \end{cases}$$

Theorem 3.3 gives a sequence of constants $\epsilon_1 = E_j(1)$ for which a guided mode exists, provided the vanishing of all linear and propagating harmonics.

B. Fixing $\epsilon_1 = \epsilon_0$ and $\mu_1 = \mu_0$ corresponds to the case of a slab with no periodic structure, having uniform material properties in $\mathcal{R}$:

$$\varepsilon(x,y,z) = \begin{cases} \alpha\epsilon_0, & z \in \mathcal{R}, \\ \epsilon_0, & z \notin \mathcal{R}, \end{cases} \quad \text{and} \quad \mu(x,y,z) = \mu_0. \tag{27}$$

We analyze this instructive case explicitly in subsection 4.1.

4. Nonexistence of guided modes in inverse structures. An "inverse structure" is one in which the speed of light is higher than the speed of light in the surrounding medium. This means that $\epsilon_1\mu_1 < \epsilon_0\mu_0$. Using the sequences of eigenvalues constructed in section 3, we prove that certain inverse structures cannot support guided modes, robust or nonrobust. Specifically, we fix $\mu_1 = \mu_0 > 0$ arbitrarily and take $0 < \epsilon_1 < \epsilon_0$. Our statement requires an additional restriction on the width of the slab that depends on the material parameters, frequency, and wavevector (33). We do not know whether this restriction arises only as a consequence of our method of proof or whether it is truly necessary.

It is known that under a certain restrictive geometric condition, guided modes do not exist in inverse structures. Theorem 3.5 of [6], extended to three-dimensional structures, amounts to the following conditions for homogeneous slabs:

a. the surface Σ of the slab has two sides, given by $z = f_1(x,y) \le 0$ and $z = f_2(x,y) \ge 0$, where the common domain of f_1 and f_2 is a subset of the square $\{-\pi \le x,y \le \pi\}$;
b. $\epsilon_1\mu_1 \le \epsilon_0\mu_0$; that is, the celerity inside the slab is greater than that outside the slab (as for a film of air within a ceramic matrix).

The first condition is a severe restriction. It excludes, among other types of structures, periodic arrays of ellipses that do not have a major axis parallel to the z-axis and structures that some line parallel to the z-axis intersects in more than one segment.

In addition, it is shown in [6, Theorem 4.1] that, for a given wavevector $\boldsymbol{\kappa}$, the set of frequencies for which a guided mode exists is greater than $|\boldsymbol{\kappa}|/n_+$, where n_+ is the maximum value of $\epsilon\mu$. It follows that, if $\epsilon_1\mu_1 < \epsilon_0\mu_0$, then robust guided modes do not exist.

Our approach to proving the nonexistence of both types of modes for $\mu_1 = \mu_0$ is first to compute explicitly the eigenvalues $\alpha_j(\epsilon_0)$, corresponding to the case $\epsilon_1 = \epsilon_0$, in which the slab $\tilde{\Omega}_{\text{aug}}$ has no genuine periodicity and then to use the restriction on the width (33) to prove that the eigenvalues are all greater than or equal to 1. Finally, since the eigenvalues are decreasing as a function of ϵ_1, we observe that $\alpha_j(\epsilon_1) > 1$ for $\epsilon_1 < \epsilon_0$. As $\alpha = 1$ corresponds to the unaugmented periodic slab structure $\tilde{\Omega}$ with material constant ϵ_1 surrounded by a medium with constant ϵ_0, we conclude that no guided modes exist in $\tilde{\Omega}$ for $\epsilon_1 < \epsilon_0$.

Indeed, if $\epsilon_1 > \epsilon_0$, we have seen that nonrobust modes exist in symmetric structures if κ_1 or κ_2 vanishes. In fact, this restriction is not necessary: in [6], nonrobust modes are constructed for arbitrary Bloch wavevectors for two-dimensional slabs.

4.1. Eigenvalues for a flat slab. We explicitly compute the eigenvalues α_j and corresponding eigenfunctions ψ_j when $\epsilon_1 = \epsilon_0$ and $\mu_1 = \mu_0$ (see (27)). In this situation, the eigenfunctions satisfy the strong form of the Helmholtz equation in $\mathcal{R}$

(with $\psi = \psi_j$ and $\alpha = \alpha_j$):

$$\begin{aligned} &(\nabla + i\boldsymbol{\kappa})^2\psi + \alpha\epsilon_0\mu_0\omega^2\psi = 0 \text{ in } \mathcal{R}, \\ &\psi \in X \quad \text{and} \quad \partial_n\psi|_\Gamma = B\psi, \\ &\psi \text{ has periodic boundary conditions in } x \text{ and } y. \end{aligned} \tag{28}$$

Since $\varepsilon(x, y, z)$ is constant in x and y and $\mathcal{R}$ is bounded by planes parallel to the three coordinate planes, the method of separation of variables is applicable. The separable solutions have the simple form

$$\psi(x, y, z) = \left(A_{mn}e^{\eta_{mn}z} + B_{mn}e^{-\eta_{mn}z}\right)e^{i(mx+ny)}, \quad m, n \in \mathbb{Z}, \tag{29}$$

in which

$$\eta_{mn}^2 = (m + \kappa_1)^2 + (n + \kappa_2)^2 - \alpha\epsilon_0\mu_0\omega^2 \tag{30}$$

and $\operatorname{Im}\eta_{mn} > 0$ if $\eta_{mn}^2 < 0$ and $\operatorname{Re}\eta_{mn} > 0$ if $\eta_{mn}^2 > 0$. If $\eta_{mn} = 0$, then

$$\psi(x, y, z) = (A_{mn} + B_{mn}z)\,e^{i(mx+ny)}. \tag{31}$$

Each solution of the Helmholtz equation with periodic boundary conditions is a series superposition of separable solutions,

$$\psi(x, y, z) = \sum_{m,n=-\infty}^{\infty} \phi_{mn}(z)e^{i(mx+ny)},$$

in which ϕ_{mn} is of the form shown in (29) or (31). Moreover, the conditions that $\psi \in X$ and $\partial_n\psi\,|_\Gamma = B\psi$ impose *independent* conditions on the Fourier harmonics indexed by m and n on the boundary Γ:

$$\begin{aligned} &\left(\widehat{\psi|_\Gamma}\right)_{mn} = 0 \quad \text{for } (m, n) \in \tilde{\mathcal{P}}, \\ &\left(\partial_z\widehat{\psi|_{\Gamma_1}}\right)_{mn} = \nu_{mn}\left(\widehat{\psi|_{\Gamma_1}}\right)_{mn} \text{ and } \left(\partial_z\widehat{\psi|_{\Gamma_2}}\right)_{mn} = -\nu_{mn}\left(\widehat{\psi|_{\Gamma_2}}\right)_{mn} \text{ for } (m, n) \notin \tilde{\mathcal{P}}. \end{aligned} \tag{32}$$

Because of this, if ψ satisfies the Helmholtz equation as well as the boundary conditions (28), then each separable component (29) of ψ in its series representation also satisfies both. Therefore, each solution of (28) is composed of separable solutions.

To find the values of α that admit such solutions and the solutions themselves, we impose the condition (32) on the separable solution (29) or (31) for each pair (m, n). For each fixed $(m, n) \in \tilde{\mathcal{P}}$, the condition (32) is possible only for values of α for which $\eta_{mn}^2 < 0$, which give oscillatory solutions in the interval from z_1 to z_2. In addition, in order for (32) to hold, η_{mn} must be of the form $\eta_{mn} = i\left(\frac{j\pi}{z_2 - z_1}\right)$ for some j (independent of (m, n)), and we thus arrive at a sequence of eigenvalues $\alpha = \alpha_{mnj}$ satisfying

$$\alpha_{mnj}\epsilon_0\mu_0\omega^2 = \left(\frac{j\pi}{z_2 - z_1}\right)^2 + (m + \kappa_1)^2 + (n + \kappa_2)^2, \quad j = 1, 2, 3, \ldots .$$

It is straightforward to deduce from the condition (32) that the constants A_{mn} and B_{mn} in (29) have the same modulus, so that, by multiplying the solution by a unitary number $e^{i\theta}$, we may take $\phi_{mnj}(z)$ to be a shifted sine function inside the region $\mathcal{R}$. These solutions do not represent guided modes because they do not satisfy the second

part of Condition 2.5 requiring the normal derivative of ϕ_{mnj} to vanish, and therefore their extensions to all of $\mathcal{S}$ do not decay as $|z| \to \infty$.

For $(m,n) \notin \tilde{\mathcal{P}}$, the condition (32) amounts to matching a solution that is decaying as $z \to -\infty$ for $z < z_1$ to one that is decaying as $z \to \infty$ for $z > z_2$ through a solution in the interval from z_1 to z_2. This is possible only if the solution in this interval is oscillatory, and this is achievable only when $\eta_{mn}^2 < 0$. By enforcing the decay of the solution as $|z| \to \infty$, we obtain a sequence of eigenvalues $\alpha = \alpha_{mnj}$ with $\alpha_{mnj} \to \infty$ as $j \to \infty$ satisfying

$$\tan \zeta(z_2 - z_1) = \frac{2\nu_{mn}\zeta}{\zeta^2 - \nu_{mn}^2}, \qquad \zeta = \left(\alpha_{mnj}\epsilon_0\mu_0\omega^2 - (m+\kappa_1)^2 - (n+\kappa_2)^2\right)^{1/2}.$$

Again, by multiplying the solution by a unitary number, we may take $\phi_{mnj}(z)$ to be a shifted sine function inside the region $\mathcal{R}$. These solutions satisfy Condition 2.5, even when $\tilde{\mathcal{P}} \neq \emptyset$, as they involve only one Fourier harmonic.

The union of the sequences $\{\alpha_{mnj}\}$, arranged in increasing order, gives the sequence $\{\alpha_j\}$ that we seek.

As ϵ_1 is perturbed away from ϵ_0, the structure attains a genuine periodicity, and separable solutions are no longer valid. Typically all Fourier harmonics are represented in the eigenfunctions so that the guided modes disappear in a regime admitting linear or propagating harmonics. As we have seen, however, antisymmetric nonrobust modes persist, for example, in symmetric structures for which there is only one propagating harmonic, the rest being evanescent.

4.2. Nonexistence of guided modes. We use the foregoing analysis to prove a theorem stating that guided modes do not exist in certain structures in which the interior product of the material coefficients $\mu_1\epsilon_1$ is greater than the exterior product $\mu_0\epsilon_0$.

THEOREM 4.1 (nonexistence of guided modes). *Let $0 < \mu_1 \leq \mu_0$ and $0 < \epsilon_1 \leq \epsilon_0$, and let the frequency ω and wavevector $\boldsymbol{\kappa} = \langle \kappa_1, \kappa_2, 0\rangle$ be given with $\boldsymbol{\kappa}$ in the first Brillouin zone: $-\frac{1}{2} \leq \kappa_1,\kappa_2 < \frac{1}{2}$. Suppose that the slab structure $\tilde{\Omega}$ (Figure 2 with $\alpha=1$) lies between two planes $\{z = z_1\}$ and $\{z = z_2\}$ satisfying*

$$(z_2 - z_1)(\epsilon_0\mu_0\omega^2 - \kappa_1^2 - \kappa_2^2)^{1/2} \leq \pi \tag{33}$$

in the case that $\epsilon_0\mu_0\omega^2 - \kappa_1^2 - \kappa_2^2 \geq 0$, that is, $\tilde{\mathcal{P}} \neq \emptyset$ (otherwise, there is no restriction). Then the slab with material properties

$$\varepsilon(r) = \begin{cases} \epsilon_0, & r \notin \tilde{\Omega}, \\ \epsilon_1, & r \in \tilde{\Omega}, \end{cases} \qquad \text{and} \qquad \mu(r) = \begin{cases} \mu_0, & r \notin \tilde{\Omega}, \\ \mu_1, & r \in \tilde{\Omega}, \end{cases}$$

admits no guided modes at the given frequency and wavevector.

Proof. We begin showing that, for $\epsilon_1 = \epsilon_0, \mu_1 = \mu_0$, the slab admits no guided modes. This is the case of a flat slab analyzed in subsection 4.1. Recall the definition of ν_{mn}^2,

$$\nu_{mn}^2 = -\epsilon_0\mu_0\omega^2 + (m+\kappa_1)^2 + (n+\kappa_2)^2,$$

and define, for each $\alpha > 0$, as before,

$$\eta_{mn}^2(\alpha) = -\alpha\epsilon_0\mu_0\omega^2 + (m+\kappa_1)^2 + (n+\kappa_2)^2.$$

In subsection 4.1, we have seen that the eigenvalues α_j (for $\epsilon_1 = \epsilon_0, \mu_1 = \mu_0$) correspond to eigenfunctions containing a single Fourier harmonic, and we wish to show that all of these eigenvalues are greater than or equal to 1.

For those pairs (m,n) for which $\nu_{mn}^2 > 0$, corresponding to the evanescent Fourier harmonics ($(m,n) \notin \tilde{\mathcal{P}}$), we have seen in subsection 4.1 that the matching conditions at $z = z_1$ and $z = z_2$ require that $\eta_{mn}^2(\alpha) < 0$. From the definitions of ν_{mn}^2 and $\eta_{mn}^2(\alpha)$, we conclude that $\alpha > 1$, so that all the eigenvalues corresponding to the evanescent harmonics are at least greater than 1.

For $(m,n) \in \tilde{\mathcal{P}}$, we still require that $\eta_{mn}^2(\alpha) < 0$. Since $\boldsymbol{\kappa}$ is taken to lie in the first symmetric Brillouin zone, that is, $-1/2 \le \kappa_1 < 1/2$ and $-1/2 \le \kappa_2 < 1/2$, we have (recall that $\operatorname{Im}(\nu_{mn}(\alpha)) > 0$ (p. 692))

$$-i\nu_{mn} \le -i\nu_{00} \quad \text{for all} \quad (m,n) \in \tilde{\mathcal{P}}.$$

According to the discussion of the preceding subsection, to satisfy the boundary conditions at $z = z_1$ and $z = z_2$, α must be chosen such that

$$-i\eta_{mn}(\alpha) = \left(\frac{j\pi}{z_2 - z_1}\right), \quad j \text{ a positive integer.}$$

From condition (33), we obtain

$$-i\eta_{mn}(\alpha) = \frac{j\pi}{z_2 - z_1} \ge j\,(\epsilon_0\mu_0\omega^2 - \kappa_1^2 - \kappa_2^2)^{1/2} \ge -i\nu_{00} \ge -i\nu_{mn},$$

from which it follows that $\alpha \ge 1$, so that all the eigenvalues corresponding to the propagating harmonics are at least 1. As we have mentioned in the previous subsection, these eigenvalues do not correspond to guided modes.

Since the eigenvalues α_j are strictly decreasing in ϵ_1, as well as in μ_1 (Lemma 3.2), we have $\alpha_j > 1$ if both $\epsilon_1 \le \epsilon_0$ and $\mu_1 \le \mu_0$, which proves the theorem. ∎

5. Numerical computations. We compute guided modes for the Helmholtz equation. These are scalar functions u satisfying Condition 2.2, for which $\alpha = 1$. We focus on the case of one propagating Fourier harmonic, and we consider a two-dimensional reduction, in which the slab is constant in the y direction and $\kappa_2 = 0$. In this case, only the $(m,0)$ Fourier harmonics enter the fields. Our method begins with a geometry Ω that is symmetric about the yz-plane (in the two-dimensional reduction to the x and z variables, this implies symmetry about the z-axis) and given values of $\boldsymbol{\kappa}$, ϵ_0, μ_0, μ_1, and ω. The code then computes the values of ϵ_1 which give rise to a solution of the first part of Condition 2.6; in other words, it computes one of the values $E_j(\alpha)$. The second part of the condition is automatically satisfied because only one harmonic is propagating, namely that with $(m,n) = (0,0)$. The corresponding antisymmetric nonrobust guided mode u is also computed.

We use a finite element solver in the finite rectangular region $\mathcal{R}$ for the eigenvalue problem, with ϵ_1 as the eigenvalue, for the Helmholtz equation in two variables, x and z:

$$(\nabla + i\kappa)\cdot(\nabla + i\kappa)u + \epsilon_0\mu_0\omega^2 u = 0 \quad \text{in } \mathcal{R}\setminus\Omega, \tag{34}$$

$$(\nabla + i\kappa)\cdot(\nabla + i\kappa)u = -\epsilon_1\mu_0\omega^2 u \quad \text{in } \Omega, \tag{35}$$

$$\mu_1\,(\partial_n u_+ + (i\boldsymbol{\kappa}\cdot n)u) = \mu_0\,(\partial_n u_- + (i\boldsymbol{\kappa}\cdot n)u) \quad \text{on } \Sigma, \tag{36}$$

$$u(-\pi,z) = u(\pi,z) \quad \text{and} \quad \partial_n u(-\pi,z) = -\partial_n u(\pi,z), \tag{37}$$

$$\partial_n u + \nu_0 u = 0 \quad \text{on the edges } z = z_1 \text{ and } z = z_2. \tag{38}$$

Note that since we assumed that only one harmonic propagates, $\nu_0 = i\sqrt{\epsilon_0\mu_0\omega^2 - |\kappa|^2}$, and condition (38) expresses that there are no incoming harmonics impinging the

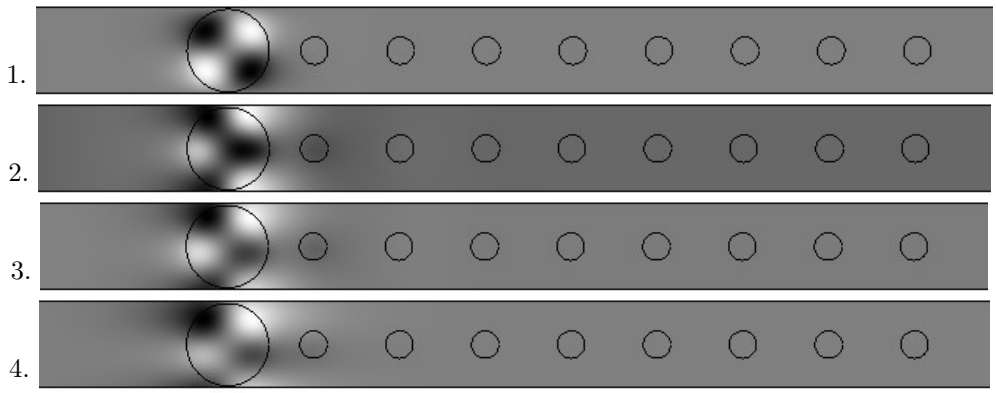


FIG. 3. *Four guided modes in a two-dimensional structure investigated in* [2, Figure 2]. *One period is shown; the structure continues periodically in the vertical direction on the page.* 1: *A nonrobust antisymmetric guided mode at Bloch wavevector zero,* $(\kappa_1, \omega) = (0.0, 0.4017)$. 2 *and* 3: *Nonrobust guided modes at nonzero wave numbers in the direction perpendicular to the line of symmetry,* $(\kappa_1, \omega) = (0.14, 0.3863), (0.22, 0.3707)$. 4: *A robust guided mode,* $(\kappa_1, \omega) = (0.44, 0.3306)$.

rectangular zone $\mathcal{R}$. In fact, condition (38) is a first approximation of the Dirichlet-to-Neumann operator B. If only one harmonic is allowed to propagate and if the rectangular region $\mathcal{R}$ is chosen to be wide enough, it is reasonable to believe that this first approximation leads to an exponentially small error. Numerical methods hinging on this boundary approximation idea have been used in the literature, for example by Kriegsmann [10] and Volkov and Kriegsmann [11], albeit in the case of regular transmission problems instead of eigenvalue problems.

We discretize (34)–(38) by finite elements on a meshing of $\mathcal{R}$, and then solve the discretized problem as an eigenvalue problem in ϵ_1. A function u satisfying (34)–(38) satisfies

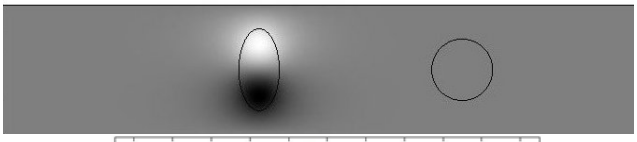
$$0 = \nu_0 \int_\Gamma |u|^2 - \int_{\mathcal{R}} |(\nabla + i\kappa)u|^2 + \epsilon_0\mu_0\omega^2 \int_{\mathcal{R}\setminus\Omega} |u|^2 + \epsilon_1\mu_0\omega^2 \int_\Omega |u|^2.$$

Thus, loosely speaking, if the imaginary part of ϵ_1 is very small, u and $\partial_n u$ are very small on Γ, which is made up by the two narrow edges of the rectangle $\mathcal{R}$. If the rectangle $\mathcal{R}$ is long enough, this simulates the exponential decay expected from a Bloch solution to the Helmholtz equation that is a guided mode.

We first use our numerical method to reproduce a computation of eigenvalues for bound states, which appeared in [2]. The geometry under consideration is that of a dielectric made up of one large circle of radius 3 and eight small circles of radius 1. (The circles are cross sections of rods that extend infinitely in the y direction.) Their centers lie on the line $x = 0$, and two consecutive centers are 2π units of length apart. Fixing $\mu_0 = \mu_1 = 1$, $\epsilon_0 = 1$, $\epsilon_1 = 12$, it was found in [2] that guided modes (referred to as "bound states" in that paper) exist for the pairs $(\kappa_1 = 0.0, \omega = 0.4017)$, $(\kappa_1 = 0.14, \omega = 0.3863)$, $(\kappa_1 = 0.22, \omega = 0.3707)$, and $(\kappa_1 = 0.44, \omega = 0.3306)$. Thus with our present numerical method, we fix $\mu_0 = \mu_1 = 1$, $\epsilon_0 = 1$, and (κ_1, ω) at one of these pairs, and compute values for ϵ_1 for which there appears to be a guided mode. The computations lead to $\epsilon_1 \approx 12.0$, as expected. The corresponding eigenfunction is plotted using grayscale coloring in Figure 3.

The first of these guided modes, at $(\kappa_1 = 0.0, \omega = 0.4017)$, is antisymmetric about the yz-plane. It is nonrobust because it exists in the (κ_1, ω)-regime of one

1.



2.

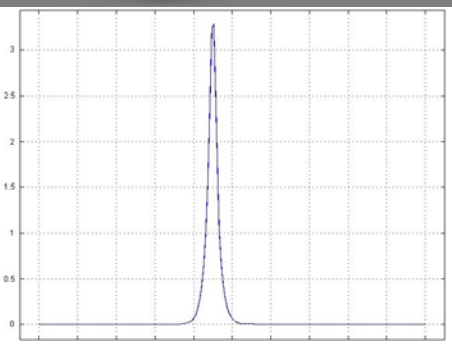


FIG. 4. 1: *Eigenfunction for the first real eigenvalue* $\epsilon_1 \approx 9.762$ *for the parameters* $\omega = 1, \epsilon_0 = 1, \mu_0 = \mu_1 = 1, \kappa = (0, 0)$. 2: *Cross section of the solution along the line* $z = \pi/2$*; the magnitude of the solution is plotted.*

propagating Fourier harmonic, which is suppressed by the symmetry of the structure and the vanishing of κ_1. The last of these modes, at $(\kappa_1 = 0.44, \omega = 0.3306)$, is a robust guided mode, for it exists in the (κ_1, ω)-regime in which all Fourier harmonics are evanescent. A dispersion relation for these is shown in [2].

The other two nonrobust modes are not discussed in the analysis in this paper, for they are in a (κ_1, ω)-regime of one propagating harmonic, but the wave number κ_1 in the x direction is not zero. However, at these two pairs, the coefficient for the one propagating Fourier harmonic happens to be zero; that is, the second part of Condition 2.6 happens to be satisfied. These values of κ_1 and ω occur at points of a complex dispersion curve calculated in [2, Figure 7.2, part 7], at which the imaginary part of the frequency appears to vanish. The existence of nonrobust modes at nonzero wave numbers is proved in the final section of [6].

We also show computations involving geometries that are not exclusively circular. We choose to place two dielectrics, one shaped as an ellipse of focal lengths 1 and 2 and centered at $(-5, \pi)$, the other shaped as a circle of a radius 1 and centered at $(5, \pi)$. Their boundaries appear in Figures 4, 5, and 6. We pick the values $z_1 = -50, z_2 = 50$ to bound the rectangle $\mathcal{R}$. We first assume that $\omega = 1, \epsilon_0 = 1, \mu_0 = \mu_1 = 1, \kappa = (0, 0)$, ensuring that only one harmonic mode propagates. Thus, we know that nonrobust guided modes do exist at certain values of ϵ_1.

The first guided mode that we find corresponds to $\epsilon_1 \approx 9.762$, and is plotted in Figure 4. Values were obtained with a mesh containing 4592 elements. Numerical convergence was verified by either quadrupling the number of mesh elements or changing $z_1 = -50, z_2 = 50$ into $z_1 = -60, z_2 = 60$. These refinements did not change the first four digits of the numerical value for ϵ_1. The numerical method employed finds complex eigenvalues and sorts them in increasing real part order. Some of those eigenvalues do not have a small imaginary part: we ignore them, as they are unrelated to the solutions we are trying to compute. We verify decay of the solution as we move away from the dielectrics. This is illustrated in the graph in Figure 4, which shows the absolute value of the solution along the line $z = \pi/2$. We also show the graphs of the second and third guided modes, still for the same values of $\omega, \epsilon_0, \mu_0, \mu_1, \kappa_1$. They appear in Figure 5.

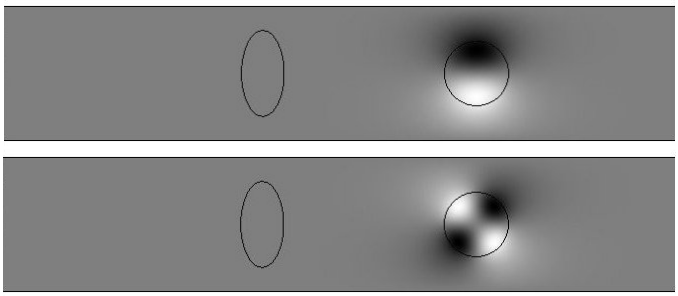


FIG. 5. *Eigenfunctions for the second and third real eigenvalues* $\epsilon_1 \approx 11.00$ *and* $\epsilon_1 \approx 25.66$ *for the parameters* $\omega = 1, \epsilon_0 = 1, \mu_0 = \mu_1 = 1, \kappa_1 = 0$. *These are nonrobust guided modes.*

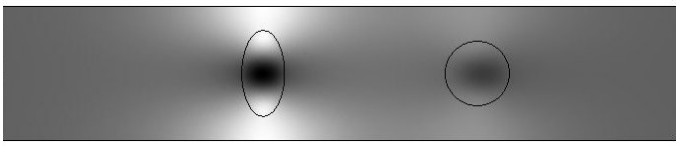


FIG. 6. *Eigenfunction for the third real eigenvalue* $\epsilon_1 \approx 11.42$ *for the parameters* $\omega = 0.3, \epsilon_0 = 1, \mu_0 = 1, \mu_1 = 3, \kappa_1 = 0.4$. *This is a robust guided mode.*

Finally, we compute a robust guided mode with $\kappa_1 \neq 0$. To guarantee existence, we choose values for the material parameters such that no harmonic can propagate. More precisely, we select $\omega = 0.3, \epsilon_0 = 1, \mu_0 = 1, \mu_1 = 3, \kappa_1 = 0.4$. The third eigenfunction for that case is plotted in Figure 6.

6. Appendix: Proofs of theorems.

Proof of Theorem 2.7 (equivalence of strong and weak conditions). Many of the arguments are standard in the literature on elliptic equations; details of the relevant theory can be found in [9, Chapter 8], for example. We confine discussion to the basic elements of the proof and those aspects that are unique to this problem.

i. That the strong formulation satisfies the weak is a matter of application of the divergence theorem (integration by parts). The relevant identity is

$$\nabla\cdot\left[\left(\frac{1}{\mu}\left(\nabla+i\boldsymbol{\kappa}\right)u\right)\bar{v}\right] = \left[\left(\nabla+i\boldsymbol{\kappa}\right)\cdot\left(\frac{1}{\mu}\left(\nabla+i\boldsymbol{\kappa}\right)u\right)\right]\bar{v} + \frac{1}{\mu}\left(\nabla+i\boldsymbol{\kappa}\right)u\cdot\left(\nabla-i\boldsymbol{\kappa}\right)\bar{v}. \tag{39}$$

Applying this identity for a function u that satisfies the strong Condition 2.2, the left-hand side of the weak Condition 2.4.i becomes

$$\begin{aligned} -\int_{\mathcal{R}}\left[\frac{1}{\mu}\left(\left(\nabla+i\boldsymbol{\kappa}\right)^2 u\right)+\varepsilon\omega^2 u\right]\bar{v} \\ +\frac{1}{\mu_0}\int_{\Gamma}\left(Bu+\partial_n u_-\right)T\bar{v} - \int_{\partial\mathcal{R}\setminus\Gamma}\frac{1}{\mu}\partial_n u T\bar{v} \\ +\int_{\Sigma}\left[\frac{1}{\alpha\mu_1}\left(\partial_n u_- + (i\boldsymbol{\kappa}\cdot n)u\right) - \frac{1}{\alpha\mu_0}\left(\partial_n u_+ + (i\boldsymbol{\kappa}\cdot n)u\right)\right]T\bar{v}. \end{aligned} \tag{40}$$

The integral over $\mathcal{R}$ vanishes by properties (i)–(iii), the integral over Γ by property (v), that over $\partial\mathcal{R}\setminus\Gamma$ by properties (vii)–(viii), and the integral over Σ by property (vi).

ii. Let u satisfy Condition 2.4. The functions v of class C^∞ with compact support in $\mathcal{R}\setminus\Sigma$ are contained in $H^1(\mathcal{R})$, and this is sufficient to establish that u satisfies the Helmholtz equation in $\mathcal{R}\setminus\Sigma$ (i)–(iii) and that $u\in H^2(\mathcal{R})$ [9, section 8.3]. Thus u has values on $\partial\mathcal{R}$ (including the interior side of Γ) and Σ that are of class $H^{3/2}$ and normal derivatives of class $H^{1/2}$. Integration by parts, using the Helmholtz equation away from these boundaries, establishes properties (iv)–(viii). The extension of u to all of $\mathcal{S}$ is achieved by the formula (22).

The form of the extension (22) of u to $\mathcal{S}$ outside of $\mathcal{R}$ shows that $a^-_{mn}=b^+_{mn}=0$ for all (m,n). The relation (15) expressing conservation of energy, which is obtained by integration by parts with $v=u$, shows that $a^+_{mn}=b^-_{mn}=0$ for $(m,n)\in\mathcal{P}$. The additional requirement in Condition 2.4 for $\nu^2_{mn}=0$ establishes $a^+_{mn}=b^-_{mn}=0$ for $(m,n)\in\tilde{\mathcal{P}}$. Therefore each harmonic with $(m,n)\in\tilde{\mathcal{P}}$ in the expansion (9) has vanishing value and normal derivative on Γ, implying that $u\in X$ and Condition 2.5.ii are satisfied.

iii. The functions v of class C^∞ with compact support in $\mathcal{R}\setminus\Sigma$ are contained in X, and again we obtain that u satisfies the Helmholtz equation in $\mathcal{R}\setminus\Sigma$ and $u\in H^2(\mathcal{R})$. It suffices to prove that, for each $(m,n)\in\mathcal{P}$, the weak form in Condition 2.5 holds for v such that $(\widehat{v|_{\Gamma_1}})_{mn}=1$, $(\widehat{v|_{\Gamma_1}})_{m'n'}=0$ for $(m',n')\neq(m,n)$, and $(\widehat{v|_{\Gamma_2}})_{m'n'}=0$ for all (m',n') (and similarly with Γ_1 and Γ_2 interchanged). Applying integration by parts for such v together with the Helmholtz equation for u yields for the left-hand side of the equation in the weak Condition 2.4,

$$\text{left-hand side}=\int_{\Gamma_1}(\partial_n u+Bu)\bar v=\big((\partial_n\widehat{u|_{\Gamma_1}})_{mn}-\nu_{mn}(\widehat{u|_{\Gamma_1}})_{mn}\big)=0,$$

in which $(\partial_n\widehat{u|_{\Gamma_1}})_{mn}=0$ by Condition 2.5.ii and $(\widehat{u|_{\Gamma_1}})_{mn}=0$ because $u\in X$.

iv. To prove Condition 2.5.i, it suffices to prove the equality for $v\in X^{\mathrm{sym}}$, which follows from the observation that the integrands are antisymmetric over the regions of integration. To prove part (ii), we observe that the Fourier coefficients with m even are zero because u is antisymmetric in the x-variable. ∎

Recall the sesquilinear forms in $H^1_{\mathrm{per}}(\mathcal{R})$,

$$A(u,v)=\int_{\mathcal{R}}\frac{1}{\mu}\,(\nabla+i\boldsymbol{\kappa})\,u\cdot(\nabla-i\boldsymbol{\kappa})\,\bar v+\frac{1}{\mu_0}\int_\Gamma(BTu)(T\bar v), \tag{41}$$

$$\ell(u,v)=\int_{\mathcal{R}\setminus\Omega}\epsilon_0\omega^2u\bar v+\int_\Omega\epsilon_1\omega^2u\bar v, \tag{42}$$

and that A depends on $\boldsymbol{\kappa}$, ω, ϵ_0, μ_0, and μ_1 (the dependence on ω and ϵ_0 is through B—see (19) and (10)), and ℓ depends on ω, ϵ_0, and ϵ_1; neither form depends on α.

LEMMA 6.1 (estimates). *There exist positive numbers C and δ such that, for all $u,v\in H^1(\mathcal{R})$,*

i. $\min\{\epsilon_0,\epsilon_1\}\|u\|^2_{L^2}\le\ell(u,u)\le\max\{\epsilon_0,\epsilon_1\}\|u\|^2_{L^2}$ *(equivalence of $\|\cdot\|_{L^2}$ and $\ell(\cdot,\cdot)$),*

ii. $|A(u,v)| \le C\|u\|_{H^1}\|v\|_{H^1}$ *(boundedness of A),*
iii. $\delta\|u\|_{H^1}^2 \le A(u,u)$ *(coercivity of A).*

These constants depend on the parameters $\boldsymbol{\kappa}$, ω, ϵ_0, μ_0, *and* μ_1.

Proof.

i. Part (i) is straightforward to verify.
ii. Because the trace operator $T : H^1(\mathcal{R}) \to H^{1/2}(\Gamma)$ and the operator $B : H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$ are bounded, there is a constant C_1 such that

$$\left|\int_\Gamma (BTu)(T\bar{v})\right| \le C_1\|u\|_{H^1}\|v\|_{H^1}.$$

This, together with the estimate

$$\begin{aligned}\min\{\mu_0,\mu_1\}\left|\int_{\mathcal{R}} \frac{1}{\mu}\,(\nabla + i\boldsymbol{\kappa})\,u\cdot(\nabla - i\boldsymbol{\kappa})\,\bar{v}\right| &\le \left|\int_{\mathcal{R}} (\nabla + i\boldsymbol{\kappa})\,u\cdot(\nabla - i\boldsymbol{\kappa})\,\bar{v}\right| \\ &\le \|\,(\nabla + i\boldsymbol{\kappa})\,u\|_{L^2}\,\|\,(\nabla + i\boldsymbol{\kappa})\,v\|_{L^2} \\ &\le (\|\nabla u\|_{L^2} + |\boldsymbol{\kappa}|\|u\|_{L^2})\,(\|\nabla u\|_{L^2} + |\boldsymbol{\kappa}|\|u\|_{L^2}) \le C_2\|u\|_{H^1}\|u\|_{H^1},\end{aligned}$$

proves the estimate.
iii. Suppose, to the contrary, that there exists a sequence $\{u_n\}_{n=0}^{\infty}$ from X such that

$$\|u_n\|_{H^1} = 1 \quad \text{and} \quad A(u_n,u_n) \to 0 \text{ as } n\to\infty. \tag{43}$$

Because of the compact embedding of X into $L^2(\mathcal{R})$, we simply assume that there is a function $u \in L^2(\mathcal{R})$ such that

$$\|u - u_n\|_{L^2(\mathcal{R})} \to 0,$$

and from the definition of A, we see that

$$\|\,(\nabla + i\boldsymbol{\kappa})\,u_n\|_{L^2(\mathcal{R})} \to 0,$$

whence

$$\|\nabla u_n + i\boldsymbol{\kappa}u\|_{L^2(\mathcal{R})} \to 0.$$

It follows that $u_n \to u$ in the H^1-norm so that $u \in X$ and

$$(\nabla + i\boldsymbol{\kappa})\,u \;=\; 0,$$

from which we obtain

$$u = C_3 e^{i(\kappa_1 x + \kappa_2 y)} \tag{44}$$

for some constant C. From the convergence of u_n to u in X, the boundedness of T and B, and the definition of A, we have

$$\int_\Gamma (BTu)T\bar{u} = \lim_{n\to\infty}\int_\Gamma (BTu_n)T\bar{u}_n = 0.$$

Since B is a positive operator on $H^{1/2}(\Gamma)$, we obtain $Tu = 0$, and the form (44) therefore gives $u = 0$. This is in contradiction to the supposition that $\|u_n\|_{H^1} = 1$, and (43) is therefore untenable. ∎

Because of the equivalence of the norms $\ell(f,f)$ and $\|f\|_{L^2(\mathcal{R})}$ (Lemma 6.1.ii), we may define $L^2(\mathcal{R},\ell)$ to be the linear space of functions in $L^2(\mathcal{R})$ endowed with the inner product $\ell(f,g)$.

Proof of Theorem 3.1. By the Lax–Milgram theorem, there exists a linear operator $K : L^2(\mathcal{R},\ell) \to X$ such that, for each $f \in L^2(\mathcal{R})$,

$$A(Kf,v) = \ell(f,v) \quad \text{for all } v \in X,$$

and by the sesquilinearity of A and ℓ, we also have

$$A(v,Kf) = \ell(v,f) \quad \text{for all } v \in X.$$

K is self-adjoint in the inner product $\ell(\cdot,\cdot)$ because

$$\ell(Kf,g) = A(Kf,Kg) = \ell(f,Kg) \quad \text{for all } f,\, g \in L^2(\mathcal{R}).$$

In fact, K is positive because

$$\ell(Kf,f) = A(Kf,Kf) > 0 \quad \text{for all } f \in L^2(\mathcal{R}).$$

K is injective because, if $Kf = 0$, then $\ell(f,v) = 0$ for all $v \in X$, and since X contains the functions of class C^∞ with compact support in $\mathcal{R}$, $f = 0$ almost everywhere, so that $f = 0$ in $L^2(\mathcal{R})$. The estimate

$$\delta\|Kf\|_{H^1}^2 \le A(Kf,Kf) = \ell(f,Kf) \le C\|f\|_{L^2}\|Kf\|_{L^2} \le C\|f\|_{L^2}\|Kf\|_{H^1}$$

gives us

$$\delta\|Kf\|_{H^1} \le C\|f\|_{L^2},$$

which shows that K is compact as an operator on $L^2(\mathcal{R})$. As the $L^2(\mathcal{R})$-norm with respect to Lebesgue measure on $\mathcal{R}$ and the norm in $L^2(\mathcal{R},\ell)$ are equivalent, K is compact as an operator on $L^2(\mathcal{R},\ell)$. The spectrum of K therefore consists of a nonincreasing sequence of eigenvalues $\{\lambda_j\}_{j=0}^\infty$ converging to zero, in which eigenvalues are repeated according to multiplicity, and there is a corresponding sequence of eigenfunctions $\{\psi_j\}_{j=0}^\infty$ that form an orthonormal Hilbert-space basis for $L^2(\mathcal{R},\ell)$.

By definition of A and K, we see that, for $\alpha \in \mathbb{R}$,

$$A(u,\cdot) = \alpha\ell(u,\cdot) \iff Ku = \alpha^{-1}u, \tag{45}$$

in which the dot in the second argument of the forms indicates action on X. The sequence of eigenvalues α_j we seek is therefore

$$\{\alpha_j = \lambda_j^{-1}\}_{j=0}^\infty.$$

By Lemma 6.1.i and 6.1.iii,

$$\ell(u,u) \le C_4 A(u,u),$$

which shows that $\alpha_0 > 0$.

We now show that the eigenvalues and eigenfunctions arise from minimization of the Rayleigh quotient. Define

$$\beta_0 = \inf_{u \in X, u \neq 0} J(u).$$

We prove that there exists a nonzero function $\phi_0 \in X$ such that

$$\beta_0 = J(\phi_0) > 0.$$

Let $\{u_n\}_{n=1}^{\infty}$ be a minimizing sequence; that is, $u_n \neq 0$ for each n and $\lim_{n\to\infty} J(u_n) = \beta_0$. By homogeneity of $J(u)$, we may assume that $\ell(u_n, u_n) = 1$ for each n, and since the sequence $\{J(u_n)\}$ is bounded, $\{A(u_n, u_n)\}$ is also bounded. Lemma 6.1.iii shows that $\{\|u_n\|_{H^1}\}$ is bounded.

By the compact embedding of $H^1_{\mathrm{per}}(\mathcal{R})$ into $L^2(\mathcal{R})$, there is a subsequence that is strongly convergent in $L^2(\mathcal{R})$; we simply assume therefore that $\{u_n\}$ is strongly convergent in $L^2(\mathcal{R})$, say to a function ϕ_0. By the second inequality in Lemma 6.1.i, $\ell(\phi_0, \phi_0) = 1$. We now prove that $\|u_n - u_m\|_{H^1} \to 0$ as $m, n \to \infty$. The parallelogram law holds for A:

$$A(u_m - u_n, u_m - u_n) = A(u_m, u_m) + A(u_n, u_n) - A(u_m + u_n, u_m + u_n). \tag{46}$$

Because $\ell(u_n, u_n) = 1$, $A(u_n, u_n) = J(u_n)$, and the sum of the first two terms on the right-hand side of (46) converges to $4\beta_0$. By definition of β_0 and because of the second inequality in Lemma 6.1.i,

$$A(u_m + u_n, u_m + u_n) \geq \beta_0 \ell(u_m + u_n, u_m + u_n) \to \beta_0 \ell(2u, 2u) = 4\beta_0 \quad \text{as } m, n \to \infty.$$

We thus obtain $A(u_m - u_n, u_m - u_n) \to 0$ as $m, n \to \infty$, and from Lemma 6.1.iii, $\|u_n - u_m\|_{H^1} \to 0$. Therefore, $u_n \to \phi_0 \in X$ in the H^1-norm. Part (ii) shows that $A(u_n, u_n) \to A(\phi_0, \phi_0)$ as $n \to \infty$, and therefore

$$J(\phi_0) = \lim_{n\to\infty} J(u_n) = \beta_0.$$

To define β_1 and ϕ_1, we set Y_1 to be the orthogonal complement of $\mathrm{span}\{\phi_0\}$ in X with respect to the sesquilinear form $A(u, v)$,

$$Y_1 = \{v \in X : A(\phi_0, v) = 0\},$$

and define

$$\beta_1 = \inf_{u \in Y_1, u \neq 0} J(u).$$

The proof of the existence of a minimizer ϕ_1 in Y_1 is essentially the same as the proof of the existence of ϕ_0. Continuing in this way, we obtain a sequence $\{Y_j\}$ of subspaces of X, numbers β_j, and functions ϕ_j such that

$$Y_j = \{v \in X : A(\phi_k, v) = 0 \text{ for } k = 0, \ldots, j-1\}$$

and

$$\beta_j = \inf_{u \in Y_j, u \neq 0} J(u) = J(\phi_j), \quad \phi_j \in Y_j.$$

Taking the first variation of the relation $A(u, u) = J(u)\ell(u, u)$ at $u = \phi_j$ and using the fact that J is minimized by ϕ_j in Y_j and that $A(\phi_k, \phi_j) = \ell(\phi_k, \phi_j) = 0$ for $k = 0, \ldots, j-1$, we obtain

$$A(\phi_j, v) = \beta_j \ell(\phi_j, v) \quad \text{for all } v \in X. \tag{47}$$

By definition, ϕ_{j+1} minimizes the same functional as ϕ_j, but over a smaller set, and therefore the sequence $\{\beta_j\}$ is nondecreasing:

$$0 < \beta_0 \le \beta_1 \le \cdots \le \beta_j \le \cdots .$$

Because of (45) and (47), we have

$$\{\beta_j\}_{j=0}^{\infty} \subseteq \{\alpha_j\}_{j=0}^{\infty} \quad \text{and} \quad \operatorname{span}\{\phi_j : j = 0, \ldots, \infty\} \subseteq \operatorname{span}\{\psi_j : j = 0, \ldots, \infty\}.$$

To show for $j = 0, \ldots, \infty$ that $\alpha_j = \beta_j$, that ψ_j can be taken to be equal to ϕ_j, and that $X_j = Y_j$, we prove that any eigenvalue α with eigenfunction $0 \neq \psi \in X$, in the sense that

$$A(\psi, v) = \alpha \ell(\psi, v) \quad \text{for all } v \in X,$$

is necessarily one of the β_j and that ψ is in the span of $\{\phi_j : \beta_j = \alpha\}$. If, to the contrary, $\alpha \neq \beta_j$ for all n, then $A(\phi_j, \psi) = 0$ for all n because

$$A(\psi, \phi_j) = \alpha \ell(\psi, \phi_j) \quad \text{and} \quad A(\phi_j, \psi) = \beta_j \ell(\phi_j, \psi),$$

whence we obtain, from conjugating the first relation and keeping in mind that $\alpha \ge \alpha_0 > 0$ and $\beta_j \ge \alpha_0 > 0$,

$$(\alpha^{-1} - \beta_j^{-1}) A(\phi_j, \psi) = 0.$$

Since $\alpha \neq \beta_j$, we obtain $A(\phi_j, \psi) = 0$, as desired. This implies that $\psi \in Y_{j+1}$ so that

$$\alpha = \frac{A(\psi, \psi)}{\ell(\psi, \psi)} \ge \inf_{u \in Y_j, u \neq 0} J(u) = \beta_j \quad \text{for all } j,$$

which is impossible because $\beta_j \to \infty$. Therefore we may let k be

$$k = \max\{j : \beta_j = \alpha\}.$$

We still have $A(\psi, \phi_j) = 0$ for all j with $\beta_j \neq \alpha$. If we also have $A(\psi, \phi_j) = 0$ for all j with $\beta_j = \alpha$, then

$$\alpha = J(\psi) \ge \inf_{u \in Y_{k+1}, u \neq 0} J(u) = \beta_{k+1} > \beta_k \quad \text{(a contradiction)}.$$

We now see that ψ, which was taken to be an *arbitrary* nonzero element of the eigenspace for α, is such that $A(\psi, \phi_j)$ for some j with $\beta_j = \alpha$. This implies that the eigenspace for α is in fact equal to $\operatorname{span}\{\phi_j : \beta_j = \alpha\}$.

The last part of the theorem on the symmetric and antisymmetric eigenfunctions is proved analogously by replacing X by X^{sym} and X^{ant} and using the fact that these two spaces are orthogonal with respect to the sesquilinear form $A(\cdot, \cdot)$. There are no essential changes in the proof. ∎

The form ℓ depends on the parameter ϵ_1; we make this dependence explicit by introducing the variable ϵ:

$$\ell_\epsilon(u, v) = \int_{\mathcal{R} \setminus \Omega} \epsilon_0 \omega^2 u \bar{v} + \int_\Omega \epsilon \omega^2 u \bar{v}, \tag{48}$$

$$J_\epsilon(u) = \frac{A(u, u)}{\ell_\epsilon(u, u)} = \frac{\int_{\mathcal{R}} \frac{1}{\mu} |(\nabla + i\boldsymbol{\kappa}) u|^2 + \frac{1}{\mu_0} \int_\Gamma (BTu)(T\bar{u})}{\epsilon_0 \omega^2 \int_{\mathcal{R} \setminus \Omega} |u|^2 + \epsilon \omega^2 \int_\Omega |u|^2}. \tag{49}$$

The eigenvalues and eigenfunctions also depend on ϵ, and we denote them by $\alpha_j(\epsilon)$ and $\psi_j(\epsilon)$. Normalizing the eigenfunctions so that $\ell_\epsilon(\psi_j(\epsilon), \psi_j(\epsilon)) = 1$, we have

$$\ell_\epsilon(\psi_j(\epsilon), \psi_k(\epsilon)) = \delta_{jk}, \qquad A(\psi_j(\epsilon), \psi_k(\epsilon)) = 0 \quad \text{for } j \neq k.$$

The compact operator $K = K_\epsilon$ also depends on ϵ,

$$A(K_\epsilon f, v) = \ell_\epsilon(f, v) \quad \text{for all } v \in X, \tag{50}$$

and the eigenvalues of K_ϵ are $\{\alpha_j(\epsilon)^{-1}\}_{j=0}^\infty$ with corresponding eigenvectors $\{\psi_j(\epsilon)\}_{j=0}^\infty$.

LEMMA 6.2. *K_ϵ is continuous in ϵ with respect to the operator norm on K_ϵ.*

Proof. Let $\epsilon_1 > 0$ be given. For an arbitrary variation $\Delta\epsilon > 0$ with $0 < |\Delta\epsilon| < \epsilon_1$, set

$$\Delta K = K_{\epsilon_1+\Delta\epsilon} - K_{\epsilon_1} \quad \text{and} \quad \Delta\ell = \ell_{\epsilon_1+\Delta\epsilon} - \ell_{\epsilon_1}.$$

Applying the defining relation (50) for K_ϵ to $K_{\epsilon_1+\Delta\epsilon}$ and K_{ϵ_1}, with $v = \Delta K f$, and subtracting yields the relation

$$A(\Delta K f, \Delta K f) = \Delta\ell(f, \Delta K f). \tag{51}$$

We have the following lower estimate for the left-hand side of (51),

$$\delta\|\Delta K f\|_{L^2}^2 \leq \delta\|\Delta K f\|_{H^1}^2 \leq A(\Delta K f, \Delta K f),$$

and upper estimate for the right-hand side:

$$|\Delta\ell(f, \Delta K f)| = |\Delta\epsilon| \left| \int_\Omega f \Delta K \bar{f} \right| \leq |\Delta\epsilon| \, \|f\|_{L^2} \, \|\Delta K f\|_{L^2}.$$

Putting these inequalities together, we obtain

$$\|\Delta K f\|_{L^2} \leq \frac{|\Delta\epsilon|}{\delta} \|f\|_{L^2},$$

so that $\|\Delta K\| \leq |\Delta\epsilon|/\delta$. ∎

We now prove the lemma of section 3 that states that the eigenvalues α are continuous strictly decreasing functions of ϵ_1, and $\alpha_j \to 0$ as $\epsilon_1 \to \infty$. A similar result was stated for the μ_1 dependency, but we omit the proof in that case, as it is similar.

Proof of Lemma 3.2. By Lemma 6.2, K_ϵ is continuous in ϵ with respect to the operator norm on K_ϵ, and its spectrum is the set $\{\alpha_j(\epsilon)^{-1}\}_{j=0}^\infty$. Because the eigenvalues of compact operators are continuous functions of the operator in the operator norm (Kato [12, Chapter IV, section 3.5]), we conclude that the functions $\alpha_j(\epsilon)$ are continuous functions of ϵ.

To prove that $\alpha_j(\epsilon)$ is strictly decreasing in ϵ, let ϵ_1 and ϵ_2 be given with $0 < \epsilon_1 < \epsilon_2$, and let an integer $N \geq 0$ be given. Define

$$V_N = \text{span}\{\psi_j(\epsilon_1) : 0 \leq j \leq N\}$$

($V_0 = \{0\}$), in which the eigenvectors $\psi_j(\epsilon_1)$ are orthonormal with respect to $\ell_{\epsilon_1}(\cdot,\cdot)$ and orthogonal with respect to $A(\cdot,\cdot)$:

$$\ell_{\epsilon_1}(\psi_j(\epsilon_1), \psi_k(\epsilon_1)) = \delta_{jk}, \qquad A(\psi_j(\epsilon_1), \psi_k(\epsilon_1)) = 0 \quad \text{for } j \neq k.$$

For each $\psi \in V_N$ with $\ell_{\epsilon_1}(\psi,\psi) = 1$, there are numbers a_j, for $0 \le j \le N$, such that

$$\psi = \sum_{j=0}^{N} a_j \psi_j(\epsilon_1), \qquad \sum_{j=0}^{N} |a_j|^2 = 1,$$

and we obtain $A(\psi,\psi) = \sum_{j=0}^{N} |a_j|^2 A(\psi_j(\epsilon_1), \psi_j(\epsilon_1))$ so that

$$A(\psi,\psi) = J_{\epsilon_1}(\psi) = \sum_{j=0}^{N} |a_j|^2 J_{\epsilon_1}(\psi_j(\epsilon_1)) = \sum_{j=0}^{N} |a_j|^2 \alpha_j(\epsilon_1) \le \alpha_N(\epsilon_1). \tag{52}$$

From the definition of J_ϵ, it is evident that $J_{\epsilon_2}(\phi) \le J_{\epsilon_1}(\phi)$ for each $\phi \in X$; however, we need to show strict inequality for $\phi \in V_N$, which requires showing that, for each nonzero $\phi \in V_N$, it is not true that ϕ is equal to zero almost everywhere on Ω. To this end, let

$$\phi = \sum_{j=0}^{N} b_j \psi_j(\epsilon_1) = 0 \quad \text{a.e. in } \Omega.$$

Set $\beta_j = \alpha_j(\epsilon_1)\epsilon_1\mu_1\omega^2$. As the $\psi_j(\epsilon_1)$ are smooth in Ω, for each $k = 0, \dots, N$, we have

$$0 = \prod_{\beta_j \ne \beta_k} \left((\nabla + i\boldsymbol{\kappa})^2 + \beta_j \right) \phi = \prod_{\beta_j \ne \beta_k} (-\beta_k + \beta_j) \sum_{\beta_j = \beta_k} b_j \psi_j(\epsilon_1) \quad \text{in } \Omega.$$

Since $\prod_{\beta_j \ne \beta_k} (-\beta_k + \beta_j) \ne 0$, we obtain $\sum_{\beta_k = \beta_h} b_k \psi_k(\epsilon_1) = 0$ in Ω. However, $\sum_{\beta_j = \beta_k} b_j \psi_j(\epsilon_1)$ satisfies Condition 2.2, and therefore $\sum_{\beta_j = \beta_k} b_j \psi_j(\epsilon_1)$ is zero in $\mathcal{R}$. As the $\psi_j(\epsilon_1)$ are linearly independent, we infer that $b_j = 0$ for j such that $\beta_j = \beta_k$. As k was chosen arbitrarily from $\{0, \dots, N\}$, we obtain $b_j = 0$ for $0 \le j \le N$. We conclude that $\psi = \sum_{j=0}^{N} a_j \psi_j(\epsilon_1)$ from above is not zero in $L^2(\Omega)$. It follows now from the definitions of J_ϵ and ℓ_ϵ and from (52) that $J_{\epsilon_2}(\psi) < J_{\epsilon_1}(\psi) \le \alpha_N(\epsilon_1)$, and by the homogeneity of J_{ϵ_2} we obtain

$$J_{\epsilon_2}(\phi) < J_{\epsilon_1}(\phi) \le \alpha_N(\epsilon_1) \quad \text{for all } \phi \in V_N. \tag{53}$$

Define, for each $\epsilon > 0$,

$$X_N(\epsilon) = \{ v \in X : A(\psi_j(\epsilon), v) = 0 \text{ for } j = 0, \dots, N-1 \}.$$

The dimension of $V_N \cap X_N(\epsilon_2)$ is at least 1; let ϕ be a nonzero vector in this intersection. We obtain

$$\alpha_N(\epsilon_2) = \inf_{u \in X_N(\epsilon_2), u \ne 0} J_{\epsilon_2}(u) \le J_{\epsilon_2}(\phi) < \alpha_N(\epsilon_1),$$

and we have proved that $\alpha_N(\epsilon)$ is a decreasing function of ϵ.

We now prove that $\alpha_N(\epsilon)$ tends to zero as ϵ tends to infinity. We define the set

$$S = \{ \psi \in V_N : \ell_{\epsilon_1}(\psi,\psi) = 1 \}.$$

S is compact in $L^2(\mathcal{R}, \ell_{\epsilon_1})$ and therefore also in $L^2(\mathcal{R})$. Since $\int_\Omega |\psi|^2$ is continuous in $L^2(\mathcal{R})$ and $\int_\Omega |\psi|^2 \ne 0$ for all $\psi \in S$, there is a number M such that

$$0 < M < \int_\Omega |\psi|^2 \quad \text{for all } \psi \in S.$$

Therefore

$$\ell_\epsilon(\psi,\psi) \geq \epsilon\omega^2 \int_\Omega |\psi|^2 > \epsilon\omega^2 M \quad \text{for all } \psi \in S,$$

and, using (52) for $\epsilon > \epsilon_1$,

$$J_\epsilon(\psi) = \frac{A(\psi,\psi)}{\ell_\epsilon(\psi,\psi)} < \frac{\alpha_N(\epsilon_1)}{\epsilon\omega^2 M} \quad \text{for all } \psi \in S.$$

The dimension of $V_N \cap X_N(\epsilon)$ is at least 1. Let ϕ be a nonzero vector in this intersection, which we may take to be in S. We then obtain

$$\alpha_N(\epsilon) = \inf_{u \in X_N(\epsilon), u \neq 0} J_\epsilon(u) \leq J_\epsilon(\phi) < \frac{\alpha_N(\epsilon_1)}{\epsilon\omega^2 M}. \qquad \square \tag{54}$$

The estimate (54) shows that the eigenvalues decay at least proportionally to $1/\epsilon$.

REFERENCES

[1] S. G. Tikhodeev, A. L. Yablonskii, E. A. Muljarov, N. A. Gippius, and T. Ishihara, *Quasiguided modes and optical properties of photonic crystals slabs*. Phys. Rev. B, 66 (2002), paper 045102

[2] S. P. Shipman and S. Venakides, *Resonant transmission near non-robust periodic slab modes*, Phys. Rev. E, 71 (2005), paper 026611.

[3] A. Krishnan, T. Thio, T. J. Kim, H. J. Lezec, T. W. Ebbesen, P. A. Wolff, J. Pendry, L. Martin-Moreno, and F. J. Gar ía-Vidal, *Evanescently coupled resonance in surface plasmon enhanced transmission*, Opt. Comm., 200 (2001), pp. 1–7.

[4] J. A. Porto, F. J. Garc ía-Vidal, and J. B. Pendry, *Transmission resonances on metallic gratings with very narrow slits*, Phys. Rev. Lett., 84 (1999), pp. 2845–2848.

[5] J. Broeng, D. Mogilevstev, S. E. Barkou, and A. Bjarklev, *Photonic crystal fibers: A new class of optical waveguides*, Optical Fiber Tech., 5 (1999), pp. 305–330.

[6] A.-S. Bonnet-Bendhia and F. Starling, *Guided waves by electromagnetic gratings and nonuniqueness examples for the diffraction problem*, Math. Methods Appl. Sci., 17 (1994), pp. 305–338.

[7] J. Jost, *Partial Differential Equations*, Springer, New York, 2002.

[8] S. H. Gould, *Variational Methods for Eigenvalue Problems*, University of Toronto Press, Toronto, 1957.

[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, Heidelberg, 2001.

[10] G. A. Kriegsmann, *The Galerkin approximation of the iris problem: Conservation of power*, Appl. Math. Lett., 10 (1997), pp. 41–44.

[11] D. Volkov and G. A. Kriegsmann, *Scattering by a perfect conductor in a waveguide: Energy preserving schemes for integral equations*, IMA J. Appl. Math., 71 (2006), pp. 898–923.

[12] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, Berlin, 1995.

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 714–730

ESTIMATES FOR ELECTRIC FIELDS BLOWN UP BETWEEN CLOSELY ADJACENT CONDUCTORS WITH ARBITRARY SHAPE*

KIHYUN YUN†

Abstract. It may be well known in practice that high stress concentrations occur in fiber-reinforced composites. There have been several works by analysis to estimate for the stresses between closed spaced fibers. However, the known results on stiff fibers have until now been restricted to the particular case of circular cross-sections. Thus, we extend the blow-up results on the stresses specialized only for disks to the general case of arbitrary shapes. Moreover, we prove that the blow-up rate of the general case is exactly the same as that of disks. Nevertheless, from the viewpoint of methodology, the technique we use is significantly different from the previous one restricted to the case of disks. Referring to antiplane shear problems, these works are reduced to the gradient estimates for the solution to the conductivity problem containing two closely spaced conductors which represent the cross-sections of fibers. We establish a novel representation for the solution on conductors by a probability function. Based on this, the general blow-up results are derived by a simpler method.

Key words. gradient estimates, blow-up, arbitrary shape, conductivity problems, stresses, composite materials

AMS subject classifications. 35J25, 73C40

DOI. 10.1137/060648817

1. Introduction.

This paper is concerned with high stress concentrations between closely spaced stiff fibers in an infinite matrix. According to Budiansky and Carrier [8], unexpectedly low strengths in longitudinal shear have been reported for brittle-matrix, fiber-reinforced composites, and it has been suggested that this might be explained by high stress concentrations between neighboring fibers (see also [5, 9]). However, according to Keller [12], it is difficult to calculate numerically the stresses in a narrow region because the stresses are much larger than elsewhere. Several approaches by analysis have been developed, but the blow-up results on the stresses are restricted to the particular case where fibers have circular cross-sections. Until now there has not been any established result associated with a large class of shapes. This paper presents the blow-up result for a class of shapes which is general enough. Moreover, the blow-up rate is exactly the same as the one for disks.

We consider two parallel elastic fibers embedded in an infinite elastic matrix. We suppose that D_1 and D_2 are very closely spaced inclusions in $\mathbb{R}^2$ which are ϵ apart, representing the cross-sections of the fibers, and the shear moduli of the inclusions are constants a_1 and a_2, different from the constant outside shear modulus 1. Referring to a problem of antiplane shear, we get the following conductivity equation for a given harmonic function H in $\mathbb{R}^2$:

$$\begin{cases} \nabla \cdot \left\{ \left(1 + \sum_{i=1,2} (a_i - 1)\chi(D_i) \right) \nabla u \right\} = 0, \\ u(x) - H(x) = O(|x|^{-1}) \ \text{ as } |x| \to \infty, \end{cases} \tag{1.1}$$

*Received by the editors January 1, 2006; accepted for publication (in revised form) November 14, 2006; published electronically March 2, 2007. This work was supported by the Korea Research Foundation grant KRF-2005-214-C00184 funded by the Korean Government (MOEHRD).
http://www.siam.org/journals/siap/67-3/64881.html
†Center for Nonlinear Analysis and Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (kyun@andrew.cmu.edu).

where the function u represents the out-of-plane elastic displacement. For applications to the composite materials, our work focuses on the stresses, represented by ∇u, particularly in the case when ∇H is a uniform field, i.e., $H(x) = A \cdot x$ for some constant vector A. The question of interest is to establish the optimal estimate on $|\nabla u|$ as the separation distance ϵ approaches 0.

We give a brief description of remarkable works by analysis on gradient estimates for solutions. For finite and strictly positive shear moduli (or conductivities) a_1 and a_2, it has been shown by Bonnetier and Vogelius in [7] that $|\nabla u|$ remains bounded for circular touching inclusions with comparable radii. Li and Vogelius derived in [15] a uniform upper bound on $|\nabla u|$ that is independent of the distance ϵ between D_1 and D_2, assuming that the moduli a_1 and a_2 stay away from 0 and ∞. It may be noted that this result of [15] holds for an arbitrary number of inclusions with arbitrary shape and is not restricted to two-dimensional space. Moreover, this result has been extended to elliptic systems by Li and Nirenberg in [14].

However, to explain high stresses occurring between stiff fibers, we should pay attention to the case of the extreme valued shear moduli. For two circular inclusions with $a_1 = a_2 = \infty$ or $a_1 = a_2 = 0$, it has been shown by Ammari, Kang, and Lim [2] and Ammari et al. [4] (see also [8]) that $|\nabla u|$ blows up as the distance ϵ approaches 0 for a special uniform field ∇H. Moreover, the optimal rate of blow-up is $\epsilon^{-1/2}$. These results on blow-up are specialized only for the case of circular inclusions. Thus there has been a strong need for a result that is not only associated with a large class of shapes but also has the same blow-up rate as circular inclusions.

In this paper we present the desirable result: for two inclusions D_1 and D_2 whose shapes are arbitrary enough, $|\nabla u|$ blows up as the distance ϵ approaches 0 for a special uniform field ∇H, and the blow-up rate is exactly $\epsilon^{-1/2}$, which is the known rate in the circular cases.

We now proceed to state the main results of this paper. To do so we need to make our notation and assumption more precise. Let D_{right} be a bounded domain in $\mathbb{R}^+ \times \mathbb{R}$ that is strictly convex at the unique left endpoint $(0,0)$ of this domain, and let D_{left} be a bounded domain in $\mathbb{R}^- \times \mathbb{R}$ that has a right endpoint at $(0,0)$ and a C^2 boundary. In addition, we assume that $\varphi : \mathbb{C}\backslash B_1(0) \to \mathbb{R}^2\backslash D_{\text{right}}$ is a conformal mapping such that $\varphi \in C^2(\mathbb{C}\backslash B_1(0))$ and $\varphi'(z) \neq 0$ for $z \in \partial B_1(0)$ (refer to the Riemann mapping theorem in [1]). We shall not distinguish between $\mathbb{R}^2$ and $\mathbb{C}$ in this paper. Let the domain D_1 and D_2 be as follows:

$$D_1 = D_{\text{right}} + \frac{1}{2}\epsilon \quad \text{and} \quad D_2 = D_{\text{left}} - \frac{1}{2}\epsilon.$$

To consider the case when $a_1 = a_2 = \infty$, given any harmonic function H in $\mathbb{R}^2$, (1.1) is rewritten in the following form:

$$\begin{cases} \triangle u = 0 \quad \text{in } \mathbb{R}^2\backslash\overline{(D_1 \cup D_2)}, \\ u(x) - H(x) = O(|x|^{-1}) \quad \text{as } |x| \to \infty, \\ u|_{\partial D_i} = C_i \text{ (constant), and} \\ \int_{\partial D_i} \partial_\nu u \, ds = 0 \quad \text{for } i = 1, 2. \end{cases} \tag{1.2}$$

The solution u can also be interpreted physically as the voltage potential outside uncharged conductors D_1 and D_2 under the action of applied electric field ∇H (see [10]).

THEOREM 1.1. *Assume the conditions above. Let u be the solution to* (1.2) *for $H(x_1, x_2) = x_1$. If the distance ϵ is sufficiently small, then there exists a strictly*

positive constant C_* independent of ϵ such that

$$|u|_{\partial D_1} - u|_{\partial D_2}| \geq C_*\sqrt{\epsilon} \tag{1.3}$$

and, owing to the mean value theorem,

$$\max_{-\frac{\epsilon}{2}<x_1<\frac{\epsilon}{2}} |\nabla u(x_1, 0)| \geq C_* \left(\frac{1}{\sqrt{\epsilon}}\right). \tag{1.4}$$

As mentioned earlier, we extend the result known only for disks to the general case of inclusions with arbitrary shape. Nevertheless, the technique we use is significantly different from the previous one. In the case of disks, the authors took advantage of Kelvin transform and properties of layer potential which are specialized only for circles in [2, 4, 8]. It is difficult to apply this method to other shaped inclusions even though they are ellipses. Thus, we need to provide a new method for the general case. To do so, we establish a new and easy representation for the difference $u|_{\partial D_1} - u|_{\partial D_2}$ in Lemma 2.3 by a probability function $\partial_\nu w$. Based on this, the blow-up result (1.4) can be derived by a method simpler than the previous ones such as asymptotic expansions related to discontinuous conductivity in [2, 3, 4, 6, 11, 16]. On the other hand, applying the new representation to the recent result of Ammari et al. in [4], we can derive the optimal upper bound of $|\nabla u|$ in Theorem 1.2.

THEOREM 1.2. *Let H and u be the same as in Theorem* 1.1*. In addition, we assume that D_{left} also has a conformal mapping $\psi : \mathbb{C}\backslash B_1(0) \to \mathbb{R}^2\backslash D_{\text{left}}$ with the same regularity conditions as φ. If the distance ϵ is sufficiently small, then we have constants C_1^* and C_2^* independent of ϵ such that*

$$|u|_{\partial D_1} - u|_{\partial D_2}| \leq C_1^*\sqrt{\epsilon} \tag{1.5}$$

and

$$\|\nabla u\|_{L^\infty(\mathbb{R}^2\backslash(D_1\cup D_2))} \leq C_2^* \left(\frac{1}{\sqrt{\epsilon}}\right). \tag{1.6}$$

To summarize the two theorems, Theorem 1.1 provides that $|\nabla u|$ blows up as the distance $\epsilon \to 0$ for a special uniform field and the blow-up rate is not less than $\epsilon^{-1/2}$. It follows from the bound (1.6) in Theorem 1.2 that the rate is exactly $\epsilon^{-1/2}$.

We now consider the case when $a_1 = a_2 = 0$. It is probably relevant to consider the case when $a_1 = a_2 = 0$, because the fibers are there for reinforcement. However, the solution to this case can be interpreted physically as the voltage potential outside nonconductors D_1 and D_2 under the action of applied electric field $\nabla\widetilde{H}$ (see [13]). For any given harmonic function $\widetilde{H}$ in $\mathbb{R}^2$, let $\widetilde{u}$ be the unique solution to the following Neumann problem:

$$\begin{cases} \triangle\widetilde{u} = 0 \quad \text{in } \mathbb{R}^2\backslash\overline{(D_1 \cup D_2)}, \\ \widetilde{u}(x) - \widetilde{H}(x) = O(|x|^{-1}) \quad \text{as } |x| \to \infty, \\ \partial_\nu\widetilde{u} = 0 \ \text{ on } \partial D_i \quad \text{for } i = 1, 2, \end{cases} \tag{1.7}$$

where $\partial_\nu\widetilde{u}$ is the normal derivative of $\widetilde{u}$.

THEOREM 1.3. *Assume that D_{left} is the same as in Theorem* 1.2*, and that C_* and C_2^* are the constants used in Theorems* 1.1 *and* 1.2*, respectively. Let $\widetilde{u}$ be the*

solution to (1.7) *for* $\widetilde{H}(x_1, x_2) = x_2$. *If the distance* ϵ *is small enough, then we have the estimates*

$$\max_{-\frac{\epsilon}{2}<x_1<\frac{\epsilon}{2}} |\nabla \widetilde{u}(x_1, 0)| \geq C_* \left(\frac{1}{\sqrt{\epsilon}}\right) \tag{1.8}$$

and

$$\|\nabla \widetilde{u}\|_{L^\infty(\mathbb{R}^2\setminus(D_1\cup D_2))} \leq C_2^* \left(\frac{1}{\sqrt{\epsilon}}\right). \tag{1.9}$$

Proof. Let u be the voltage potential for $H = x_1$ in Theorem 1.1. By Lemma 2.1, we have

$$u = H + V_1 + V_2,$$

where V_i is a harmonic function in $\mathbb{R}^2\setminus D_i$ and $V_i(x) = O(|x|^{-1})$ as $|x| \to \infty$ for $i = 1, 2$. Owing to Poincaré's theorem, we have a well-defined conjugate harmonic function of V_i, denoted by $\widetilde{V_i}$, for $i = 1, 2$ such that

$$\widetilde{V_1}(x) + \widetilde{V_2}(x) = O(|x|^{-1}) \quad \text{as } |x| \to \infty.$$

Let $\widetilde{u} = \widetilde{H} + \widetilde{V_1} + \widetilde{V_2}$. Then $\widetilde{u}$ satisfies (1.7) and is also a harmonic conjugate function of u. Hence, we have

$$|\nabla u| = |\nabla \widetilde{u}| \ \text{ in } \mathbb{R}^2\setminus\overline{(D_1 \cup D_2)}.$$

Therefore, we have completed the proof. ∎

2. Proof of Theorem 1.1. In this section we will give a proof of the inequality (1.3). The proof is based on (2.6) and Lemma 2.3 which present an interesting representation for the difference $u|_{\partial D_1} - u|_{\partial D_2}$ by a probability function $\partial_\nu w$. Thus we choose a constant C satisfying the inequality $\partial_\nu w \geq C\sqrt{\epsilon}$ (2.15). This inequality completes the proof.

We start by representing the voltage potential u as a function related to H. To do so, we define the operator $R_1 : C^\infty(\mathbb{R}^2\setminus\overline{D_2}) \to C^\infty(\mathbb{R}^2\setminus\overline{D_1}) \cap C(\mathbb{R}^2\setminus D_1)$ as

$$\begin{cases} \triangle R_1(v) = 0 \quad \text{in } \mathbb{R}^2\setminus\overline{D_1}, \\ R_1(v)(x) = O(|x|^{-1}) \quad \text{as } \ |x| \to \infty, \\ (v - R_1(v))|_{\partial D_1} = C \text{ (constant)}, \end{cases} \tag{2.1}$$

where C is a constant dependent on v, and we also define $R_2 : C^\infty(\mathbb{R}^2\setminus\overline{D_1}) \to C^\infty(\mathbb{R}^2\setminus\overline{D_2}) \cap C(\mathbb{R}^2\setminus D_2)$ similarly. It follows from Green's theorem that

$$\int_{\partial_{D_1}} \frac{\partial R_1(v)}{\partial \nu}\, ds = 0.$$

By definition, $H - R_1(H)$ can be interpreted physically as the voltage potential due only to the presence of D_1, under the action of applied electric field ∇H. Since the voltage potential u is due not only to D_1 but also to D_2, we take advantage of R_2. We thus expect $u \sim H - R_1(H) - R_2(H)$. But since $H - R_1(H) - R_2(H)$ is not constant on the boundaries ∂D_1 and ∂D_2, we expect $u \sim H - R_1(H) - R_2(H) + R_2R_1(H) + R_1R_2(H)$ again. These steps can proceed inductively. The process provides the following lemma.

LEMMA 2.1. *For a harmonic function H defined in $\mathbb{R}^2$, the voltage function u is represented as follows:*

$$\begin{aligned} u &= H - R_1(H) - R_2(H) + R_1R_2(H) + R_2R_1(H) \\ &\quad - R_1R_2R_1(H) - R_2R_1R_2(H) + \cdots \\ &= H - R_1(H) - R_2(H) + R_1R_2(H) + R_2R_1(H) \\ &\quad + \sum_{n=1}^{\infty} (R_1R_2)^n(-R_1(H) + R_1R_2(H)) + (R_2R_1)^n(-R_2(H) + R_2R_1(H)). \end{aligned} \tag{2.2}$$

Proof. We choose an interior point p of D_1. Let $\Omega = \{\frac{x-p}{|x-p|^2} \mid x \in \mathbb{R}^2 \backslash D_1\} \cup \{0\}$ and $\Omega_\epsilon = \{\frac{x-p}{|x-p|^2} \mid x \in D_2\} \cup \{0\}$. Then we have $\Omega_\epsilon \subset \Omega$ and the distance $\mathrm{d}(\Omega_\epsilon, \partial\Omega) > 0$. Hence, by the maximum principle and standard estimates, one can choose a positive constant $c < 1$ such that

$$\max_{\Omega_\epsilon} h - \min_{\Omega_\epsilon} h \leq c \left(\max_{\Omega} h - \min_{\Omega} h \right)$$

for any harmonic function h defined in Ω. It follows that

$$\max_{\partial D_2} R_2R_1(v) - \min_{\partial D_2} R_2R_1(v) \leq c \left(\max_{\partial D_1} R_1 v - \min_{\partial D_1} R_1 v \right) \tag{2.3}$$

and, since $R_1(v)(x) = O(|x|^{-1})$ as $|x| \to \infty$, we have

$$\|R_1(v)\|_{L^\infty} \leq \max_{\partial D_1} R_1(v) - \min_{\partial D_1} R_1(v) \tag{2.4}$$

for any $v \in C^\infty(\mathbb{R}^2 \backslash D_2)$. And we can obtain similar results for D_2 and R_2. Since $0 < c < 1$, the expansion (2.2) is well defined and satisfies (1.2). □

In the particular case of circular inclusions, it follows from Lemma 2.1 and Kelvin transforms that the main result (1.4) holds (see [2]). However, it is not easy to apply the asymptotic expansion (2.2) directly to the general case of arbitrary shape. On this account, we would make the expansion (2.2) simpler. We define the operator $K_1 : C^\infty(\mathbb{R}^2 \backslash \overline{D_2}) \to C^\infty(\mathbb{R}^2 \backslash \overline{D_1}) \cap C(\mathbb{R}^2 \backslash D_1)$ as follows:

$$\begin{cases} \triangle K_1(v) = 0 \quad \text{in } \mathbb{R}^2 \backslash \overline{D_1}, \\ K_1(v)(x) \text{ converges to some constant as } \ |x| \to \infty, \\ v = K_1(v) \text{ on } \partial D_1. \end{cases}$$

And we also define K_2 similarly. By the definitions of R_i and K_i $(i = 1, 2)$, we have

$$R_i(v) = K_i(v) + \text{some constant},$$

where the constant is dependent on v.

LEMMA 2.2. *We have*

$$\begin{aligned} u|_{\partial D_1} &= C_0 + \lim_{n\to\infty} (K_2K_1)^n(H)|_{\partial D_1}, \\ u|_{\partial D_2} &= C_0 + \lim_{n\to\infty} (K_1K_2)^n(H)|_{\partial D_2}, \end{aligned} \tag{2.5}$$

where C_0 is a constant.

Proof. Since $K_i(v) = v$ on ∂D_i for $i = 1, 2$, we have

$$\begin{aligned}&(R_1R_2)^n(-R_1(H) + R_1R_2(H)) + (R_2R_1)^n(-R_2(H) + R_2R_1(H))\\ &= c_n + (K_1K_2)^n(-K_1(H) + K_1K_2(H)) + (K_2K_1)^n(-K_2(H) + K_2K_1(H))\end{aligned}$$

for $n = 0, 1, 2, 3, \ldots$, where

$$\begin{aligned}c_n = &\{-(R_1R_2)^nR_1(H) + (R_2R_1)^nR_2R_1(H)\}\,|_{\partial D_2}\\ &+ \{-(R_2R_1)^nR_2(H) + (R_1R_2)^nR_1R_2(H)\}\,|_{\partial D_1}.\end{aligned}$$

By (2.3) and (2.4), we have $\sum_{n=0}^{\infty} |c_n| < \infty$. We thus set $C_0 = \sum_{n=0}^{\infty} c_n$. Then

$$\begin{aligned}u &= H - R_1(H) - R_2(H) + R_1R_2(H) + R_2R_1(H)\\ &\quad + \sum_{n=1}^{\infty}(R_1R_2)^n(-R_1(H) + R_1R_2(H)) + (R_2R_1)^n(-R_2(H) + R_2R_1(H))\\ &= C_0 + H - K_1(H) - K_2(H) + K_1K_2(H) + K_2K_1(H)\\ &\quad + \sum_{n=1}^{\infty}(K_1K_2)^n(-K_1(H) + K_1K_2(H)) + (K_2K_1)^n(-K_2(H) + K_2K_1(H))\\ &= C_0 + \{H - K_1(H)\} + \{-K_2(H) + K_1K_2(H)\} + K_2K_1(H) + \cdots\\ &= C_0 + \lim_{n\to\infty}(K_2K_1)^n(H) \text{ on } \partial D_1. \qquad \square\end{aligned}$$

Hence we conclude that

$$u|_{\partial D_1} - u|_{\partial D_2} = \lim_{n\to\infty}(K_2K_1)^n(H)|_{\partial D_1} - \lim_{n\to\infty}(K_1K_2)^n(H)|_{\partial D_2}. \tag{2.6}$$

In what follows, we consider

$$\lim_{n\to\infty}(K_2K_1)^n(H)|_{\partial D_1}.$$

Now we present an interesting representation for $\lim_{n\to\infty}(K_2K_1)^n(H)|_{\partial D_1}$ in the following lemma. Based on this, the main result would be derived without any asymptotic analysis and layer potentials.

LEMMA 2.3. *Let w be the solution of the following problem:*

$$\begin{cases} \triangle w = 0 \quad \textit{in } \mathbb{R}^2\backslash\overline{(D_1 \cup D_2)},\\ w(x) = O(|x|^{-1}) \quad \textit{as} \quad |x| \to \infty,\\ w|_{\partial D_1} = c_{1\epsilon} \textit{ (constant)},\\ w|_{\partial D_2} = c_{2\epsilon} \textit{ (constant)},\\ \int_{\partial D_1} \partial_\nu w \, ds = 1. \end{cases} \tag{2.7}$$

Then we have

$$\lim_{n\to\infty}(K_2K_1)^n(H)|_{\partial D_1} = \int_{\partial D_1}(\partial_\nu w)H ds.$$

Proof. By definition, we have

$$H = K_1(H) \quad \text{on } \partial D_1.$$

Since $K_1(H)$ is harmonic in D_2 and w is constant on ∂D_2, we have

$$\int_{\partial D_2} w\partial_\nu K_1(H)ds = 0.$$

And since $K_1(H)$ converges to some constant as $|x| \to \infty$ and w is constant on ∂D_1, we also have

$$\int_{\partial D_1} w\partial_\nu K_1(H)ds = 0.$$

It follows from Green's theorem that

$$\int_{\partial D_1} (\partial_\nu w)Hds + \int_{\partial D_2} (\partial_\nu w)K_1(H)ds = 0.$$

Similarly we have

$$\int_{\partial D_2} (\partial_\nu w)K_1(H)ds + \int_{\partial D_1} (\partial_\nu w)K_2K_1(H)ds = 0.$$

Hence we obtain

$$\begin{aligned}
\int_{\partial D_1} (\partial_\nu w)Hds &= \int_{\partial D_1} (\partial_\nu w)K_2K_1(H)ds \\
&= \int_{\partial D_1} (\partial_\nu w)(K_2K_1)^n(H)ds \quad \text{for} \quad n = 1, 2, 3, \dots \\
&= \int_{\partial D_1} (\partial_\nu w) \lim_{n\to\infty} (K_2K_1)^n(H)ds \\
&= \lim_{n\to\infty} (K_2K_1)^n(H). \qquad \square
\end{aligned}$$

Thus we focus on $\partial_\nu w$. We define $\varphi_1 : \mathbb{C}\backslash B_1(0) \to \mathbb{R}^2\backslash D_1$ as $\varphi_1 = \varphi + \frac{\epsilon}{2}$ and the conformal mapping $\Phi : \overline{B_1(0)}\backslash\{0\} \to \mathbb{R}^2\backslash D_1$ as

$$\Phi(z) = \varphi_1\left(\frac{z}{|z|^2}\right),$$

where $\varphi : \mathbb{C}\backslash B_1(0) \to \mathbb{R}^2\backslash D_{\text{right}}$ is the conformal mapping defined in the introduction. Then we have

$$\Phi^{-1}\left(\mathbb{R}^2\backslash\overline{(D_1 \cup D_2)}\right) \subset B_1(0)$$

and

$$\Phi^{-1}(\partial D_1) = \partial B_1(0).$$

We consider the solution W to the following Dirichlet problem:

$$\begin{cases} \triangle W = 0 \quad \text{in } B_1(0)\backslash\overline{\Phi^{-1}(D_2)}, \\ W = 1 \quad \text{on } \Phi^{-1}(\partial D_1) = \partial B_1(0), \\ W = -1 \quad \text{on } \Phi^{-1}(\partial D_2). \end{cases} \tag{2.8}$$

Let $M = \frac{1}{2}(c_{1\epsilon} - c_{2\epsilon})$ and $c_* = \frac{1}{2}(c_{1\epsilon} + c_{2\epsilon})$. Then, even though $w(\Phi(z))$ is not defined at 0, it follows from the decreasing behavior of w at infinity that

$$\begin{cases} w(\Phi(z)) = MW(z) + c_* \quad \text{for } z \in \Phi^{-1}\left(\mathbb{R}^2 \backslash \overline{(D_1 \cup D_2)}\right), \\ \partial_{\nu_{(x)}} w(x) = \left(\frac{M}{|\varphi'(z)|}\right) \partial_{\nu_{(z)}} W(z) \quad \text{for } z \in \partial B_1(0), \end{cases} \tag{2.9}$$

where $x = \Phi(z)$.

Without loss of generality, we assume that

$$B_{r_2}(r_2 + \epsilon - 1) \subset \Phi^{-1}(D_2) \subset B_{r_1}(r_1 + \epsilon - 1), \tag{2.10}$$

where r_1 and r_2 are independent of ϵ for any sufficiently small $\epsilon \geq 0$. (See Lemma 4.1 in the appendix for details.) Then we consider the solution U_1 and U_2 to the following equations:

$$\begin{cases} \triangle U_i = 0 \quad \text{in } B_1(0) \backslash \overline{B_{r_i}(r_i + \epsilon - 1)}, \\ U_i = 1 \quad \text{on } \Phi^{-1}(\partial D_1) = \partial B_1(0), \\ U_i = -1 \quad \text{on } \partial B_{r_i}(r_i + \epsilon - 1) \ \text{ for } i = 1, 2. \end{cases} \tag{2.11}$$

By the maximum principle, we have

$$U_1 \leq W \leq U_2 \ \text{ in } B_1(0) \backslash B_{r_1}(r_1 + \epsilon - 1),$$

and by Hopf's lemma, we have

$$\partial_\nu U_2 \leq \partial_\nu W \leq \partial_\nu U_1 \ \text{ on } \partial B_1(0). \tag{2.12}$$

Lemma 2.4. *We have the conformal mappings Ψ_1 and Ψ_2 such that*

$$\begin{cases} \Psi_1(B_1(0)) = \Psi_2(B_1(0)) = B_1(0), \\ B_{r_1}(r_1 + \epsilon - 1) = \Psi_1\left(B_{1-\alpha_1\sqrt{\epsilon}+o(\sqrt{\epsilon})}(0)\right), \\ B_{r_2}(r_2 + \epsilon - 1) = \Psi_2\left(B_{1-\alpha_2\sqrt{\epsilon}+o(\sqrt{\epsilon})}(0)\right), \end{cases} \tag{2.13}$$

and for $i = 1, 2$,

$$\Psi_i^{-1}(z) = \left(\beta_i\sqrt{\epsilon} + o\left(\sqrt{\epsilon}\right)\right) \frac{1}{z - (-1 - \gamma_i\sqrt{\epsilon} + o(\sqrt{\epsilon}))} + \kappa_i \quad \text{as } \epsilon \to 0, \tag{2.14}$$

where α_i, β_i, and γ_i are strictly positive constants.

Proof. See section 4.1 in the appendix. □

Moreover, by (2.11) and (2.13), we have

$$U_i(\Psi_i(t)) = -2(\log(1 - \alpha_i\sqrt{\epsilon} + o(\sqrt{\epsilon})))^{-1} \log(|t|) + 1$$

for $t \in B_1(0)$ and

$$\begin{aligned} \int_{\partial B_1(0)} \partial_{\nu(z)} U_i(z) ds(z) &= \int_{\partial B_1(0)} \partial_{\nu(t)} U_i(\Psi_i(t)) ds(t) \\ &= -4\pi(\log(1 - \alpha_i\sqrt{\epsilon} + o(\sqrt{\epsilon})))^{-1} \quad \text{for } i = 1, 2. \end{aligned}$$

Therefore, by (2.9) and (2.12), we have

$$\begin{aligned}\int_{\partial B_1(0)} M\partial_{\nu_{(z)}} W(z)ds(z) &= \int_{\partial B_1(0)} \partial_{\nu(z)} w(\Phi(z))ds(z)\\ &= \int_{\partial D_1} \partial_\nu w(x)ds(x) = 1\end{aligned}$$

and

$$M \geq -\frac{\log(1-\alpha_1\sqrt{\epsilon}+o(\sqrt{\epsilon}))}{4\pi}.$$

It follows that

$$\begin{aligned}\partial_{\nu_{(z)}} w(\Phi(z)) &= M\partial\nu_{(z)}W(z)\\ \text{by (2.12)} \quad &\geq -\frac{\log(1-\alpha_1\sqrt{\epsilon}+o(\sqrt{\epsilon}))}{4\pi}\partial\nu_{(z)}U_2(z)\\ &= -\frac{\log(1-\alpha_1\sqrt{\epsilon}+o(\sqrt{\epsilon}))}{4\pi}\partial\nu_{(t)}U_2(\Psi_2(t))|(\Psi_2^{-1})'(z)|\\ &\geq \frac{\log(1-\alpha_1\sqrt{\epsilon}+o(\sqrt{\epsilon}))}{2\pi\log(1-\alpha_2\sqrt{\epsilon}+o(\sqrt{\epsilon}))}|(\Psi_2^{-1})'(z)|\\ &\geq \frac{\log(1-\alpha_1\sqrt{\epsilon}+o(\sqrt{\epsilon}))}{2\pi\log(1-\alpha_2\sqrt{\epsilon}+o(\sqrt{\epsilon}))}\min_{|z|=1}|(\Psi_2^{-1})'(z)|\\ \text{(2.15)} \qquad \text{by (2.14)} \quad &\geq C\sqrt{\epsilon},\end{aligned}$$

where $z = \Psi_2(t) \in \partial B_1(0)$. And since φ_1' is independent of ϵ, when $H(x_1, x_2) = x_1$, we have

$$\lim_{n\to\infty}(K_2K_1)^n(H)|_{\partial D_1} \geq C_*\sqrt{\epsilon}$$

and, by definition,

$$\lim_{n\to\infty}(K_1K_2)^n(H)|_{\partial D_2} \leq 0.$$

These bounds complete the proof of (1.3). □

REMARK 2.5. *We assumed that D_2 has a C^2 boundary. This regularity condition is used only for choosing r_2 of* (2.10) *in Lemma* 4.1. *We observe that one can prove our estimates on a relaxed regularity condition. For example, even when $D_2 = (-1,0)\times(-1,1) - \epsilon/2$, the estimates* (1.3), (1.4), *and* (1.8) *hold.*

3. The proof of Theorem 1.2. We first prove that the inequality (1.5) holds. This proof is the continuation of the proof of Theorem 1.1. By the argument similar to (2.15), we have

$$\begin{aligned}\partial_{\nu_{(z)}} w(\Phi(z)) &= M\partial\nu_{(z)}W(z)\\ &\leq -\frac{\log(1-\alpha_2\sqrt{\epsilon}+o(\sqrt{\epsilon}))}{4\pi}\partial\nu_{(z)}U_1(z)\\ &= -\frac{\log(1-\alpha_2\sqrt{\epsilon}+o(\sqrt{\epsilon}))}{4\pi}\partial\nu_{(t)}U_1(\Psi_1(t))|(\Psi_1^{-1})'(z)|\\ &\leq \frac{\log(1-\alpha_2\sqrt{\epsilon}+o(\sqrt{\epsilon}))}{2\pi\log(1-\alpha_1\sqrt{\epsilon}+o(\sqrt{\epsilon}))}|(\Psi_1^{-1})'(z)|\\ \text{(3.1)} \quad \text{by } \overline{z} = \frac{1}{z} \text{ on } \partial B_1(0), \quad &\leq CP(p_\epsilon\overline{z}),\end{aligned}$$

where $P(x,y)$ is a Poisson kernel and

$$\begin{aligned} p_\epsilon &= -1 + \gamma_1\sqrt{\epsilon} + o(\sqrt{\epsilon}) \\ &= -1 + O(\sqrt{\epsilon}) \in B_1(0) \ \text{ as } \epsilon \to \infty. \end{aligned}$$

One can prove by assuming $\varphi(-1) = 0$ and the regularity conditions of φ on the boundary, instead of the assumption (2.10), that this bound (3.1) holds and $p_\epsilon = -1 + O(\sqrt{\epsilon}) \in B_1(0)$ as $\epsilon \to 0$.

From the definition of φ, $(\varphi(\frac{1}{z}) + c)^{-1}$ is extended to a conformal mapping defined in $B(0)$ for some constant c that attains zero value at $z = 0$. Hence $\varphi(z)$ can be rewritten as follows:

$$\varphi(z) = a_1 z + h\left(\frac{1}{z}\right) \quad \text{for } z \in \mathbb{C} \setminus B_1(0), \tag{3.2}$$

where h is analytic in $B_1(0)$ and a_0 is a nonzero constant. Then for $H(x_1, x_2) = x_1$, we have

$$H(\varphi_1(z)) = \Re\left(a_1 z + h\left(\frac{1}{z}\right)\right) + \frac{\epsilon}{2}.$$

Then we define $\mathbf{H}$ on $\overline{B}_1(0)$ as follows:

$$\mathbf{H}(z) = \Re\left(a_1\overline{z} + h\left(z\right)\right) + \frac{\epsilon}{2} \quad \text{for } z \in \overline{B}_1(0). \tag{3.3}$$

It is easy to see that $\mathbf{H}(z) = H(\varphi_1(\overline{z}))$ on $\partial B_1(0)$, and that $\mathbf{H}$ is a harmonic function in $B_1(0)$ and belongs to $C^1(\overline{B_1(0)})$. Then it follows from Lemma 2.3 that

$$\begin{aligned} 0 \le \lim_{n\to\infty} (K_2K_1)^n(H)|_{\partial D_1} &= \int_{\partial D_1} \partial_\nu w(x) H(x) ds(x) \\ &\le C_1 \int_{\partial B_1(0)} P(p_\epsilon, \overline{z}) H(\varphi_1(z)) ds(z) \\ &= C_1 \ \mathbf{H}(\overline{p_\epsilon}) \\ \text{by } \mathbf{H}(\varphi_1(-1)) = \frac{\epsilon}{2}, \ &\le C_2\sqrt{\epsilon}. \end{aligned} \tag{3.4}$$

Applying the same argument to ψ and D_{left}, we also have

$$0 \ge \lim_{n\to\infty} (K_1K_2)^n(H)|_{\partial D_2} \ge C_2{}'\sqrt{\epsilon}.$$

These bounds are reduced to the inequality (1.5), that is,

$$|u|_{\partial D_1} - u|_{\partial D_2}| \le C_1^*\sqrt{\epsilon}.$$

REMARK 3.1. *We suggest another method to get the inequalities* (3.4) *and* (1.5). *We divide the integration in* (3.4) *into two parts as follows:*

$$\begin{aligned} \int_{\partial B_1(0)} P(p_\epsilon, \overline{z}) H(\varphi_1(z)) ds(z) = &\int_{\partial B_1(0) \ \textit{and} \ |z+1| \le \sqrt[4]{\epsilon}} P(p_\epsilon, \overline{z}) H(\varphi_1(z)) ds(z) \\ &+ \int_{\partial B_1(0) \ \textit{and} \ |z+1| > \sqrt[4]{\epsilon}} P(p_\epsilon, \overline{z}) H(\varphi_1(z)) ds(z). \end{aligned}$$

Then the inequality (3.4) *can also be proved directly without* (3.2). *Moreover, one can prove by using the same partition of integration that the bound* (1.5) *still holds for any harmonic function* H *with* $\partial_{x_2} H(0) = 0$.

We now derive the inequality (1.6). We divide u into four parts as follows:

$$u = x_1 + u_0 + u_1 + u_2$$

such that for $i = 0, 1, 2$, $\Delta u_i = 0$ in $\mathbb{R}^2 \backslash \overline{D_1 \cup D_2}$ and $u_i = O(1)$ as $|x| \to \infty$ with the boundary conditions

$$\begin{cases} u_0 &= u \text{ (constants)} \quad \text{on } \partial D_1 \cup \partial D_2, \\ u_1 &= -x_1 \text{ on } \partial D_1 \quad \text{and } u_1 = 0 \text{ on } \partial D_2, \\ u_2 &= -x_1 \text{ on } \partial D_2 \quad \text{and } u_2 = 0 \text{ on } \partial D_1. \end{cases}$$

Hence, we would estimate them separately.

Estimate for u_0. It follows from the maximum principle for analytic functions that

$$\|\nabla u_0\|_{L^\infty(\mathbb{R}^2 \backslash (D_1 \cup D_2))} \leq \|\partial_\nu u_0\|_{L^\infty(\partial D_1 \cup \partial D_2)}.$$

Thus we estimate $\|\partial_\nu u_0\|_{L^\infty(\partial D_1)}$. There exists a constant C independent of ϵ such that

$$\|\partial_\nu u_0\|_{L^\infty(\partial D_1)} \leq C \|\partial_{\nu(z)} u_0(\Phi(z))\|_{L^\infty(\partial B_1(0))},$$

where Φ is as defined in the proof of Theorem 1.1. By the argument similar to that of (2.12) and (3.1), we can choose a constant C such that for $z \in \partial B_1(0)$

$$\begin{aligned} \left|\partial_{\nu(z)} u_0(\Phi(z))\right| &\leq \left| \frac{1}{2} (u|_{\partial D_1} - u|_{\partial D_2}) \partial_{\nu(z)} U_1(z) \right| \\ \text{by (1.5) and the Poisson kernel } P(p_\epsilon, \cdot), \quad &\leq C \frac{1}{\sqrt{\epsilon}}, \end{aligned}$$

where U_1 is as defined in (2.11). Since u_0 is constant on ∂D_i for $i = 1, 2$, we conclude that

$$\|\nabla u_0\|_{L^\infty(\mathbb{R}^2 \backslash (D_1 \cup D_2))} \leq C_a \frac{1}{\sqrt{\epsilon}}, \tag{3.5}$$

where C_a is a constant independent of ϵ.

Estimate for u_1. Let $\psi_2 : \mathbb{C} \backslash B_1(0) \to \mathbb{R}^2 \backslash D_2$ be the conformal mapping defined by $\psi_2 = \psi - \frac{\epsilon}{2}$ as φ_1. Since $u_1(\psi_2(z)) = 0$ on $\partial B_1(0)$, $u_1(\psi_2(z))$ can be extended harmonically to $\mathbb{C} \backslash (\mathbf{B} \cup \psi_2^{-1}(D_1))$ as follows:

$$-u_1 \left(\psi_2 \left(\frac{z}{|z|^2} \right) \right) \quad \text{for } z \in B_1(0) \backslash \mathbf{B},$$

where $\mathbf{B} = \{ z \mid \frac{z}{|z|^2} \in \psi_2^{-1}(D_1) \}$. The symmetry of the extended $u_1(\psi_2(z))$ occurs on $\partial \mathbf{B}$ and $\partial \psi_2^{-1}(D_1)$. This yields

$$\|\nabla u_1\|_{L^\infty(\mathbb{R}^2 \backslash (D_1 \cup D_2))} \leq C \|\nabla u_1\|_{L^\infty(\partial D_1)},$$

where C is independent of ϵ. We now estimate $\|\partial_\nu u_1\|_{L^\infty(\partial D_1)}$, because the tangential derivative of u_1 on ∂D_1 is not only fixed by $H(x_1, x_2) = x_1$ but also independent of ϵ. Using a linear fractional transform, without loss of generality, we can also assume that

$$B_{r_4}(-1-\epsilon-r_4) \subset \varphi_1^{-1}(D_2) \subset B_{r_3}(-1-\epsilon-r_3),$$

where r_3 and r_4 are independent of ϵ. Then we consider a harmonic function V_3 and V_4 as follows:

$$\begin{cases} \Delta V_i = 0 \quad \text{in } \mathbb{R}^2 \backslash \overline{(B_1(0) \cup B_{r_i}(-1-\epsilon-r_i))}, \\ V_i(z) = O(1) \quad \text{as } |z| \to \infty, \\ V_i(z) = u_1(\varphi_1(z)) = -H(\varphi_1(z)) \quad \text{for } z \in \partial B_1(0), \\ V_i(z) = 0 \quad \text{for } z \in \partial B_{r_i}(-1-\epsilon-r_i) \end{cases}$$

for $i = 3, 4$. Since $u_1(\varphi_1(z)) \le 0$ for $z \in \partial B_1(0)$, it follows from Hopf's lemma that

$$\partial_\nu V_3(z) \le \partial_{\nu(z)} u_1(\varphi_1(z)) \le \partial_\nu V_4(z) \quad \text{for } z \in \partial B_1(0). \tag{3.6}$$

Therefore, we now estimate $\partial_\nu V_3(z)$ and $\partial_\nu V_4(z)$. By a definition similar to K_1 and K_2 in the previous proof, we define the operator $K_\alpha : C^\infty(\mathbb{R}^2 \backslash \overline{B_{r_3}(-1-\epsilon-r_3)}) \to C^\infty(\mathbb{R}^2 \backslash \overline{B_1(0)}) \cap C(\mathbb{R}^2 \backslash B_1(0))$ as

$$\begin{cases} \triangle K_\alpha(v) = 0 \quad \text{in } \mathbb{R}^2 \backslash \overline{B_1(0)}, \\ K_\alpha(v)(z) = O(1) \quad \text{as } |z| \to \infty, \\ (v - K_\alpha(v))|_{\partial B_1(0)} = 0, \end{cases}$$

and $K_\beta : C^\infty(\mathbb{R}^2 \backslash \overline{B_1(0)}) \to C^\infty(\mathbb{R}^2 \backslash \overline{B_{r_3}(-1-\epsilon-r_3)}) \cap C(\mathbb{R}^2 \backslash B_{r_3}(-1-\epsilon-r_3))$ is defined similarly. Indeed K_α is the Kelvin transform for $B_1(0)$ and K_β is the Kelvin transform for $B_{r_3}(-1-\epsilon-r_3)$ simply. Then we define a harmonic function U on $\mathbb{R}^2 \backslash (B_{r_3}(-1-\epsilon-r_3) \cup B_1(0))$ as follows:

$$U = -K_\alpha \mathbf{H}(\overline{z}) + K_\beta K_\alpha \mathbf{H}(\overline{z}) - \sum_{n=1}^{\infty} (I - K_\beta)(K_\alpha K_\beta)^n K_\alpha \mathbf{H}(\overline{z}),$$

where $\mathbf{H}$ is defined at (3.3). By an argument similar to Lemma 2.2 or [2, 4], we have

$$\begin{cases} \triangle U = 0 \quad \text{in } \mathbb{R}^2 \backslash \overline{(B_{r_3}(-1-\epsilon-r_3) \cup B_1(0))}, \\ U(z) = O(1) \quad \text{as } |z| \to \infty, \\ U = 0 \quad \text{for } z \in \partial B_{r_3}(-1-\epsilon-r_3), \\ (U + H(\varphi(z)))\,|_{B_1(0)} = \text{a constant with order } O(\sqrt{\epsilon}) \quad \text{as } \epsilon \to 0. \end{cases}$$

We note that U is a solution to the equation with the circular inclusions. Thus, we can apply Theorem 1.1 of Kang et al. [4] in U. Then we have

$$\|\nabla U\|_{L^\infty\left(\mathbb{R}^2 \backslash (B_{r_3}(-1-\epsilon-r_3) \cup B_1(0))\right)} \le C \left(\frac{1}{\sqrt{\epsilon}}\right).$$

On the other hand, we have

$$\begin{cases} \triangle(V_3 - U) = 0 \quad \text{in } \mathbb{R}^2 \backslash \overline{(B_{r_3}(-1-\epsilon-r_3) \cup B_1(0))}, \\ (V_3 - U)(z) = O(1) \quad \text{as } \ |z| \to \infty, \\ (V_3 - U)(z) = 0 \quad \text{for } z \in \partial B_{r_3}(-1-\epsilon-r_3), \\ (V_3 - U)|_{\partial B_1(0)} = \text{a constant with order } O(\sqrt{\epsilon}) \quad \text{as } \epsilon \to 0. \end{cases}$$

By the same argument as *Estimate for* u_0, we have

$$\|\nabla(V_3 - U)\|_{L^\infty\left(\mathbb{R}^2 \backslash (B_{r_3}(-1-\epsilon-r_3) \cup B_1(0))\right)} \leq C\left(\frac{1}{\sqrt{\epsilon}}\right).$$

These bounds are reduced to

$$\|\nabla V_3\|_{L^\infty\left(\mathbb{R}^2 \backslash (B_{r_3}(-1-\epsilon-r_3) \cup B_1(0))\right)} \leq C\left(\frac{1}{\sqrt{\epsilon}}\right).$$

Similarly, we also obtain

$$\|\nabla V_4\|_{L^\infty\left(\mathbb{R}^2 \backslash (B_{r_4}(-1-\epsilon-r_4) \cup B_1(0))\right)} \leq C\left(\frac{1}{\sqrt{\epsilon}}\right).$$

By (3.6), these bounds yield

$$\|\nabla u_1\|_{L^\infty(\mathbb{R}^2 \backslash (D_1 \cup D_2))} \leq C_\beta\left(\frac{1}{\sqrt{\epsilon}}\right), \tag{3.7}$$

where C_β is a constant independent of ϵ.

Estimate for u_2. This estimate is derived in the same way as u_1.

Therefore, by (3.5) and (3.7) we have completed the proof of (1.6). □

4. Appendix. In this appendix we make up the parts omitted in the proof of Theorem 1.1. We first consider Lemma 2.4.

4.1. How to construct the conformal mapping Ψ_1^{-1}. We now prove Lemma 2.4; that is, we present a method for constructing Ψ_1^{-1}. For convenience, we use two steps to derive it.

• *Step* 1. To make $\Psi_1^{-1}(B_1(0))$ and $\Psi_1^{-1}(B_{r_1}(r_1+\epsilon-1))$ concentric balls, we find

$$f_1(z) = \frac{1}{z - (-1 - \tau(\epsilon)\sqrt{\epsilon})}$$

with $f_1(-1) - f_1(-1+\epsilon) = f_1(-1+\epsilon+2r_1) - f_1(1)$. As ϵ approaches 0, we have

$$\begin{aligned} \lim_{\epsilon \to 0}\left(\frac{1}{\tau(\epsilon)}\right)^2 &= \lim_{\epsilon \to 0} f_1(-1+\epsilon+2r_1) - f_1(1) \\ &= \left(\frac{1}{\gamma_1}\right)^2. \end{aligned}$$

Therefore, we have

$$f_1(z) = \frac{1}{z - (-1 - \gamma_1\sqrt{\epsilon} + o(\sqrt{\epsilon}))}.$$

• *Step* 2. To make $\Psi_1^{-1}(B_1(0))$ a unit disk, we find

$$f_2(z) = \lambda(\epsilon)\sqrt{\epsilon} f_1(z)$$

with $f_2(-1) - f_2(1) = 2$. Then, as ϵ approaches 0, we have

$$\begin{aligned} 2 &= \lim_{\epsilon\to 0} \lambda(\epsilon)\ \sqrt{\epsilon}\ f_1(-1) \\ &= \lim_{\epsilon\to 0} \lambda(\epsilon)\frac{1}{\gamma_1} = \frac{\beta_1}{\gamma_1}. \end{aligned}$$

Therefore, we have

$$f_2(z) = (\beta_1\sqrt{\epsilon} + o(\sqrt{\epsilon}))\frac{1}{z - (-1 - \gamma_1\sqrt{\epsilon} + o(\sqrt{\epsilon}))}$$

and

$$f_2(B_1(0)) = B_1(0) - k_1 \ \text{ for some constant } k_1.$$

Hence, we obtain

$$\Psi_1^{-1}(z) = f_2(z) + k_1,$$

satisfying

$$\Psi_1^{-1}(B_1(0)) = B_1(0). \qquad \square$$

4.2. How to construct the radii r_1 and r_2 satisfying (2.10). To construct them, we provide the variable ϵ to the definitions used in the main proof. Thus D_1 and D_2 are rewritten in the form $D_1(\epsilon) = D_{\text{right}} + \frac{\epsilon}{2}$ and $D_2(\epsilon) = D_{\text{left}} - \frac{\epsilon}{2}$. In addition, we assume that $\varphi : \mathbb{C}\backslash B_1(0) \to \mathbb{R}^2\backslash D_{\text{right}}$ is a bijective conformal mapping with $\varphi(-1) = 0$. Then φ_1 and Φ are rewritten as $\varphi_{1\epsilon} = \varphi + \epsilon$ and $\Phi_\epsilon = \varphi_{1\epsilon}(\frac{z}{|z|^2})$.

For any r and z with $0 < r < 1$ and $r \le |z| \le 1$, we define the ball with radius r, denoted by $B_r[z]$, in $B_1(0)$ whose boundary attains the minimal distance r to $\partial B_1(0)$ at z, i.e.,

$$\mathrm{d}(B_r[z],\ \partial B_1(0)) = \mathrm{d}(z,\ \partial B_1(0)) = r \quad \text{and} \quad z \in \partial B_r[z],$$

where $\mathrm{d}(A, B)$ is the distance between A and B.

LEMMA 4.1. *Let $\epsilon' = d(\partial B_1(0), \partial\Phi_\epsilon^{-1}(D_2(\epsilon)))$, that is, the distance between $\partial B_1(0)$ and $\partial\Phi_\epsilon^{-1}(D_2(\epsilon))$. Then we have a constant C such that*

$$\epsilon < C\epsilon' \qquad \text{for } \ \epsilon < 1. \tag{4.1}$$

We suppose that for each $\epsilon > 0$, z_ϵ is the point at which $\partial\Phi_\epsilon^{-1}(D_2(\epsilon))$ attains the minimal distance ϵ' to $\partial B_1(0)$, i.e., $z_\epsilon \in \partial\Phi_\epsilon^{-1}(D_2(\epsilon))$ and $\epsilon' = d(z_\epsilon, \partial B_1(0))$. Then we have r_1, r_2, and $\epsilon_0 > 0$ such that

$$B_{r_2}[z_\epsilon] \subset \Phi_\epsilon^{-1}(D_2(\epsilon)) \subset B_{r_1}[z_\epsilon] \quad \text{for } \epsilon < \epsilon_0. \tag{4.2}$$

If we use this result (4.2) instead of the assumption (2.10), then we can prove by the same derivation as (2.15) that $\partial_{\nu_{(z)}} w(\Phi(z)) \ge C\sqrt{\epsilon'}$. It follows from (4.1) that (2.15) holds.

Proof. To prove the inequality (4.1), we suppose that p_ϵ is the closest point at which $\partial B_1(0)$ attains the minimal distance ϵ' to $\partial\Phi_\epsilon^{-1}(D_2(\epsilon))$, i.e., $\epsilon' = \mathrm{d}(p_\epsilon, z_\epsilon)$ and $p_\epsilon \in \partial B_1(0)$. Then we obtain

$$\begin{aligned}\epsilon &\le \mathrm{d}(\Phi_\epsilon(p_\epsilon), \Phi_\epsilon(z_\epsilon))\\ &= \mathrm{d}(\Phi_0(p_\epsilon), \Phi_0(z_\epsilon))\\ &\le C\mathrm{d}(p_\epsilon, z_\epsilon) = C\epsilon' \quad \text{for } \epsilon < 1,\end{aligned}$$

where C is a strictly positive constant. Hence, we have completed the proof of (4.1).

To construct the radius r_1, we may assume that the boundary $\partial(\varphi^{-1}(D_{\text{left}}))$ near 0 is the graph of equation $z_1 = f(z_2)$, defined on I, as follows:

$$\begin{cases} f(z_2) + z_2 i \in \partial(\varphi^{-1}(D_{\text{left}})) \quad \text{for} \quad z_2 \in I,\\ f(0) = -1 \quad \text{and} \quad f'(0) = 0,\end{cases}$$

where I is a sufficient small open interval containing 0. Owing to $\varphi \in C^2(\mathbb{C}\backslash B_1(0))$ and the strict convexity of D_{right} at 0, we obtain

$$\frac{d^2 f}{dz_2{}^2}(0) < \left[-\frac{d^2\sqrt{1-z_2^2}}{dz_2{}^2}\right]_{z_2=0}.$$

Hence, one can choose $R_0 > 0$ such that

$$\Phi_0^{-1}\left(\cup_{0\le\epsilon\le 2} D_2(\epsilon)\right) \subset B_{R_0}(-1+R_0) \subset B_1(0).$$

It may be noted that

$$\text{(4.3)} \qquad \begin{cases} \Phi_\epsilon^{-1}(D_2(\epsilon)) = \Phi_0^{-1}(D_2(2\epsilon)) \subset B_{R_0}(-1+R_0) \quad \text{for } \epsilon < 1,\\ R_0 < 1 \quad \text{and} \quad -1 \in \partial B_{R_0}(-1+R_0).\end{cases}$$

Let $\mathbf{S} = \{z \in B_1(0) \mid \Phi_\epsilon(z) \in \partial D_1(\epsilon) \text{ for some } \epsilon \in [0,1]\}$. Then there is a neighborhood N_1 of 0 such that for each $z \in \mathbf{S}\cap N_1$, ϵ is uniquely determined by $\Phi_\epsilon(z) \in \partial D_1(\epsilon)$, and for each ϵ, the boundary $\partial\Phi_\epsilon^{-1}(D_1(\epsilon))$ is connected in N_1. And we define $\kappa(z)$ as the curvature of $\partial\Phi_\epsilon^{-1}(D_1(\epsilon))$ at $z \in \mathbf{S}\cap N_1$. By the continuity of $\kappa(z)$ and $\kappa(0) > R_0^{-1}$, there is a neighborhood N_2 of 0 such that

$$\text{(4.4)} \qquad \begin{cases} \frac{2}{R_0+1} < \kappa(z) \quad \text{for } z \in \mathbf{S} \cap N_2,\\ \text{the boundary} \quad \partial\Phi_\epsilon^{-1}(D_1(\epsilon)) \text{ is connected in } N_2.\end{cases}$$

Let $r_1 = \frac{R_0+1}{2}$. Then we have

$$\cup_{0\le\epsilon\le 1}\Phi_\epsilon^{-1}\left(D_1(\epsilon)\right) \subset B_{R_0}(-1+R_0) \subset B_{r_1}[-1]$$

and can choose a small constant $\delta > 0$ such that

$$\begin{cases} \cup_{0\le\epsilon\le 1}\,\Phi_\epsilon^{-1}\left(D_1(\epsilon)\right) \subset B_{r_1}[z] \cup N_2 \quad \text{for any } z \in B_\delta(-1)\cap B_1(0),\\ B_\delta(-1)\cap B_1(0) \subset N_2.\end{cases}$$

By (4.3) and the relation of $\partial B_1(0)$ and $B_{R_0}(-1+R_0)$, we can obtain a sufficiently small constant $\epsilon_0 > 0$ such that

$$z_\epsilon \in B_\delta(-1) \quad \text{for each } \epsilon < \epsilon_0,$$

where z_ϵ is mentioned above. It follows that

$$\begin{aligned}\Phi_\epsilon^{-1}(D_1(\epsilon)) &\subset (\Phi_\epsilon^{-1}(D_1(\epsilon)) \cap N_2) \cup (\Phi_\epsilon^{-1}(D_1(\epsilon)) \cap N_2^c) \\ &\subset (\Phi_\epsilon^{-1}(D_1(\epsilon)) \cap N_2) \cup B_{r_1}[z_\epsilon] \\ \text{by (4.4)} &\subset B_{r_1}[z_\epsilon].\end{aligned}$$

This means that r_1 is the desirable radius.

To choose the radius r_2, we define $\kappa : [0, \epsilon_0] \times \partial D_{\text{left}} \to \mathbb{R}^+$ by

$$\kappa(\epsilon, z) = \text{the curvature of } \Phi_\epsilon^{-1}\left(\partial D_2(\epsilon)\right) \text{ at } \Phi_\epsilon^{-1}\left(z - \frac{\epsilon}{2}\right).$$

We set $r_2 = \left(\sup\left\{\kappa(\epsilon, z) \mid (\epsilon, z) \in [0, \epsilon_0] \times \partial D_{\text{left}}\right\}\right)^{-1}$. It is easy to prove that

$$B_{r_2}[z_\epsilon] \subset \Phi_\epsilon^{-1}(D_2(\epsilon)) \quad \text{for } \epsilon < \epsilon_0.$$

Therefore we have completed the proof. □

Acknowledgments. The author would like to express his gratitude to Professor Hyeonbae Kang, who suggested the problem studied in this paper and gave useful comments. The author is also grateful to Professor David Kinderlehrer for several helpful discussions and advice. The author proved an estimate only for symmetric inclusions in the first draft. Professor Kinderlehrer's advice stimulated the author to complete this paper for general shapes. The author would also like to thank the referees for their thorough reading of the paper and constructive comments.

REFERENCES

[1] L. Ahlfors, *Complex Analysis*, 3rd ed., McGraw-Hill, New York, 1979.
[2] H. Ammari, H. Kang, and M. Lim, *Gradient estimates for solutions to the conductivity problem*, Math. Ann., 332 (2005), pp. 277–286.
[3] H. Ammari, H. Kang, E. Kim, and M. Lim, *Reconstruction of closely spaced small inclusions*, SIAM J. Numer. Anal., 42 (2005), pp. 2408–2428.
[4] H. Ammari, H. Kang, H. Lee, J. Lee, and M. Lim, *Optimal Estimates for the Electric Field in Two Dimensions*, preprint, 2006; available online from http://www.arxiv.org/abs/math.AP/0610653.
[5] I. Babuška, B. Andersson, P. Smith, and K. Levin, *Damage analysis of fiber composites. I. Statistical analysis on fiber scale*, Comput. Methods Appl. Mech. Engrg., 172 (1999), pp. 27–77.
[6] M. F. Ben Hensen and E. Bonnetier, *Asymptotic formulas for the voltage potential in a composite medium containing closer or touching disks of small diameter*, Multiscale Model. Simul., 4 (2005), pp. 250–277.
[7] E. Bonnetier and M. Vogelius, *An elliptic regularity result for a composite medium with "touching" fibers of circular cross-section*, SIAM J. Math. Anal., 31 (2000), pp. 651–677.
[8] B. Budiansky and G. F. Carrier, *High shear stresses in stiff-fiber composites*, J. Appl. Mech., 51 (1984), pp. 733–735.
[9] J. G. Goree and A. B. Wilson, Jr., *Transverse shear loading in an elastic matrix containing two circular cylindrical inclusions*, Trans. ASEM J. Appl. Mech., 34 (1967), pp. 511–513.
[10] J. D. Jackson, *Classical Electrodynamics*, 3rd ed., Wiley, New York, 1999.
[11] H. Kang and J. K. Seo, *Layer potential technique for the inverse problems*, Inverse Problems, 12 (1996), pp. 267–278.
[12] J. B. Keller, *Stresses in narrow regions*, Trans. ASME J. Appl. Mech., 60 (1993), pp. 1054–1056.
[13] J. B. Keller, *Conductivity or a medium containing a dense array of perfectly conducting spheres or cylinders or nonconducting cylinders*, J. Appl. Phys., 34 (1963), pp. 991–993.
[14] Y. Y. Li and M. Nirenberg, *Estimates for elliptic systems from composite material*, Comm. Pure Appl. Math., 56 (2003), pp. 892–925.

[15] Y. Y. Li and M. Vogelius, *Gradient estimates for solutions to divergence form elliptic equations with discontinuous coefficients*, Arch. Ration. Mech. Anal., 153 (2000), pp. 91–151.
[16] M. Vogelius and D. Volkov, *Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 723–748.

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 731–756

MATHEMATICAL ANALYSIS OF AGE-STRUCTURED HIV-1 DYNAMICS WITH COMBINATION ANTIRETROVIRAL THERAPY*

LIBIN RONG†, ZHILAN FENG†, AND ALAN S. PERELSON‡

Abstract. Various classes of antiretroviral drugs are used to treat HIV infection, and they target different stages of the viral life cycle. Age-structured models can be employed to study the impact of these drugs on viral dynamics. We consider two models with age-of-infection and combination therapies involving reverse transcriptase, protease, and entry/fusion inhibitors. The reproductive number $\mathcal{R}$ is obtained, and a detailed stability analysis is provided for each model. Interestingly, we find in the age-structured model a different functional dependence of $\mathcal{R}$ on ϵ_{RT}, the efficacy of a reverse transcriptase inhibitor, than that found previously in nonage-structured models, which has significant implications in predicting the effects of drug therapy. The influence of drug therapy on the within-host viral fitness and the possible development of drug-resistant strains are also discussed. Numerical simulations are performed to study the dynamical behavior of solutions of the models, and the effects of different combinations of antiretroviral drugs on viral dynamics are compared.

Key words. human immunodeficiency virus type 1, antiretroviral therapy, drug resistance, optimal viral fitness, age-structured model, stability analysis

AMS subject classifications. 35L60, 45D05, 92C37, 92C45, 92C50

DOI. 10.1137/060663945

1. Introduction.

Since the discovery of the human immunodeficiency virus type 1 (HIV-1) in the early 1980s, the disease has spread in successive waves to most regions around the globe. It is reported that HIV has infected more than 60 million people, and over a third of them subsequently died [10]. Considerable scientific effort has been devoted to the understanding of viral pathogenesis, host/virus interactions, immune response to infection, and antiretroviral therapy.

Over the last decade, there has been a great effort in the mathematical modeling of HIV infection and treatment strategies. These models mainly investigated the dynamics of the target cells and infected cells, viral production and clearance, and the effects of antiretroviral drugs treatment. Perelson et al. [44] and Ho et al. [22] used a simple mathematical model to analyze a set of viral load data collected from infected patients after the administration of a protease inhibitor, and the virion clearance rate, the rate of loss of productively cells, and the viral production rate were estimated. These estimates were minimal estimates since the effects of antiretroviral drugs were assumed to be 100% effective, and cells were assumed to produce new virus immedi-

*Received by the editors June 28, 2006; accepted for publication (in revised form) January 3, 2007; published electronically March 2, 2007. Portions of this work were performed under the auspices of the U.S. Department of Energy under contract DE-AC52-06NA25396. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/siap/67-3/66394.html

†Department of Mathematics, Purdue University, West Lafayette, IN 47907 (rong@math.purdue.edu, zfeng@math.purdue.edu). The manuscript was finalized when the first author visited the Theoretical Biology and Biophysics Group, Los Alamos National Laboratory in 2006. The research of the second author was supported in part by NSF grant DMS-0314575 and the James S. McDonnell Foundation 21st Century Science Initiative.

‡Theoretical Biology and Biophysics, Los Alamos National Laboratory, MS K710, Los Alamos, NM 87545 (asp@lanl.gov). The research of this author was supported by NIH grants AI28433 and RR06555.

ately after they were infected [35, 36]. In order to characterize the time between the infection of target cells and the production of virus particles, an intracellular delay was introduced by Herz et al. [19] in a mathematical model to analyze the clinical data. Subsequently, Culshaw and Ruan [3] investigated the effect of the time delay on the stability of the endemical equilibrium in their model. Criteria were presented to guarantee the asymptotic stability of the infected steady state independent of the time delay. In [35], Nelson, Murray, and Perelson studied a generalized model that included a discrete delay and allowed for less than perfect drug effects. The estimation of kinetic parameters underlying HIV infection was improved by the use of a delay differential equation model. In [31, 32], the authors used a gamma distribution function to describe a continuous delay between infection and viral production and found no change in the estimate of δ, the death rate of productively infected cells. However, Nelson and Perelson [36] extended this model and showed that the constancy of δ was due to the assumption of 100% drug effectiveness. When drug effectiveness was less than 100%, the estimate of δ depended on the delay, i.e., the variance and mean of the assumed gamma distribution. Recently, a model including both pharmacokinetics and the intracellular delay has been employed to obtain new estimates of intracellular delay and the antiviral efficacy of ritonavir [7].

Age-structured models have also been developed to study the epidemiology of HIV. Thieme and Castillo-Chavez [52] kept track of an individual's infection age to study the effect of infection-age-dependent infectivity on the dynamics of HIV transmission in a homogeneously mixing population. Kirschner and Webb [24] proposed a model that incorporated age structure into the infected cells to account for the mechanism of AZT (zidovudine) treatment. Recently, for the within-host dynamics of HIV, age-structured models have received increasing interest due to their greater flexibility in modeling viral production and mortality of infected cells [16, 34]. Nelson et al. [34] considered an age-structured model that allowed for variations in the production rate of virus particles and the death rate of infected T cells. For a specific form of the viral production function and constant death rate of infected cells, the authors performed a local stability analysis of the nontrivial equilibrium point. They used numerical simulations to illustrate that the time to reach the peak viral level depended not only on the initial conditions but also on the speed at which viral production achieves its maximum value. Based on this age-structured model, Gilchrist, Coombs, and Perelson [16] used the various life history trade-offs between viral production and clearance of infected cells to derive the within-host relative viral fitness.

In this article, we develop two age-structured models to study HIV-1 infection dynamics. These models extend the existing age-structured models [16, 24, 34] by incorporating combination therapies to study the influence of antiretroviral therapy on the evolution of HIV-1. The first model includes therapy with a combination of a reverse transcriptase (RT) inhibitor and a protease inhibitor, while the second model includes an entry inhibitor and a protease inhibitor. To account for the fact that reverse transcription takes place in the early stage of infection before an infected T cell produces virus particles, we divide the infected cells into two subclasses. One subclass represents the cells that have been infected by the virus but in which reverse transcription has not been completed. The other subclass contains infected cells that have finished the reverse transcription process and are capable of producing new virions. Our stability analysis is performed for a general form of both the viral production rate and the mortality rate of infected cells. The stability of the infection-free or the infected steady state is shown to depend on the reproductive ratio $\mathcal{R}$ being

smaller or greater than 1. The formulation of this reproductive ratio also provides an appropriate measure for the within-host viral fitness, which can be used to explore the optimal viral production rate for which $\mathcal{R}$ is maximized.

We also discuss the possible influence of treatment for drug-sensitive strains of HIV-1 on the development of drug-resistant strains of the pathogen. Clinical studies have suggested that prolonged treatment with a single antiretroviral drug may be associated with the emergence of resistant virus [20, 26, 27, 28, 39]. The impact of drug treatment on the dynamics of resistant stains of pathogens has been studied using age-independent mathematical models (see, for example, [11, 26, 55]). We show that if viral production is linked to resistance, then higher treatment efficacy with antiretroviral agents (such as protease inhibitors) may lead to the establishment of multiple viral strains with a wider range of resistance levels.

The organization of the remaining part is as follows. In section 2, we formulate a mathematical model for HIV-1 infection that generalizes the age-structured model proposed in [34] by incorporating an RT inhibitor and a protease inhibitor. Section 3 is devoted to the analysis of our model, including the existence and stability of both the infection-free and the infected steady states. In section 4, another model including therapy with a new class of drugs, fusion/entry inhibitors, is developed. Stability properties of the steady states are also obtained in this section. In section 5, we derive a criterion for invasion by drug-resistant strains and explore how drug treatment may affect the optimal viral fitness of resistant strains. Some numerical simulations are presented in section 6 to illustrate/extend our analytical results. We also compare the treatment effects of these two combination antiretroviral therapies. Section 7 contains concluding remarks.

2. The model with RT and protease inhibitors. HIV infection begins by the attachment of a virus to a $CD4^+$ cell. Inside the cell, the HIV-1 enzyme RT makes a DNA copy of the virus's RNA genome. During this process, if an RT inhibitor is present, then the viral genome will not be copied into DNA, and therefore the host cell will not produce new virus. When the virus replicates, its DNA is read out to produce viral proteins. A large polyprotein is made, and a viral protease is needed to cut the long polypeptide chain into individual components that are needed to produce infectious virus particles. If the HIV-1 protease is inhibited, the newly produced virus will be noninfectious.

From the above description of the HIV life cycle and the roles of various inhibitors, it is clear that the infection age of an infected cell can be important for the study of HIV dynamics under the influence of antiretroviral drug treatment. In [34] the following age-structured model of HIV infection (without drug treatment) was proposed:

$$
\begin{aligned}
&\frac{d}{dt}T(t) = s - dT - kVT,\\
&\frac{\partial}{\partial t}T^*(a,t) + \frac{\partial}{\partial a}T^*(a,t) = -\delta(a)T^*(a,t),\\
&\frac{d}{dt}V(t) = \int_0^\infty p(a)T^*(a,t)da - cV,\\
&T^*(0,t) = kVT,
\end{aligned}
\tag{2.1}
$$

where $T(t)$ denotes the concentration of uninfected target T cells at time t, $T^*(a,t)$

denotes the concentration of infected T cells of infection age a (i.e., the time that has elapsed since an HIV virion has penetrated the cell) at time t, and $V(t)$ denotes the concentration of infectious virus at t. s is the recruitment rate of healthy T cells, d is the per capita death rate of uninfected cells, $\delta(a)$ is the age-dependent per capita death rate of infected cells, c is the clearance rate of virions, k is the rate at which an uninfected cell becomes infected by an infectious virus, and $p(a)$ is the viral production rate of an infected cell with age a.

The functional forms of the viral production kernel, $p(a)$, and the death rate of infected cells, $\delta(a)$, need to be determined experimentally [21, 34]. In [34], the authors choose the following function for the production rate:

$$p(a) = \begin{cases} p^* \left(1 - e^{-\theta(a-a_1)}\right) & \text{if } a \geq a_1, \\ 0 & \text{else,} \end{cases} \tag{2.2}$$

where θ determines how quickly $p(a)$ reaches the saturation level p^*, and a_1 is the age at which reverse transcription is completed.

To incorporate the two types of treatments mentioned above, we divide the class of infected cells, $T^*(a,t)$, into two subclasses: $T^*_{preRT}(a,t)$ and $T^*_{postRT}(a,t)$. $T^*_{preRT}(a,t)$ represents the density of cells that have been "infected" by an HIV virion but in which reverse transcription has not been completed at infection age a. An RT inhibitor could allow a preRT cell to revert back to an uninfected cell (because if reverse transcription fails to complete, cellular nucleases will degrade the HIV RNA that entered the cell) or reduce the probability that a preRT cell progresses to the postRT state [9]. $T^*_{postRT}(a,t)$ represents the density of infected cells that have progressed to the postRT phase at infection age a. The densities of the preRT and postRT cells are related by a function $\beta(a)$ $(0 \leq \beta(a) \leq 1)$ that describes the proportion of infected cells that have not completed reverse transcription, i.e.,

$$T^*_{preRT}(a,t) = \beta(a)T^*(a,t), \qquad T^*_{postRT}(a,t) = (1-\beta(a))T^*(a,t). \tag{2.3}$$

We assume that $\beta(a) \in L^1[0,\infty)$ is a nonincreasing function with the following properties: $0 \leq \beta(a) \leq 1$; $\beta(0) = 1$; $\beta(a) = 0$ for $a \geq a_1$; $\beta'(a) \leq 0$ a.e.

Let ϵ_{RT} and ϵ_{PI} denote the efficacy of the therapy with RT inhibitors and protease inhibitors, respectively $(0 \leq \epsilon_{RT},\ \epsilon_{PI} < 1)$. The efficacy is scaled such that zero represents complete ineffectiveness and unity represents 100% effectiveness. To study the effect of protease inhibitor, we divide the newly produced virus particles into two classes: infectious virions with concentration $V_I(t)$ and noninfectious virions with concentration $V_{NI}(t)$. New infectious virus particles are produced at the rate $\int_0^\infty (1-\epsilon_{PI})p(a)T^*_{postRT}(a,t)da$.

Let $\eta(\epsilon_{RT})$ denote the rate at which preRT cells revert to the uninfected state due to the failure of reverse transcription. The rate at which preRT cells of all ages become uninfected is then given by $\int_0^\infty \eta(\epsilon_{RT})T^*_{preRT}(a,t)da$.

The reversion rate $\eta(\epsilon_{RT})$ is an increasing function of drug efficacy ϵ_{RT}. In the absence of drug therapy, we assume there are no infected cells going back to the uninfected class, i.e., $\eta(0) = 0$. As the limit case, when RT inhibitors are 100% effective $(\epsilon_{RT} \to 1)$, $\eta(\epsilon_{RT})$ should be very large. We shall discuss the functional form of $\eta(\epsilon_{RT})$ more in the simulation section. Our analytical results are obtained for a general reversion rate function.

Incorporating these drugs into the equations for T, T^*, and V in model (2.1), we have

$$\begin{aligned}
&\frac{d}{dt}T(t) = s - dT - kV_IT + \int_0^\infty \eta(\epsilon_{RT})T^*_{preRT}(a,t)da,\\
&\frac{\partial}{\partial t}T^*(a,t) + \frac{\partial}{\partial a}T^*(a,t) = -\delta(a)T^*(a,t) - \eta(\epsilon_{RT})T^*_{preRT}(a,t)da,\\
&\frac{d}{dt}V_I(t) = \int_0^\infty (1-\epsilon_{PI})p(a)T^*_{postRT}(a,t)da - cV_I,\\
&\frac{d}{dt}V_{NI}(t) = \int_0^\infty \epsilon_{PI}p(a)T^*_{postRT}(a,t)da - cV_{NI},\\
&T^*(0,t) = kV_IT.
\end{aligned} \tag{2.4}$$

Notice that the variable V_{NI} does not appear in equations for other variables. Thus, we can ignore the V_{NI} equation when studying the dynamics of infection. Using the relation (2.3), we have the following system:

$$\begin{aligned}
&\frac{d}{dt}T(t) = s - dT - kV_IT + \int_0^\infty \eta(\epsilon_{RT})\beta(a)T^*(a,t)da,\\
&\frac{\partial}{\partial t}T^*(a,t) + \frac{\partial}{\partial a}T^*(a,t) = -\delta(a)T^*(a,t) - \eta(\epsilon_{RT})\beta(a)T^*(a,t),\\
&\frac{d}{dt}V_I(t) = \int_0^\infty (1-\epsilon_{PI})(1-\beta(a))p(a)T^*(a,t)da - cV_I,\\
&T^*(0,t) = kV_IT.
\end{aligned} \tag{2.5}$$

In our analysis, we allow the viral production rate $p(a)$ to be an arbitrary function that is bounded (e.g., it does not have to be a monotone function). $\delta(a)$ is also assumed to be a bounded function.

Since we are interested in the effect of combination therapy on virus dynamics, we assume that the patients are initially at steady state and the combination of drugs is administered at time 0. We choose the initial conditions to be $T(0) = T_0$, $V_I(0) = V_{I0}$, $V_{NI}(0) = 0$, and $T^*(a,0) = T^*_0(a)$, where T_0 and V_{I0} are the steady state levels of target cells and infectious virions, respectively. $T^*_0(a)$ is the age distribution of infected cells at the initial time $t = 0$, and $\int_0^\infty T^*_0(a)da$ represents the steady state level of infected cells before the onset of drug therapy.

System (2.5) can be reformulated as a system of Volterra integral equations. To simplify expressions, we introduce the following notations:

$$\begin{aligned}
&K_0(a) = e^{-\int_0^a(\delta(s)+\eta(\epsilon_{RT})\beta(s))ds}, \quad K_1(a) = \eta(\epsilon_{RT})\beta(a)K_0(a),\\
&K_2(a) = (1-\epsilon_{PI})(1-\beta(a))p(a)K_0(a), \quad \mathcal{K}_i = \int_0^\infty K_i(a)da, \quad i = 1,2.
\end{aligned} \tag{2.6}$$

$K_0(a)$ is the probability of an infected cell remaining infected at age a, hereafter the age-specific survival probability of an infected cell. $K_2(a)$ is the product of the age-specific survival probability of an infected cell and the rate at which infectious virus particles are produced by an infected cell of age a. Thus, the integral of $K_2(a)$ over all ages, i.e., $\mathcal{K}_2 = \int_0^\infty (1-\epsilon_{PI})(1-\beta(a))p(a)K_0(a)da$, gives the total number of infectious virus particles produced by one infected cell over its lifespan. For convenience, we call $\mathcal{K}_2$ the infectious virus burst size.

For mathematical convenience, we introduce a new variable, $B(t)$, to describe the rate at which an uninfected T cell becomes infected at time t,

$$B(t) = kV_I(t)T(t). \tag{2.7}$$

Integrating the T^* equation in system (2.5) along the characteristic lines, $t - a =$ constant, we get the following formula:

$$T^*(a,t) = \begin{cases} B(t-a)K_0(a) & \text{for } a < t, \\ T_0^*(a-t)\dfrac{K_0(a)}{K_0(a-t)} & \text{for } a \geq t. \end{cases} \tag{2.8}$$

Substituting (2.8) into the T and V_I equations in (2.5),

$$\begin{aligned} \frac{d}{dt}T(t) &= s - dT - B(t) + \int_0^t K_1(a)B(t-a)da + \tilde{F}_1(t), \\ \frac{d}{dt}V_I(t) &= \int_0^t K_2(a)B(t-a)da - cV_I + \tilde{F}_2(t), \end{aligned} \tag{2.9}$$

where

$$\begin{aligned} \tilde{F}_1(t) &= \int_t^\infty \eta(\epsilon_{RT})\beta(a)T_0^*(a-t)\frac{K_0(a)}{K_0(a-t)}da, \\ \tilde{F}_2(t) &= \int_t^\infty (1-\epsilon_{PI})(1-\beta(a))p(a)T_0^*(a-t)\frac{K_0(a)}{K_0(a-t)}da. \end{aligned} \tag{2.10}$$

Clearly, $\tilde{F}_i(t) \to 0$ as $t \to \infty$, $i = 1, 2$. Integrating the T equation in (2.9) and changing the order of integration, we have

$$\begin{aligned} T(t) &= T_0e^{-dt} + \int_0^t e^{-d(t-u)}\left[s - B(u) + \int_0^u B(u-\tau)K_1(\tau)d\tau + \tilde{F}_1(u)\right]du \\ &= \int_0^t \left[e^{-d(t-u)}\left(s - B(u)\right) + B(u)H_1(t-u)\right]du + F_1(t), \end{aligned} \tag{2.11}$$

where

$$H_1(t) = e^{-dt}\int_0^t e^{d\tau}K_1(\tau)d\tau, \quad F_1(t) = T_0e^{-dt} + \int_0^t e^{-d(t-u)}\tilde{F}_1(u)du. \tag{2.12}$$

Similarly, by integrating the V_I equation in (2.9), we get

$$\begin{aligned} V_I(t) &= V_{I0}e^{-ct} + \int_0^t e^{-c(t-u)}\left[\int_0^u B(u-\tau)K_2(\tau)d\tau + \tilde{F}_2(u)\right]du \\ &= \int_0^t B(u)H_2(t-u)du + F_2(t), \end{aligned} \tag{2.13}$$

where

$$H_2(t) = e^{-ct}\int_0^t e^{c\tau}K_2(\tau)d\tau, \quad F_2(t) = T_0e^{-ct} + \int_0^t e^{-c(t-u)}\tilde{F}_2(u)du. \tag{2.14}$$

Equations (2.11) and (2.13), with $B(t)$ replaced by $kV_I(t)T(t)$, form a system of Volterra integral equations that are equivalent to the original system (2.5). Hence,

for determining the existence and uniqueness of the solutions we need only consider the following system:

$$\begin{aligned} T(t) &= \int_0^t \left[e^{-d(t-u)}(s - kV_I(u)T(u)) + kV_I(u)T(u)H_1(t-u)\right]du + F_1(t), \\ V_I(t) &= \int_0^t kV_I(u)T(u)H_2(t-u)du + F_2(t), \end{aligned} \tag{2.15}$$

where H_i and F_i $(i = 1, 2)$ are given in (2.12) and (2.14).

3. Analysis of the system (2.5). In this section, we provide analytic results on the existence of positive solutions as well as possible steady states and their stability for the system (2.5) or the equivalent system (2.15).

3.1. Existence of positive solutions. Let $x(t) = (T(t), V_I(t))^\top$, where $\top$ denotes the transpose of the vector. System (2.15) can be written in the form $x(t) = \int_0^t \kappa(t-u)g(x(u))du + f(t)$, where $f(t) = (F_1(t), F_2(t))^\top$ is a continuous function from $[0,\infty)$ to $[0,\infty)^2$, κ is the 2×2 matrix with entries being locally integrable functions on $[0,\infty)$,

$$\kappa(t) = \begin{pmatrix} se^{-dt} & H_1(t) - e^{-dt} \\ 0 & H_2(t) \end{pmatrix},$$

and g is defined by $g(x) = (1, kV_IT)^\top$. Obviously, $f \in C([0,\infty); \mathbf{R}^2)$, $g \in C(\mathbf{R}^2, \mathbf{R}^2)$, and $\kappa \in L^1_{loc}([0,\infty); \mathbf{R}^{2\times 2})$. Theorem 1.1 in Gripenberg, Londen, and Staffans [17, section 12.1], shows that a continuous solution exists on a maximal interval such that the solution goes to infinity if this maximal interval is finite.

To see that all solutions will remain nonnegative for positive initial data, we use the following system (see (2.7) and (2.9)) that is also equivalent to system (2.5):

$$\begin{aligned} \frac{d}{dt}T(t) &= s - dT - B(t) + \int_0^t K_1(a)B(t-a)da + \tilde{F}_1(t), \\ \frac{d}{dt}V_I(t) &= \int_0^t K_2(a)B(t-a)da - cV_I + \tilde{F}_2(t), \\ B(t) &= kV_I(t)T(t), \end{aligned} \tag{3.1}$$

where $\tilde{F}_i$ is given in (2.10) and $\tilde{F}_i(t) > 0$, $\lim_{t\to\infty}\tilde{F}_i(t) = 0$ for $i = 1, 2$.

Suppose that there exists a $\bar{t} > 0$ such that $T(\bar{t}) = 0$ and $T(t), V_I(t) > 0$ for $0 \le t < \bar{t}$. Then $B(\bar{t}) = kV_I(\bar{t})T(\bar{t}) = 0$, $B(t) = kV_I(t)T(t) > 0$ for $0 \le t < \bar{t}$, and thus from the T equation in (3.1) we have $\frac{d}{dt}T(\bar{t}) = s + \int_0^{\bar{t}} K_1(a)B(\bar{t}-a)da + \tilde{F}_1(\bar{t}) > 0$. Hence, $T(t) \ge 0$ for all $t \ge 0$. Similarly, we can show that $V_I(t) \ge 0$ and $B(t) \ge 0$ for all $t \ge 0$ and for all positive initial data.

3.2. Steady states and their stability. We use the system (3.1) for our stability analysis. According to [30], any equilibrium of system (3.1), if it exists, must be a constant solution of the following limiting system:

$$\begin{aligned} \frac{d}{dt}T(t) &= s - dT(t) - B(t) + \int_0^\infty K_1(a)B(t-a)da, \\ \frac{d}{dt}V_I(t) &= \int_0^\infty K_2(a)B(t-a)da - cV_I, \\ B(t) &= kV_I(t)T(t). \end{aligned} \tag{3.2}$$

We mention that the introduction of the variable $B(t)$ is just for mathematical convenience. If we substitute $kV_I(t)T(t)$ for $B(t)$ in the first two equations of (3.2), then we will obtain the same stability results.

System (3.2) has two constant solutions, the infection-free steady state $\bar{E} = (\bar{T}, \bar{V}_I, \bar{B}) = (s/d, 0, 0)$, and the infected steady state $E^\diamond = (T^\diamond, V_I^\diamond, B^\diamond)$, where

$$T^\diamond = \frac{c}{k\mathcal{K}_2}, \quad V_I^\diamond = \frac{sk\mathcal{K}_2 - dc}{kc(1-\mathcal{K}_1)}, \quad B^\diamond = kT^\diamond V_I^\diamond, \tag{3.3}$$

with $\mathcal{K}_1$ and $\mathcal{K}_2$ given in (2.6). Notice that $\mathcal{K}_1$ is less than 1. Thus, $V^\diamond > 0$ if and only if $sk\mathcal{K}_2 - dc > 0$, or $\mathcal{R}_1 > 1$, where

$$\mathcal{R}_1 = \frac{sk\mathcal{K}_2}{dc}. \tag{3.4}$$

Clearly, the infected steady state (3.3) is feasible if and only if $\mathcal{R}_1 > 1$. Notice that s/d is the cell density in the absence of infection, and k and c are the cell infection and viral clearance rate, respectively. Recall that $\mathcal{K}_2$, the infectious virus burst size, gives the number of infectious virus particles produced by one infected cell over its lifespan. Therefore, $\mathcal{R}_1$ gives the reproductive ratio of the virus under the impact of drugs.

We now consider the stability of steady states. Let us first consider the infection-free steady state $\bar{E}$. The following result suggests that the population sizes of virus and infected cells will go to zero if the reproductive ratio is less than 1.

THEOREM 1. *The noninfected steady state $\bar{E}$ is locally asymptotically stable (l.a.s) if $\mathcal{R}_1 < 1$, and it is unstable if $\mathcal{R}_1 > 1$.*

Proof. The Jacobian matrix of (3.2) at the steady state $\bar{E}$ is

$$J = \begin{bmatrix} -d-\lambda & -ks/d & \hat{K}_1(\lambda) \\ 0 & -c-\lambda & \hat{K}_2(\lambda) \\ 0 & ks/d & -1 \end{bmatrix},$$

where λ is an eigenvalue and $\hat{K}_i(\lambda)$ denotes the Laplace transform of $K_i(a)$, i.e., $\hat{K}_i(\lambda) = \int_0^\infty K_i(a)e^{-\lambda a}da,\ i = 1, 2$. The corresponding characteristic equation is

$$(\lambda + d)\left(\lambda + c - \frac{sk}{d}\hat{K}_2(\lambda)\right) = 0. \tag{3.5}$$

One negative root of equation (3.5) is $\lambda = -d$, and all other roots are given by the equation

$$\lambda + c = \frac{sk}{d}\hat{K}_2(\lambda), \tag{3.6}$$

which can be rewritten as

$$\frac{\lambda}{c} + 1 = \mathcal{R}_1 \frac{\hat{K}_2(\lambda)}{\mathcal{K}_2}. \tag{3.7}$$

Notice that $|\hat{K}_2(\lambda)| \le \mathcal{K}_2$ for all complex roots λ with nonnegative real parts (i.e., $\Re\lambda \ge 0$). Hence, the modulus of the right-hand side of (3.7) is less than 1, provided that $\mathcal{R}_1 < 1$. Since the modulus of the left-hand side of (3.7) is always greater than

or equal to 1 if $\Re\lambda \geq 0$, we conclude that all roots of (3.6) have negative real parts if $\mathcal{R}_1 < 1$. It follows that $\bar{E}$ is l.a.s. when $\mathcal{R}_1 < 1$.

In the case of $\mathcal{R}_1 > 1$, let $\psi(\lambda) = \frac{\lambda}{c} + 1 - \mathcal{R}_1 \frac{\hat{\mathcal{K}}_2(\lambda)}{\mathcal{K}_2}$. Thus, any real roots of $\psi(\lambda) = 0$ are also roots of (3.6). Recognizing that $\psi(0) = 1 - \mathcal{R}_1 < 0$ and $\lim_{\lambda\to\infty} \psi(\lambda) = \infty$, we know that $\psi(\lambda) = 0$ has at least one positive root $\lambda^* > 0$, which is a positive eigenvalue of the characteristic equation (3.5). This shows that the infection-free steady state is unstable when $\mathcal{R}_1 > 1$. □

The following theorem deals with the global stability of the noninfected steady state $\bar{E}$.

THEOREM 2. *For $\mathcal{R}_1 < 1$, the noninfected steady state $\bar{E}$ is a global attractor, i.e.,* $\lim_{t\to\infty}(T(t), V_I(t), B(t)) = (s/d, 0, 0)$.

In order to prove Theorem 2, we need the following lemma, in which the following notations are used: $\varphi_\infty = \liminf_{t\to\infty} \varphi(t)$, $\varphi^\infty = \limsup_{t\to\infty} \varphi(t)$, where φ is a real-valued function on $[0, \infty)$.

LEMMA 1 (see [51]). *Let φ: $[0, \infty) \to \mathbf{R}$ be bounded and continuously differentiable. Then there exist sequences $s_n, t_n \to \infty$ as $n \to \infty$ such that $\varphi(s_n) \to \varphi_\infty$, $\varphi'(s_n) \to 0$ and $\varphi(t_n) \to \varphi^\infty$, $\varphi'(t_n) \to 0$.*

Proof of Theorem 2. It is difficult to apply Lemma 1 to the T equation of (2.5) directly. We introduce a new variable, $W(t) = T(t) + \mathcal{T}^*(t)$, where $\mathcal{T}^*(t)$ denotes the total number of infected cells at t. Notice from the T^* equation in (2.5) that $\mathcal{T}^*$ satisfies the equation $\frac{d\mathcal{T}^*}{dt} = kV_I T - \int_0^\infty \big[\delta(a) + \eta(\epsilon_{RT})\beta(a)\big] T^*(a,t)da$. Then we get $\frac{dW}{dt} = s - d(W - \mathcal{T}^*) - \int_0^\infty \delta(a)T^*(a,t)da = s - dW - \int_0^\infty (\delta(a) - d)T^*(a,t)da \leq s - dW$. The last inequality holds because of the fact that $\delta(a) \geq d$ (i.e., the death rate of infected cells $\delta(a)$ is equal to the natural death rate d plus an extra death rate due to the infection). By Lemma 1, we can choose a sequence $t_n \to \infty$ such that $W(t_n) \to W^\infty$, $W'(t_n) \to 0$. From $\frac{dW}{dt} \leq s - dW$, we have $W^\infty \leq s/d$.

Rewrite the V_I equation in (2.15) as $V_I(t) = \int_0^t kV_I(t-u)T(t-u)H_2(u)du + F_2(t)$. We use Lemma 1 to choose a sequence $s_n \to \infty$ such that $V_I(s_n) \to V_I^\infty$ as $n \to \infty$. Taking supremum limit on both sides of the above V_I equation for $t = s_n \to \infty$, we have $V_I^\infty \leq kV_I^\infty T^\infty \int_0^\infty H_2(u)du$. Noticing that $T^\infty \leq W^\infty \leq s/d$ and that $\int_0^\infty H_2(u)du = \mathcal{K}_2/c$, we get $V_I^\infty \leq ks\mathcal{K}_2 V_I^\infty/(cd) = \mathcal{R}_1 V_I^\infty$. Since $\mathcal{R}_1 < 1$, we see that $V_I^\infty = 0$. Thus, $V_I(t) \to 0$ as $t \to \infty$. It also follows that $B(t) \to 0$ since $B(t) = kV_I(t)T(t)$ and $T \leq W \leq s/d$. We use Lemma 1 again to choose a sequence $s_n \to \infty$ such that $T(s_n) \to T_\infty$ and $T'(s_n) \to 0$. Using the T equation in (3.2) we get $T_\infty \geq s/d$. But $T^\infty \leq W^\infty \leq s/d$. This shows that $T(t) \to s/d$ as $t \to \infty$, which finishes the proof of Theorem 2. □

Next, we consider the stability of the infected steady state $E^\diamond$. As noted earlier, this steady state exists if and only if $\mathcal{R}_1 > 1$. The following result suggests that the virus population will be established if the reproductive ratio is greater than 1.

THEOREM 3. *The infected steady state $E^\diamond$ is l.a.s if $\mathcal{R}_1 > 1$.*

Proof. The Jacobian at the steady state $E^\diamond$ is

$$J = \begin{bmatrix} -d - kV_I^\diamond - \lambda & -kT^\diamond & \hat{K}_1(\lambda) \\ 0 & -c - \lambda & \hat{K}_2(\lambda) \\ kV_I^\diamond & kT^\diamond & -1 \end{bmatrix}.$$

Using the notation $\mathcal{R}_1 = sk\mathcal{K}_2/dc$, the corresponding characteristic equation can be

written as

$$\begin{aligned}\Big((1-\mathcal{K}_1)\lambda + d(\mathcal{R}_1-\mathcal{K}_1)\Big)\Big(\lambda + c - c\tfrac{\hat{K}_2(\lambda)}{\mathcal{K}_2}\Big)\\ = d(\mathcal{R}_1-1)\Big((\lambda+c)\hat{K}_1(\lambda) - c\tfrac{\hat{K}_2(\lambda)}{\mathcal{K}_2}\Big),\end{aligned} \tag{3.8}$$

or

$$\Big(1+\frac{\lambda}{c}\Big)\Big(A(\lambda+d)+1-\hat{K}_1(\lambda)\Big) = \frac{\hat{K}_2(\lambda)}{\mathcal{K}_2}A\big(\lambda+d\big), \tag{3.9}$$

where $A = (1-\mathcal{K}_1)/(d(\mathcal{R}_1-1))$.

We can exclude the possibility of a nonnegative real root of (3.9) as follows. Suppose $\lambda \geq 0$. Then $\hat{K}_1(\lambda) \leq \hat{K}_1(0) = \mathcal{K}_1 < 1$. It follows that $A > 0$ and $(1+\frac{\lambda}{c})(A(\lambda+d)+1-\hat{K}_1(\lambda)) > A(\lambda+d)$. Hence, (3.9) yields $\hat{K}_2(\lambda)/\mathcal{K}_2 > 1$. However, since $\lambda \geq 0$, we have $\hat{K}_2(\lambda) \leq \hat{K}_2(0) = \mathcal{K}_2$, which leads to a contradiction. Thus, (3.9) has no nonnegative real roots.

In the next step, we will exclude the possibility that (3.9) has a complex root λ with a nonnegative real part. We prove this by contradiction. Suppose that $\lambda = x_0 + iy_0$ is a root with $x_0 \geq 0$ and $y_0 > 0$. From (3.8), we have

$$(\lambda+d)\Big(\lambda + c - c\frac{\hat{K}_2(\lambda)}{\mathcal{K}_2}\Big) \to 0 \quad \text{as} \quad \mathcal{R}_1 \to 1. \tag{3.10}$$

It follows from a similar argument as in Theorem 1 that $\lambda = x_0 + iy_0$ cannot be a root if $x_0 > 0$. Now we let $x_0 = 0$ and $y_0 > 0$. In this case, (3.10) has a negative root $-d$, and all other roots are determined by the equation $(1+\frac{\lambda}{c}) = \frac{\hat{K}_2(\lambda)}{\mathcal{K}_2}$ or

$$1 + \frac{y_0}{c}i = \frac{\int_0^\infty K_2(a)\cos(ya)da}{\mathcal{K}_2} - \frac{\int_0^\infty K_2(a)\sin(ya)da}{\mathcal{K}_2}i. \tag{3.11}$$

Comparison of the real parts of both sides yields $\cos(ya) = 1$. Thus, $\sin(ya) = 0$, which implies that (3.11) cannot hold. Therefore, (3.8) has no roots with nonnegative real parts when $\mathcal{R}_1 \to 1$.

By the continuous dependence of roots of the characteristic equation on $\mathcal{R}_1$, we know that the curve determined by the roots must cross the imaginary axis as $\mathcal{R}_1$ decreases close to 1. That is, the characteristic equation (3.8) or (3.9) has a pure imaginary root, say, iy, with $y > 0$. Replacing λ in (3.9) with iy, we see that the modulus of the left-hand side of (3.9) satisfies

$$|LHS| > \left|Ad + 1 - \int_0^\infty K_1(a)\cos(ya)da + i\left(Ay + \int_0^\infty K_1(a)\sin(ya)da\right)\right|. \tag{3.12}$$

We claim that $\int_0^\infty K_1(a)\sin(ya)da \geq 0$. In fact, notice that $\int_0^\infty K_1(a)\sin(ya)da = \int_0^{a_1} K_1(a)\sin(ya)da$, where a_1 is the age at which reverse transcription is complete. Notice also that $K_1(0) = \eta(\epsilon_{RT})$ and $K_1'(a) = \eta(\epsilon_{RT})[\beta'(a)K_0(a) + \beta(a)K_0'(a)] \leq 0$ a.e. on $[0,\infty)$. Integrating $\int_0^{a_1} K_1(a)\sin(ya)da$ by parts, we get

$$\begin{aligned}\int_0^{a_1} K_1(a)\sin(ya)da &= \frac{\eta(\epsilon_{RT})}{y} - \frac{1}{y}K_1(a_1)\cos(ya_1) + \frac{1}{y}\int_0^{a_1} K_1'(a)\cos(ya)da\\ &\geq \frac{\eta(\epsilon_{RT})}{y} - \frac{1}{y}K_1(a_1)\cos(ya_1) + \frac{1}{y}\int_0^{a_1} K_1'(a)da\\ &= \frac{1}{y}K_1(a_1)(1-\cos(ya_1)) \geq 0.\end{aligned}$$

Thus, we have $\int_0^\infty K_1(a)\sin(ya)da \geq 0$. We also observe that $1-\int_0^\infty K_1(a)\cos(ya)da \geq 1-\mathcal{K}_1 > 0$. It follows from (3.12) that $|LHS| > A|d+iy|$. On the other hand, the modulus of the right-hand side of (3.9) satisfies $|RHS| \leq A|d+iy|$. This leads to a contradiction. We conclude that the characteristic equation (3.9) has no roots with nonnegative real parts. Therefore, Theorem 3 is proved. □

4. The model with entry and protease inhibitors. Since the discovery of RT inhibitors and protease inhibitors, significant progress in drug development has been made. Recently, a new class of drugs, entry/fusion inhibitors, has been introduced [10, 18]. These compounds can block the fusion of the viral envelope to the target cell membrane and interfere with continued infection. They became available with the FDA approval of enfuvirtide (Fuzeon) in 2003.

In this section, we develop an age-structured model that takes into account the effects of both entry inhibitors and protease inhibitors. The model can be described by the following equations:

$$\begin{aligned}
&\frac{d}{dt}T(t) = s - dT - (1-\epsilon_{EI})kV_IT,\\
&\frac{\partial}{\partial t}T^*(a,t) + \frac{\partial}{\partial a}T^*(a,t) = -\delta(a)T^*(a,t),\\
&\frac{d}{dt}V_I(t) = \int_0^\infty (1-\epsilon_{PI})(1-\beta(a))p(a)T^*(a,t)da - cV_I,\\
&\frac{d}{dt}V_{NI}(t) = \int_0^\infty \epsilon_{PI}(1-\beta(a))p(a)T^*(a,t)da - cV_{NI},\\
&T^*(0,t) = (1-\epsilon_{EI})kV_IT,
\end{aligned} \tag{4.1}$$

where ϵ_{EI} represents the efficacy of the entry inhibitor. The other parameters and variables have the same meaning as in the model (2.4). We remark that the model in [34] is a special case of our model (4.1) when $\epsilon_{EI} = \epsilon_{PI} = \beta(a) = 0$. Our result applies to a general form of the viral production rate $p(a)$ and the death rate $\delta(a)$.

The existence and uniqueness of (nonnegative) solutions for the system (4.1) can be proved in a similar way as for the system (2.4). Here we present only the stability analysis. The following notations are used throughout the rest of this section:

$$K_3(a) = e^{-\int_0^a \delta(s)ds}, \quad K_4(a) = (1-\epsilon_{PI})(1-\beta(a))p(a)K_3(a), \quad \mathcal{K}_4 = \int_0^\infty K_4(a)da.$$

The following limiting system is used to derive stability results:

$$\begin{aligned}
&\frac{d}{dt}T(t) = s - dT(t) - Y(t),\\
&\frac{d}{dt}V_I(t) = \int_0^\infty K_4(a)Y(t-a)da - cV_I,\\
&Y(t) = (1-\epsilon_{EI})kV_I(t)T(t),
\end{aligned} \tag{4.2}$$

where the variable $Y(t)$ is introduced for mathematical convenience.

System (4.2) has two constant solutions (steady states): the noninfected steady state $\bar{E} = (\bar{T}, \bar{V}_I, \bar{Y}) = (s/d, 0, 0)$, and the infected steady state $E^\circ = (T^\circ, V_I^\circ, Y^\circ)$, where

$$T^\circ = \frac{c}{k(1-\epsilon_{EI})\mathcal{K}_4}, \quad V_I^\circ = \frac{sk(1-\epsilon_{EI})\mathcal{K}_4 - dc}{kc(1-\epsilon_{EI})}, \quad Y^\circ = (1-\epsilon_{EI})kT^\circ V_I^\circ.$$

Clearly, $V_I^\circ > 0$ if and only if $\mathcal{R}_2 > 1$, where $\mathcal{R}_2 = sk(1-\epsilon_{EI})\mathcal{K}_4/(dc)$ is the reproductive ratio for model (4.1). Hence, E° exists if and only if $\mathcal{R}_2 > 1$. The stability results are given in the following theorem. It can be proved similarly by previous arguments. Here we omit the proof due to the space limit.

THEOREM 4. (a) *The noninfected steady state $\bar{E}$ is a global attractor if $\mathcal{R}_2 < 1$; and it is unstable if $\mathcal{R}_2 > 1$.*

(b) *When $\mathcal{R}_2 > 1$, the infected steady state E° is l.a.s.*

Results obtained in this section and in the previous section will be used in the next section to explore the impact of drug treatment on the evolution of HIV-1.

5. Influence of drug therapy on the invasion of resistant strains. In the previous sections, we have shown that a virus population can establish itself if and only if its reproductive ratio exceeds 1. Consider an environment in which the drug-sensitive strain of HIV-1 infection is at the infected steady state $E^\diamond = (T^\diamond, V_I^\diamond, B^\diamond)$ (see (3.3)), and a small number of drug-resistant virions has been introduced into the virus population. Denote the reproductive ratio of the sensitive strain by $\mathcal{R}_s$, which is the same as $\mathcal{R}_1$ defined in (3.4). We can rewrite the population size of uninfected cells in terms of $\mathcal{R}_s$, i.e., $T^\diamond = s/(d\mathcal{R}_s)$. Assume that $\mathcal{R}_s$ is greater than 1.

Let $\tilde{\epsilon}_{RT}$ and $\tilde{\epsilon}_{PI}$ denote the efficacies of the two types of drugs for the resistant strain, respectively, and let $\tilde{p}(a)$ denote the viral production rate of the resistant strain. We can define the corresponding $\tilde{K}_0(a)$ as the age-specific survival probability of T cells infected with the resistant strain (an equivalent quantity for the sensitive strain is given in (2.6)). For ease of illustration, we assume that all other parameters are the same for both strains. We derive an invasion criterion for a resistant strain by using a heuristic argument, as is done in [16]. This criterion will be applied to different scenarios of antiretroviral therapy, such as single-drug therapy (e.g., $\epsilon_{PI} > 0$ and $\epsilon_{RT} = 0$) or combination therapy (i.e., $\epsilon_{PI} > 0$ and $\epsilon_{RT} > 0$).

Notice that $1/c$ is the average lifespan of a free virus. Thus a single resistant virus can infect on average $kT^\diamond/c$ cells in its whole life. Each of these infected cells can produce a total of

$$N_r = \int_0^\infty (1-\tilde{\epsilon}_{PI})(1-\beta(a))\tilde{p}(a)\tilde{K}_0(a)da$$

infectious drug-resistant virus particles during its lifespan (burst size). Thus, the reproductive ratio of the resistant strain at the resident equilibrium density $T^\diamond$ is

$$\mathcal{R}_r^\diamond = \frac{kT^\diamond}{c}\int_0^\infty (1-\tilde{\epsilon}_{PI})(1-\beta(a))\tilde{p}(a)\tilde{K}_0(a)da,$$

and the invasion criterion is $\mathcal{R}_r^\diamond > 1$. Substituting $s/(d\mathcal{R}_s)$ for $T^\diamond$, we obtain that the condition for the resistant strain to invade the sensitive strain is $\mathcal{R}_r > \mathcal{R}_s$, where the quantity

$$\mathcal{R}_r = \frac{s}{d}\frac{k}{c}\int_0^\infty (1-\tilde{\epsilon}_{PI})(1-\beta(a))\tilde{p}(a)\tilde{K}_0(a)da \tag{5.1}$$

represents the reproductive ratio of the resistant strain when the equilibrium density of uninfected cells is s/d (which is the value of T at the infection-free steady state).

Viral fitness is often used to describe the relative replication competence of a virus in a given environment. $\mathcal{R}_r$ can be regarded as a good measure of the fitness of a resistant virus. Thus the inequality $\mathcal{R}_r > \mathcal{R}_s$ implies that natural selection within a host favors the virus strain that maximizes its reproductive ratio.

In order to calculate the reproductive ratio, we consider the case when the viral production rate for the resistant strain has the form given in (2.2). That is,

$$\tilde{p}(a) = \begin{cases} \tilde{p}^* \left(1 - e^{-\theta(a-a_1)}\right) & \text{if } a \geq a_1, \\ 0 & \text{else,} \end{cases} \tag{5.2}$$

where $\tilde{p}^*$ is the saturation level for production of the resistant strain. Accordingly, we choose $\beta(a)$ to be

$$\beta(a) = \begin{cases} 1, & 0 \leq a < a_1, \\ 0, & a \geq a_1. \end{cases} \tag{5.3}$$

The death rate of cells is assumed to be the same for both strains with the form

$$\delta(a) = \begin{cases} \delta_0, & 0 \leq a < a_1, \\ \delta_0 + \mu, & a \geq a_1, \end{cases} \tag{5.4}$$

where δ_0 and μ are positive constants with δ_0 representing a background death rate of cells and μ representing an extra death rate for productively infected cells due to either viral cytopathicity or cell-mediated immune responses.

Drug resistance is incorporated by assuming that the efficacy of antiretroviral therapy for the resistant strain is lower than that for the drug sensitive strain by a factor between 0 and 1, i.e., $\tilde{\epsilon}_{RT} = \sigma_{RT}\epsilon_{RT}$, $\tilde{\epsilon}_{PI} = \sigma_{PI}\epsilon_{PI}$. For ease of demonstration, we assume that $\sigma_{RT} = \sigma_{PI} = \sigma$. $\sigma = 0$ corresponds to the completely resistant strain, while $\sigma = 1$ corresponds to the completely sensitive strain. Other strains have an intermediate value $0 < \sigma < 1$. Many drug-resistant HIV variants display some extent of resistance-associated loss of fitness as the resistant viral strains propagate at a reduced rate when compared to sensitive strains [2]. Therefore, there is a trade-off between drug resistance and viral production rate $\tilde{p}(a)$. We choose two types of functional forms for the cost by which the saturation level p^* is reduced in resistant strains, using the following formulas:

$$\text{Type I}: \quad \tilde{p}(a) = \sigma p^* \left(1 - e^{-\theta(a-a_1)}\right), \tag{5.5}$$

$$\text{Type II}: \quad \tilde{p}(a) = e^{-\phi\left(\frac{1}{\sigma}-1\right)} p^* \left(1 - e^{-\theta(a-a_1)}\right), \tag{5.6}$$

where ϕ is a measure for the level of cost. We provide analytic results for the Type I cost and illustrate that the qualitative properties of the two types of costs are similar. Using (5.2)–(5.5), we have the following relationship between $\mathcal{R}_r$ and $\mathcal{R}_s$ (see [13]):

$$\mathcal{R}_r = \frac{\sigma(1 - \sigma\epsilon_{PI}) e^{-\eta(\epsilon_{RT})(1-\sigma)a_1}}{1 - \epsilon_{PI}} \mathcal{R}_s. \tag{5.7}$$

We consider $\mathcal{R}_r = \mathcal{R}_r(\sigma)$ as a function of σ. A drug-resistant strain with resistance σ can invade the sensitive strain if $\mathcal{R}_r(\sigma) > \mathcal{R}_s$. Obviously, it is not easy to draw conclusions from this condition. We first derive some analytic understanding for a simpler case in which only single-drug therapy with a protease inhibitor is considered, i.e., $\epsilon_{PI} > 0$ and $\epsilon_{RT} = 0$. The case of combined therapy will be explored numerically.

(a) Single-drug therapy. In this case, since $\epsilon_{PI} > 0$ and $\epsilon_{RT} = 0$, (5.7) simplifies to $\mathcal{R}_r(\sigma) = \frac{\sigma(1-\sigma\epsilon_{PI})}{1-\epsilon_{PI}} \mathcal{R}_s$. It is easy to check that in order to have $\mathcal{R}_r(\sigma) \geq \mathcal{R}_s$ for some $\sigma \in (0, 1)$ it is necessary that $\epsilon_{PI} > \frac{1}{2}$. In fact, there exists a maximum

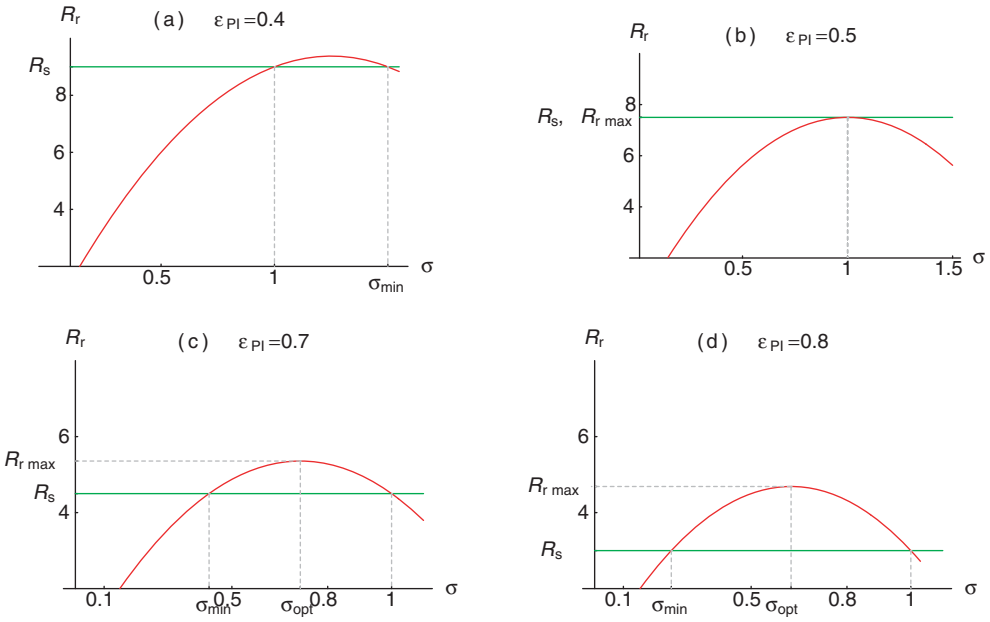


FIG. 1. *Plots of the reproductive ratio $\mathcal{R}_r$ of a resistant strain vs. the resistance σ for different treatment efficacy ϵ_{PI} (ϵ_{RT} is chosen to be 0). In* (a) *and* (b)*, it is shown that $\mathcal{R}_r < \mathcal{R}_s$ for all $\sigma < 1$. Therefore, no resistant strains can invade. In* (c) *and* (d)*, resistant strains with resistance σ in $(\sigma_{\min}, 1)$ can invade. The optimal resistance is σ_{opt} at which $\mathcal{R}_r$ reaches its maximum $\mathcal{R}_{r\max}$.*

level of resistance (corresponding to the smallest value of σ), $\sigma_{\min} = \frac{1-\epsilon_{PI}}{\epsilon_{PI}} < 1$, such that $\mathcal{R}_r(\sigma) > \mathcal{R}_s$ if and only if $\sigma_{\min} < \sigma < 1$ (see Figure 1). Clearly, if $\epsilon_{PI} < \frac{1}{2}$, then $\sigma_{\min} > 1$, and hence $\mathcal{R}_r < \mathcal{R}_s$ for all σ. This indicates that when the drug efficacy is very low, the sensitive strain is favored. The intuitive reason for this is that if the cost of resistance is high, one would not expect resistance when there is little selection pressure from the drugs. Other nonresistant strains would outcompete it under these conditions. Resistant strains can increase in frequency only when the selection pressure (drug efficacy) is high.

We can also determine an optimal resistance, σ_{opt}, which maximizes the reproductive ratio. In fact, we can easily check that $\mathcal{R}_r(\sigma)$ has only one critical point in the interval $(\sigma_{\min}, 1)$, $\sigma = \frac{1}{2\epsilon_{PI}}$, at which $\frac{d\mathcal{R}_r(\sigma)}{d\sigma} = 0$ (see Figure 1). Hence, $\sigma_{opt} = 1/(2\epsilon_{PI})$.

We summarize the following results for the case of single-drug therapy. Recall that a resistant strain with resistance σ can invade the sensitive strain if and only if $\mathcal{R}_r(\sigma) > \mathcal{R}_s$.

(i) There exists a threshold drug efficacy ϵ^*_{PI} ($\epsilon^*_{PI} = 1/2$ for Type I cost) below which no resistant strains can invade (see Figure 1(a)–(b)). Analytically, this is due to the fact that $\sigma_{\min} \geq 1$ when $\epsilon_{PI} < \epsilon^*_{PI}$. Hence, $\mathcal{R}_r(\sigma) < \mathcal{R}_s$ for all $\sigma < 1$.

(ii) When the drug efficacy is above the threshold ϵ^*_{PI}, there is a range of resistance levels for which the resistant strains are able to invade. This is because, analytically, $\sigma_{\min} < 1$ when $\epsilon_{PI} > \epsilon^*_{PI}$, and $\mathcal{R}_r(\sigma) > \mathcal{R}_s$ for all σ in $(\sigma_{\min}, 1)$.

(iii) When $\sigma_{\min} < 1$, the range of invasion strains, $(\sigma_{\min}, 1)$, increases with the drug efficacy ϵ_{PI}. The optimal resistance, σ_{opt}, decreases with the drug efficacy ϵ_{PI} (a more resistant strain corresponds to a smaller σ value; see Figure 1(c)–(d)).

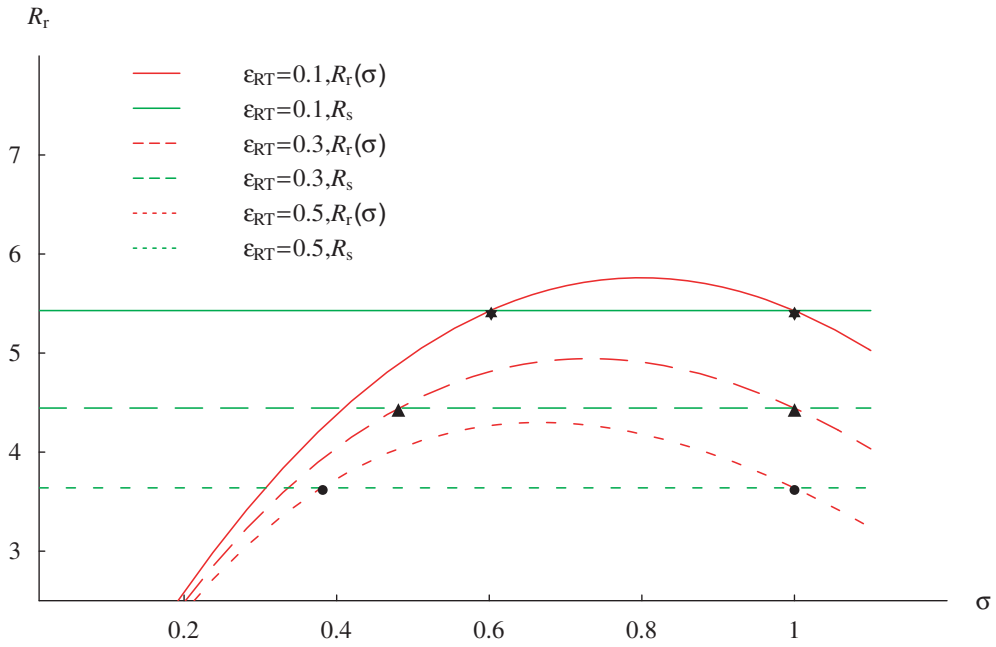


FIG. 2. *Plots of the reproductive ratio $\mathcal{R}_r$ vs. resistance σ for $\epsilon_{RT} = 0.1$ (solid), $\epsilon_{RT} = 0.3$ (long dashed), $\epsilon_{RT} = 0.5$ (short dashed). The value of ϵ_{PI} is fixed at $\epsilon_{PI} = 0.6$ for which invasion is possible in the absence of the an RT drug (i.e., if $\epsilon_{RT} = 0$). For each given ϵ_{RT}, the values of σ for which $\mathcal{R}_r(\sigma) > \mathcal{R}_s$ give the range for resistance invasion, which is the range between the two intersection points of the $\mathcal{R}_r$ curve and the $\mathcal{R}_s$ horizontal line.*

This increasing property is also clear from the formulas $\sigma_{\min} = (1 - \epsilon_{PI})/\epsilon_{PI}$ and $\sigma_{opt} = 1/(2\epsilon_{PI})$.

(iv) As the drug efficacy increases, the optimal viral fitness, $\mathcal{R}_r(\sigma_{opt})$, decreases (see Figure 1(c)–(d)).

(b) Combination therapy. We now consider the case of combination therapy, i.e., $\epsilon_{PI} > 0$ and $\epsilon_{RT} > 0$. Again, we consider $\mathcal{R}_r = \mathcal{R}_r(\sigma)$ in (5.7) as a function of σ. Then $\mathcal{R}_r(\sigma) > \mathcal{R}_s$ if and only if σ satisfies the inequality

$$\frac{\sigma(1 - \sigma\epsilon_{PI})e^{-\eta(\epsilon_{RT})(1-\sigma)a_1}}{1 - \epsilon_{PI}} > 1. \tag{5.8}$$

To explore the role of ϵ_{RT}, we fix ϵ_{PI} (e.g., $\epsilon_{PI} = 0.6$ in Figure 2). Because the numerical simulations appear qualitatively similar for different increasing reversion rate functions, we choose $\eta(\epsilon_{RT}) = \epsilon_{RT}$ for simplicity here. We will discuss the selection of the function $\eta(\epsilon_{RT})$ in the next section. Equation (5.8) cannot be solved analytically for σ. However, plots of $\mathcal{R}_r(\sigma)$ for different values of ϵ_{RT} suggest that, as ϵ_{RT} increases, the range for $\mathcal{R}_r(\sigma) > \mathcal{R}_s$ also increases (see Figure 2). Figure 3 illustrates the joint effect of ϵ_{RT} and ϵ_{PI} on the reproductive ratios $\mathcal{R}_s$ and $\mathcal{R}_r$. From the contour plot (see Figure 3(c)), we see that when the drug efficacy is low (the region in the lower-left corner in which $\mathcal{R}_s > \mathcal{R}_r > 1$) the resistant strain cannot invade. Neither strain can survive when the drug efficacy is high (the top-right region in which $\mathcal{R}_s < 1$ and $\mathcal{R}_r < 1$). In the middle region, the invasion of resistant strains is possible as $\mathcal{R}_r > \mathcal{R}_s$.

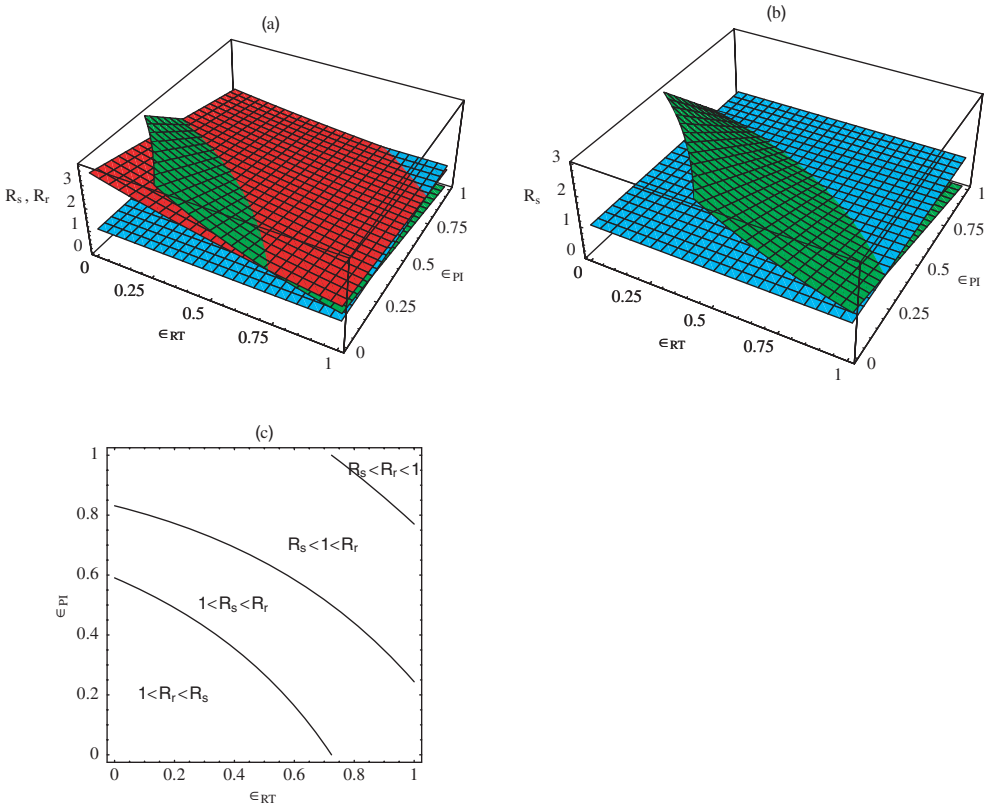


FIG. 3. *Plots of the reproductive ratios $\mathcal{R}_s$ and $\mathcal{R}_r$ as functions of ϵ_{RT} and ϵ_{PI}. Three surfaces are plotted in* (a): *$\mathcal{R}_r(\epsilon_{RT}, \epsilon_{PI})$ (the top surface near the origin), $\mathcal{R}_s(\epsilon_{RT}, \epsilon_{PI})$ (middle surface), and the constant* 1 *(the bottom surface). The intersection of the top two surfaces is the curve on which $\mathcal{R}_r = \mathcal{R}_s$. In* (b), *two surfaces, $\mathcal{R}_s(\epsilon_{RT}, \epsilon_{PI})$ and the constant* 1, *are plotted to show the curve on which $\mathcal{R}_s = 1$.* (c) *is a contour plot of the surfaces $\mathcal{R}_r(\epsilon_{RT}, \epsilon_{PI})$ and $\mathcal{R}_s(\epsilon_{RT}, \epsilon_{PI})$.*

Figure 4 shows that when the Type II cost is used, the qualitative property of the reproductive ratio $\mathcal{R}_r$ as a function of σ is very similar to that when the Type I cost is used. For example, the function $\mathcal{R}_r(\sigma)$ admits a unique $\sigma_{\min}$ and a unique σ_{opt} for sufficiently small values of ϕ.

6. Numerical results. In this section, we provide numerical simulations to confirm and/or extend our analytical results. Backward Euler and the linearized finite difference method are used to discretize the ODE and PDE, respectively, and the integral is evaluated using Simpson's rule. For all simulations, we choose the viral production rate $p(a)$ as (2.2) and $\beta(a)$ as (5.3) with $a_1 = 0.25$ days [24]. The death rate of infected cells $\delta(a)$ is assumed to be constant $\delta = 1$ day^{-1} [29], and the virion clearance rate is set to our best estimate $c = 23$ day^{-1} [45]. The other model parameters are chosen as follows [8]: $s = 10^4$ ml^{-1} day^{-1}, $d = 0.01$ day^{-1}, $k = 2.4 \times 10^{-8}$ ml day^{-1}, and the burst size is $N = 2500$.

The reversion rate function, $\eta(\epsilon_{RT})$, remains to be determined. We know $\eta(0) = 0$, and when $\epsilon_{RT} \to 1$, $\eta(\epsilon_{RT})$ should be sufficiently large such that all the preRT cells will revert back to the uninfected class. In our simulation, we assume the reversion rate function takes the following form: $\eta(\epsilon_{RT}) = -\rho \ln(1 - \epsilon_{RT})$, where the constant

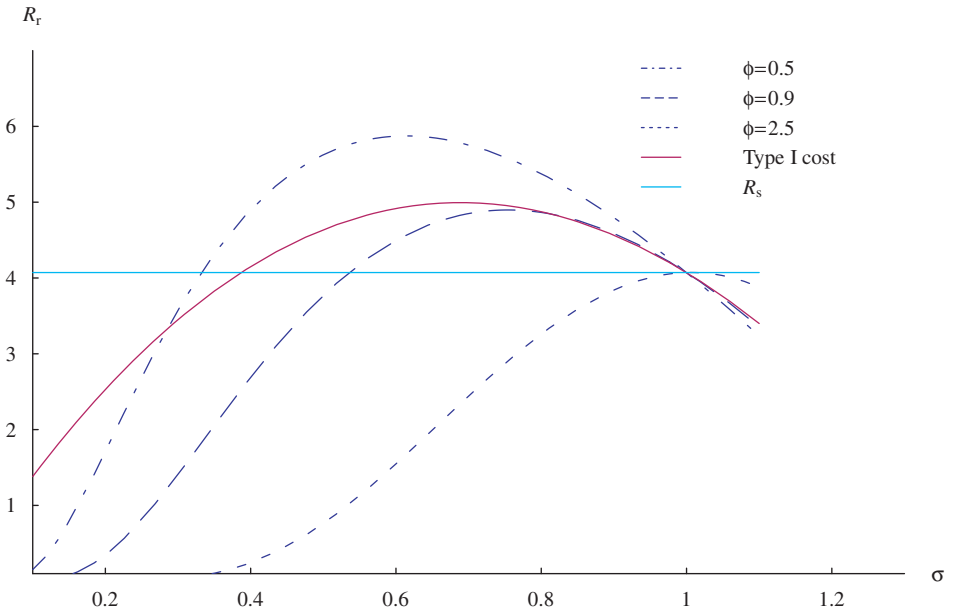


FIG. 4. *Plots of the reproductive ratio $\mathcal{R}_r$ vs. resistance σ when Type* II *cost is considered. The value of ϕ measures the cost of resistance. Invasion is possible for σ in the range between the two intersection points at which $\mathcal{R}_r = \mathcal{R}_s$. It also shows that invasion is impossible if the cost is too high (e.g., $\phi = 2.5$).*

ρ controls the steepness of the function. From the standard model in which there are only short-lived infected cells (see [44]), the viral level will be theoretically suppressed to be below the limit of viral detection (50 RNA copies ml^{-1} in the blood) in 10.2 days if RT inhibitors are assumed to be 100% effective (we assume the same parameters as above and choose the initial viral load to be 6.7038×10^5 ml^{-1}).[1] In our model (2.4), under the same initial conditions and parameters, if we choose $\rho = 2$ day^{-1}, then the viral load can reach the same limit in 10.2 days when the drug efficacy of RT inhibitors is very close to 1. Therefore, we will use the value $\rho = 2$ day^{-1} in our simulation to study the RT inhibitor's effects on the dynamics of viral load. The abilities of RT inhibitors with different ρ to suppress the viral load will be discussed later.

Figures 5 and 6 show numerical simulations of the first model (2.4) and the second model (4.1), respectively. For the calculations underlying Figure 5, the maximum age of infected cells a_{max} is chosen to be 10 days [34]. In (2.2), we choose $p^* = 6.4201 \times 10^3$ and $\theta = 1$ to guarantee the burst size is 2500 [8]. To see the influence of antiretroviral drug therapy on the viral dynamics, we choose the initial conditions to be the steady states of the standard model [44] in the absence of drug treatment. We use $T(0) = 10^6$ ml^{-1} [42] and $V(0) = 10^{-6}$ ml^{-1} [50] in the standard model to get the following steady state values: $T = 3.8333 \times 10^5$ ml^{-1}, $T^* = 6.1675 \times 10^3$ ml^{-1}, $V = 6.7038 \times 10^5$ ml^{-1}, which are used as the initial values of our models (2.4) and (4.1). The value for the efficacy of the protease inhibitor is fixed at $\epsilon_{PI} = 0.50$. Figures 5(a)–(b) and (c)–(d) are for different values of ϵ_{RT} that increase from $\epsilon_{RT} = 0.2$ (Figure 5(a)–(b)) to $\epsilon_{RT} = 0.5$ (Figure 5(c)–(d)). We observe that, when ϵ_{RT} is increased, the infection

[1] In reality, the time to reach this limit is much longer, probably due to the existence of long-lived infected cells and latently infected cells [40, 41].

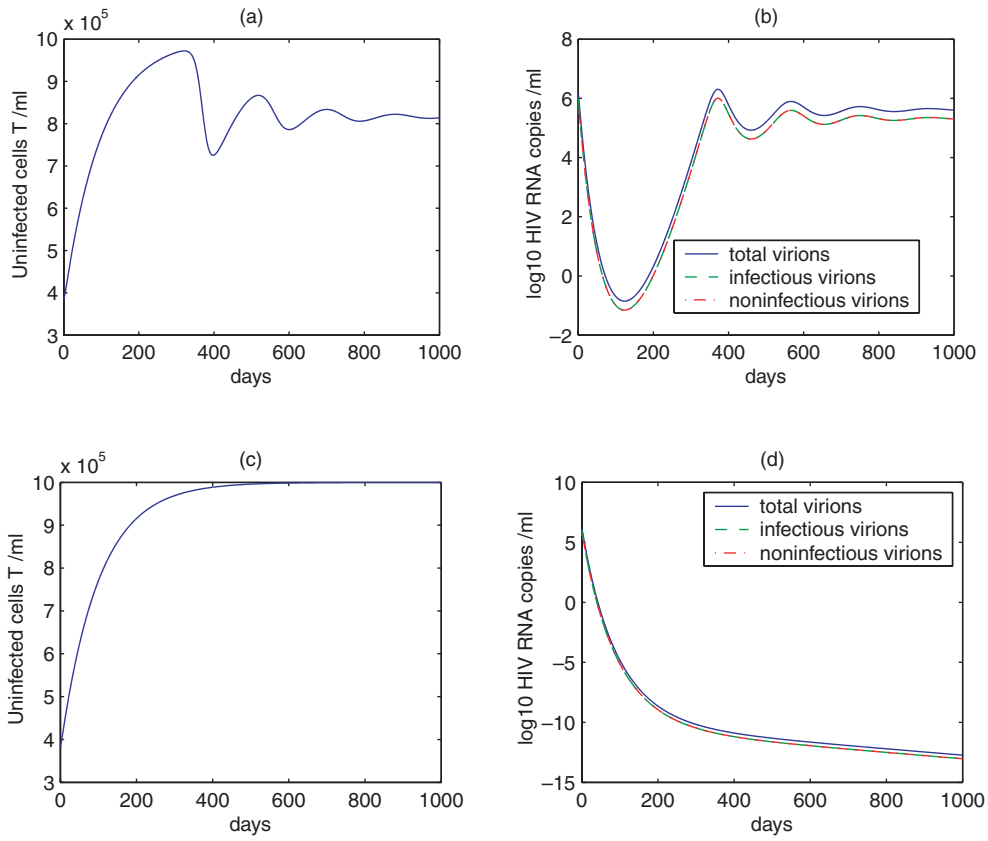


FIG. 5. *Simulation of model* (2.4) *with* $\epsilon_{PI} = 0.50$. *The upper panel:* $\epsilon_{RT} = 0.20$*; the lower panel:* $\epsilon_{RT} = 0.50$. *The other parameters for each panel are the same:* $s = 10^4\ ml^{-1}\ day^{-1}$, $d = 0.01\ day^{-1}$, $c = 23\ day^{-1}$, $k = 2.4 \times 10^{-8}\ ml\ day^{-1}$, $\delta = 1\ day^{-1}$, $p^* = 6.4201 \times 10^3\ day^{-1}$, $\theta = 1$, $T_0 = 3.8333 \times 10^5\ ml^{-1}$, $V_{I0} = 6.7038 \times 10^5\ ml^{-1}$, $V_{NI0} = 0$, $T_0^* = 6.1675 \times 10^3\ ml^{-1}$ *(see text for description). The reproductive numbers of the upper and lower panel are* 1.1666 *and* 0.9223*, respectively. The upper panel shows that the virus population stabilizes at a steady state and uninfected T cell concentration remains at* 800 μl^{-1}*, and the lower panel shows that the virus dies out and the T cell count reaches* 1000 μl^{-1}.

level at which the system stabilizes is decreased as expected. When ϵ_{RT} is greater than a threshold value ($\epsilon_{RT} = 0.41$; see also Figure 8(c)), the virus population will die out. Figure 6 shows a similar qualitative behavior of the viral load, although the efficacy of entry inhibitors has a different threshold value, $\epsilon_{EI} = 0.23$ (Figure 8(c)). The virus population persists when $\epsilon_{EI} < 0.23$ and dies out when $\epsilon_{EI} > 0.23$. This is consistent with our analytic results, as the calculation of the reproductive ratio for this set of parameters shows that $\mathcal{R}_2 > 1$ when $\epsilon_{EI} < 0.23$ and $\mathcal{R}_2 < 1$ when $\epsilon_{EI} > 0.23$. The different behaviors of the models shown in Figures 5 and 6 indicate that the entry inhibitor appears more effective than the RT inhibitor under given conditions. However, this comparison of effectiveness depends heavily on the choice of parameter ρ. If ρ is increased to 5, then the RT inhibitors can suppress viral load more effectively than entry inhibitors (see more discussion in Figure 8).

Figure 7 demonstrates how the viral load can be affected by the virion production rate $p(a)$. Each drug efficacy has a fixed value: $\epsilon_{EI} = 0.20$, $\epsilon_{PI} = 0.40$. We compare

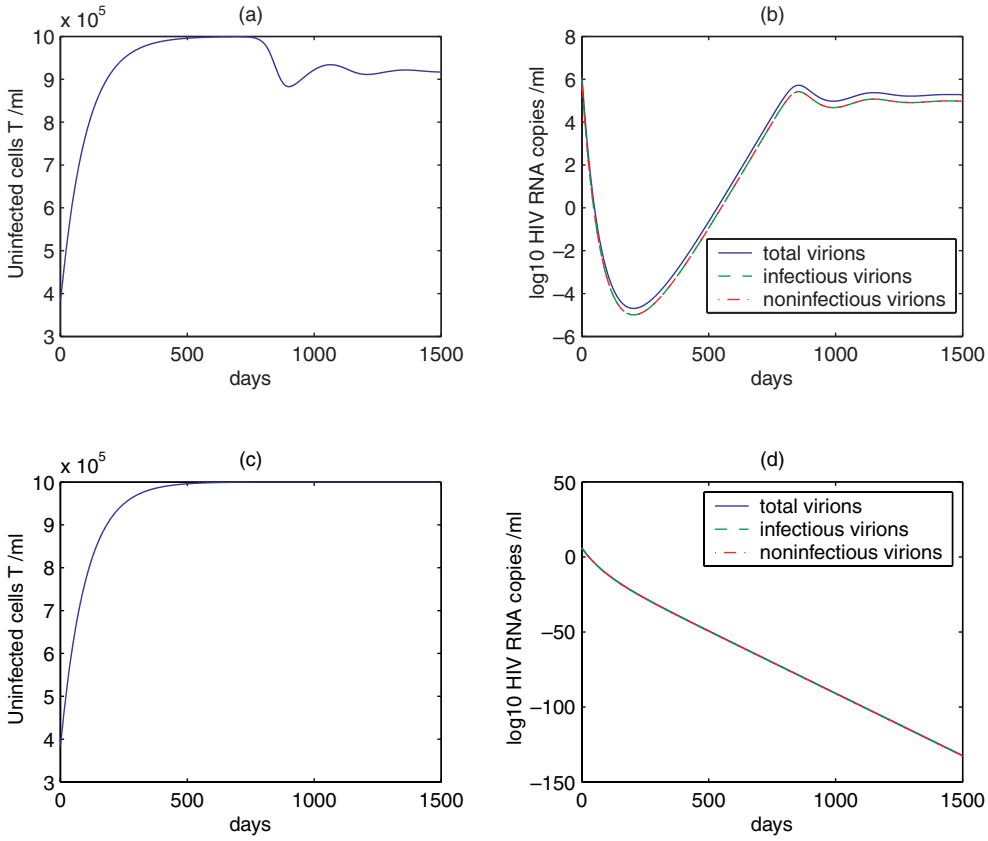


FIG. 6. *Simulation of model* (4.1) *with* $\epsilon_{PI} = 0.50$. *The upper panel:* $\epsilon_{EI} = 0.20$*; the lower panel:* $\epsilon_{EI} = 0.50$. *The other parameters are the same as those in Figure* 5. *The reproductive numbers of the upper and lower panel are* 1.0435 *and* 0.6522*, respectively. The upper panel shows that the virus population stabilizes at a lower steady state than in Figure* 5(b) *(the graphs do not show this clearly, but the numerical values show the difference) and the uninfected T cell concentration remains more than* $900\mu l^{-1}$. *The lower panel shows that the virus dies out and the T cell count reaches* $1000\mu l^{-1}$. *This implies that the entry inhibitor appears more effective than the RT inhibitor in the given conditions.*

two sets of parameters $p^* = 6.4201 \times 10^3$, $\theta = 1$ (Figure 7(a)–(b)) and $p^* = 3.5311 \times 10^3$, $\theta = 10$ (Figure 7(c)–(d)) in the viral production function (2.2), which generate the same burst size, $N = 2500$ [8]. However, the viral production rate in Figure 7(a)–(b) ramps up more slowly to the saturation level than in Figure 7(c)–(d). We observe that there is not much difference in the T cell dynamics, the viral peak, the time needed to reach the peak level, and the steady state viral load, although the nadir of the viral load in panel (d) is less than that of panel (b). This implies that varying the viral production function does not play an important role, at least in the long-term virus dynamics, given the same burst size.

Comparing Figure 7(a)–(b) with Figure 6(a)–(b), we observe that when the drug treatment becomes more effective (ϵ_{PI} increases from 0.4 to 0.5, $\epsilon_{EI} = 0.20$), the amplitude of the viral peak and the steady state viral load are decreased. However, it takes longer for the viral load to reach its peak level when the drug efficacy is higher. A possible explanation for this phenomenon is the following. Because a more effective

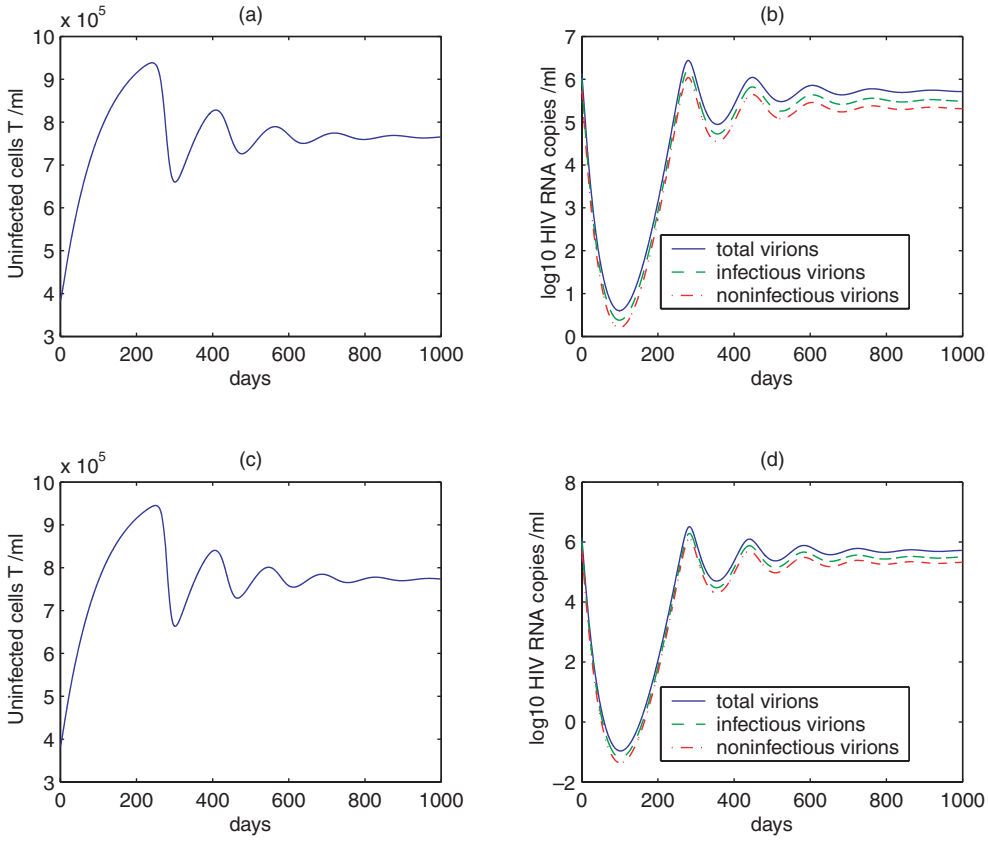


FIG. 7. *Simulation of model* (4.1) *with* $\epsilon_{EI} = 0.20, \epsilon_{PI} = 0.40$. *The upper panel:* $p^* = 6.4201 \times 10^3, \theta = 1$*; the lower panel:* $p^* = 3.5311 \times 10^3, \theta = 10$. *(The burst size of each panel is the same:* $N = 2500$*.) The other parameters are the same as those in Figure* 5. *The viral production of the lower panel ramps up more quickly to the saturation level than that of the upper panel. There is almost no difference in the viral peak, the time to reach the peak level, and the steady state viral load. This shows that the viral production function does not play an important role in the long-term viral dynamics given the same burst size.*

drug treatment (assuming that it is not potent enough to eliminate the virus) can suppress the virus more substantially, the nadir that the viral load can reach is much lower than when the treatment is more effective. Thus the time for the viral load to reach its peak level is prolonged.

In Figure 8, we compare the effects of two combination therapies on reducing the viral load. With the choice of $p(a)$ and $\beta(a)$ given in (2.2) and (5.3), we have the following reproductive numbers:

$$\mathcal{R}_1 = e^{-a_1\eta(\epsilon_{RT})}M_0, \quad \mathcal{R}_2 = (1-\epsilon_{EI})M_0, \tag{6.1}$$

where $M_0 = \frac{sk\theta}{cd\delta(\theta+\delta)}(1-\epsilon_{PI})p^*e^{-\delta a_1}$. Let $V_I^{(1)}$ and $V_I^{(2)}$ denote the viral steady states of models 1 and 2, respectively. Then $V_I^{(1)} = \frac{d(\mathcal{R}_1-1)}{k(1-\mathcal{K}_1)}$, $V_I^{(2)} = \frac{d(\mathcal{R}_2-1)}{k(1-\epsilon_{EI})}$, where $\mathcal{K}_1 = \frac{\eta(\epsilon_{RT})}{\delta+\eta(\epsilon_{RT})}(1-e^{-(\delta+\eta(\epsilon_{RT}))a_1})$. If we assume the reversion rate takes the

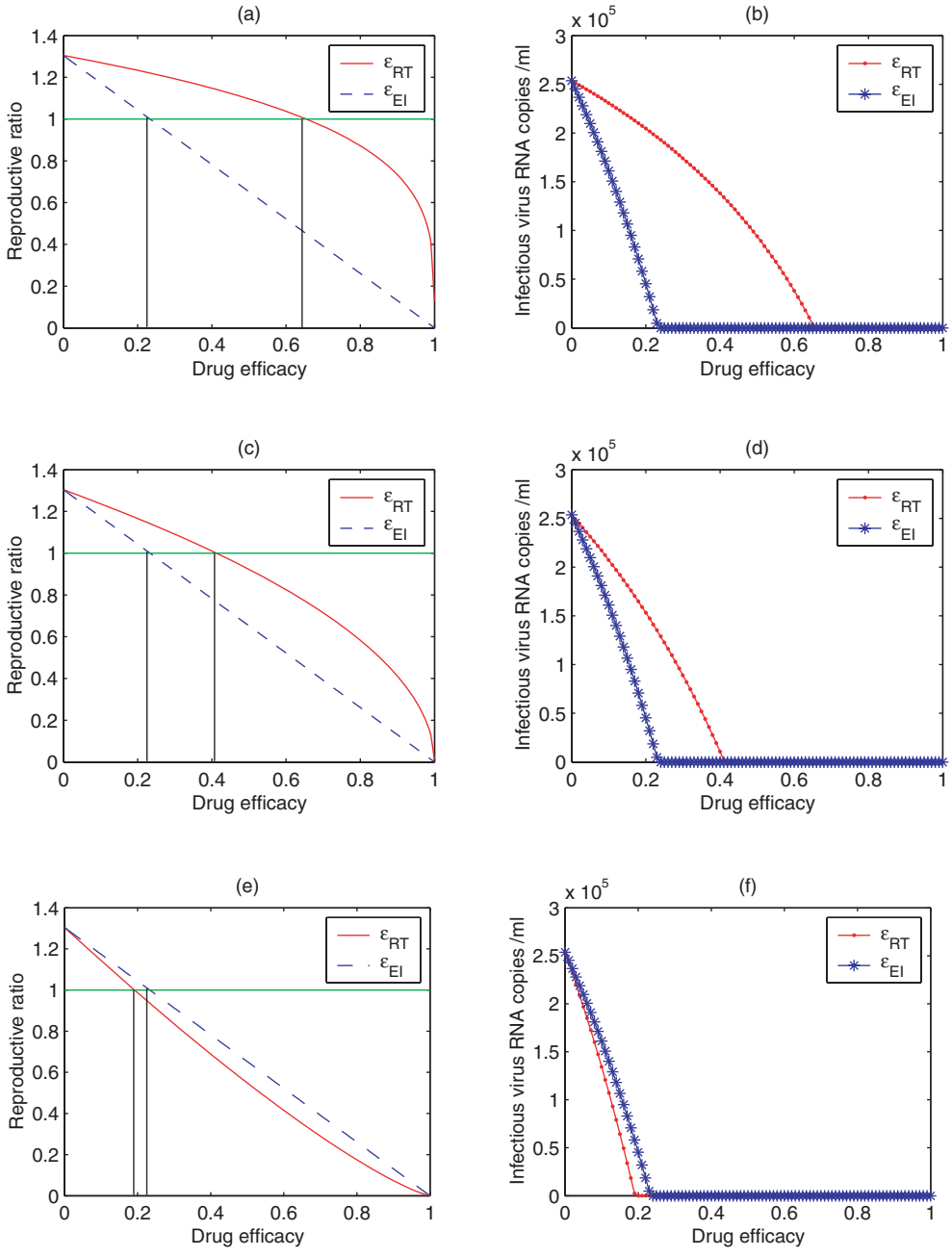


FIG. 8. *Comparison of the two combination therapies with fixed protease inhibitor drug efficacy* $\epsilon_{PI} = 0.50$. *The other parameters are the same as those in Figure* 5. *Left column: reproductive numbers* $\mathcal{R}_1$ *and* $\mathcal{R}_2$ *as the function of* ϵ_{RT} *and* ϵ_{EI}, *respectively. If* $\epsilon_{EI} > 0.23$, *then* $\mathcal{R}_2 < 1$, *and hence virus will die out. Right column: steady state* V_I *of models* (2.4) *and* (4.1) *as the function of* ϵ_{RT} *and* ϵ_{EI}, *respectively. The upper panel:* $\rho = 1$, *the threshold for* $\mathcal{R}_1 < 1$ *is* $\epsilon_{RT} > 0.65$; *the middle panel:* $\rho = 2$, *the threshold for* ϵ_{RT} *is* 0.41; *the bottom panel:* $\rho = 5$, *the threshold for* ϵ_{RT} *is* 0.19. *For a small* ρ *(*$\rho < 4$*), the entry inhibitors appear more effective than the RT inhibitors; for a large* ρ *(*$\rho > 4$*), we have the contrary result.*

form $\eta(\epsilon_{RT}) = -\rho \ln(1-\epsilon_{RT})$, then $\mathcal{R}_1$ can be simplified, and (6.1) reduces to

$$\mathcal{R}_1 = (1-\epsilon_{RT})^{a_1\rho} M_0, \quad \mathcal{R}_2 = (1-\epsilon_{EI}) M_0. \tag{6.2}$$

In Figure 8(c), we let $\epsilon_{PI} = 0.50$ and plot $\mathcal{R}_1(\epsilon_{RT})$ and $\mathcal{R}_2(\epsilon_{EI})$ as functions of ϵ_{RT} and ϵ_{EI}, respectively. We observe that there is a threshold value, $\epsilon_{RT} = 0.41$, such that $\mathcal{R}_1 > 1$ when $\epsilon_{RT} < 0.41$ and $\mathcal{R}_1 < 1$ when $\epsilon_{RT} > 0.41$. By comparison, the threshold value for entry inhibitors is $\epsilon_{EI} = 0.23$, which implies that the virus population will die out when $\epsilon_{EI} > 0.23$ (see also Figures 5 and 6). In Figure 8(d), $V_I^{(1)}(\epsilon_{RT})$ and $V_I^{(2)}(\epsilon_{EI})$ are plotted as functions of ϵ_{RT} and ϵ_{EI}, respectively. Given the same efficacy, the value of steady state $V_I^{(2)}(\epsilon_{EI})$ is less than $V_I^{(1)}(\epsilon_{RT})$. This indicates that the entry inhibitor appears more effective in reducing the viral load than the RT inhibitor in this scenario ($\rho = 2$). In fact, more information can be obtained by looking at the slopes of $\mathcal{R}_1(\epsilon_{RT})$ and $\mathcal{R}_2(\epsilon_{EI})$ in Figure 8(c) since the slope characterizes the effectiveness of drug treatment in infection control when drug efficacy is increased. From (6.2), we see that $\mathcal{R}_1$ decreases nonlinearly as ϵ_{RT} increases and the decay rate is $(1-\epsilon_{RT})^{a_1\rho}$, while $\mathcal{R}_2$ decreases linearly with the decay rate $(1-\epsilon_{EI})$ as ϵ_{EI} increases. This implies that the effectiveness of RT inhibitors depends heavily upon the reversion constant ρ. In our simulation, we choose $\rho = 2$ day^{-1} and find that the entry inhibitor is more likely able to annihilate the virus population than the RT inhibitor when the efficacy is increased by the same percentage (see Figures 5, 6, and 8(c)–(d)). However, we obtain the contrary result when ρ is chosen to be greater than $1/a_1$ (Figure 8(e)–(f)).

7. Concluding remarks. We have formulated two age-structured models for HIV-1 infection with drug treatments to study the influence of antiretroviral therapy on viral dynamics. We considered two types of combination therapies. One is the standard combination of RT inhibitors and protease inhibitors, and the other is a combination of an entry inhibitor with protease inhibitors. For each of these cases, we have calculated the reproductive ratio $\mathcal{R}_i$ ($i = 1, 2$), which is shown to determine the asymptotic stability of the infection-free steady state (when $\mathcal{R}_i < 1$) and the infected steady state (when $\mathcal{R}_i > 1$). In simple nonage-structured models, both RT and entry inhibitors are modeled in the same way, i.e., as a factor that reduces the rate of infection. Here, by considering the details of the viral life cycle, we model these inhibitors differently and explicitly. When an RT inhibitor is administered, some infected cells may have already completed reverse transcription (postRT cells). For these cells, the RT inhibitor will have no effect. For infected cells that have not completed reverse transcription (preRT cells), the RT inhibitor will prevent completion of reverse transcription and allow, under the influence of enzymes that degrade the HIV-1, a reversion of infected cells back into an uninfected state. These features of our model are novel. An entry inhibitor, enfuvirtide, has been used in a combination of RT and protease inhibitors [1, 33]. The addition of enfuvirtide to the regimen was reported to increase the antiretroviral potency in one study [33] but not the other [1]. Thus the impact of entry inhibitor use needs further evaluation, and mathematical modeling may be able to help in this regard.

We studied the impact of combination therapy using RT and protease inhibitors on the emergence of drug-resistant HIV-1 strains. The cost of resistance was assumed to be a reduced viral production rate. We calculated the reproductive ratio for the resistant strain $\mathcal{R}_r(\sigma)$ with a resistance level σ and provided a criterion for the potential invasion of resistant strains, i.e., $\mathcal{R}_r(\sigma) > \mathcal{R}_s$, in an environment where the

wild-type strain was already established. We argue that natural selection within a host favors virions that maximize the reproductive ratio, which is consistent with earlier findings (see, for example, [16]). Consequently, we show that natural selection should favor viral strains that have an intermediate level of resistance and that the optimal resistance level (σ_{opt}) decreases with increasing drug efficacy (see Figures 1 and 2). Mathematically, increasing the values of ϵ_{PI} and ϵ_{RT} results in (1) a reduction in the reproductive ratio $\mathcal{R}_s$ of the drug sensitive strain (see Figure 4) and hence a reduction in the equilibrium level of infection (see $T^\diamond = s/(dcR_s)$ and $V^\diamond = (d/k)(\mathcal{R}_s - 1)/(1 - \mathcal{K}_1)$ in (3.3)); and (2) a decrease in the optimal viral fitness $\mathcal{R}_{r\max}$ of the resistant strain and a decrease in the optimal resistance σ_{opt} (see Figures 1 and 3), and an increase in the range of resistance ($\sigma_{\min} < \sigma < 1$) for which $\mathcal{R}_r > \mathcal{R}_s$ ($\sigma_{\min}$ decreases with both ϵ_{PI} and ϵ_{RT}; see Figures 1 and 2). These are strains that are able to invade a host population. On the other hand, if ϵ_{PI} and ϵ_{RT} are small such that $\sigma_{\min}$ is greater than 1, then $\mathcal{R}_r < \mathcal{R}_s$ for all resistance σ. These strains will not be maintained in a population. It should be noted that the condition under which drug-resistant virus variants are selected in the presence of drug pressure is very complex due to various factors [48], e.g., drug potency [46], adherence to combination antiretroviral medications [14, 15], spatial heterogeneity [23], and the increasing levels of transmitted resistant virus [4]. The management of HIV-infected patients requires a better understanding of the mechanisms underlying the emergence of drug resistance. HIV resistance testing has proved helpful in clinical practice and is rapidly being incorporated into standard HIV care [47, 49].

We have also examined the effect of drug efficacy on viral dynamics by numerical simulations. As the drug efficacy increases, the steady state of viral load as well as the amplitude of the damped oscillations that characterize the approach to equilibrium decrease, which shows that an effective drug treatment will detectably lower the plasma viral load after the administration (see Figures 5, 6, and 8(c)–(d)). Moreover, the age-dependent virion production rate can also have an effect on the viral dynamics (see Figure 7).

We compared the effects of various treatments on reducing the viral population in plasma. The effectiveness of an RT inhibitor was proved to rely heavily on the reversion rate, $\eta(\epsilon_{RT})$, at which preRT cells revert back to an uninfected state because of the inhibitor. In fact, the reproductive ratio in the presence of an RT inhibitor, $\mathcal{R}_1$, is proportional to the factor $e^{-a_1\eta(\epsilon_{RT})}$ (a_1 is the age at which reverse transcription is complete), while the reproductive ratio in the presence of an entry inhibitor, $\mathcal{R}_2$, is proportional to the factor $(1 - \epsilon_{EI})$. Thus the comparison of the effectiveness of RT and entry inhibitors depends on a_1 and the functional form of $\eta(\epsilon_{RT})$. We chose a specific function $-\rho\ln(1-\epsilon_{RT})$ for $\eta(\epsilon_{RT})$ in our simulations. Then the reproductive ratio $\mathcal{R}_1$ is proportional to $(1 - \epsilon_{RT})^{a_1\rho}$. Given the same drug efficacy, an entry inhibitor appears to be more effective in reducing viral load than an RT inhibitor (see Figures 5, 6, and 8(c)–(d)) if ρ is chosen to be 2 day^{-1}. We get the contrary result if we choose ρ such that $\rho > 1/a_1$ ($\rho = 5$ day^{-1} in Figure 8(e)–(f)).

Another comparison of the effectiveness of RT inhibitors is between that obtained here using our age-structured model and the previous results in the literature based on the nonage-structured "standard model" (see [37, 44]). The reproductive ratio in the presence of an RT inhibitor for the standard model is $\mathcal{R} = (1-\epsilon_{RT})kps/(dc\delta)$ [5], where p is the constant production rate and δ is the constant death rate of productively infected cells. Thus $\mathcal{R}$ decreases linearly as the drug efficacy increases, and the decay rate is $(1-\epsilon_{RT})$. A similar comparison follows for the drug effect of RT inhibitors on suppressing the viral load between our model and the standard model. The functional

form of the reversion rate $\eta(\epsilon_{RT})$ awaits future studies. These findings might be helpful in designing treatment for the control of HIV infections. In the current model, we have not included both the wild-type and drug-resistant strains explicitly [54]. This will be done in the future work.

REFERENCES

[1] M. A. Boyd, N. M. Dixit, U. Siangphoe, N. E. Buss, M. P. Salgo, J. M. Lange, P. Phanuphak, D. A. Cooper, A. S. Perelson, and K. Ruxrungtham, *Viral decay dynamics in HIV-infected patients receiving ritonavir-boosted saquinavir and efavirenz with or without enfuvirtide: A randomized, controlled trial (HIV-NAT* 012*)*, J. Infect. Dis., 194 (2006), pp. 1319–1322.

[2] F. Clavel, E. Race, and F. Mammano, *HIV drug resistance and viral fitness*, Adv. Pharmacol., 49 (2000), pp. 41–66.

[3] R. V. Culshaw and S. Ruan, *A delay-differential equation model of HIV infection of CD*4^+ *T-cells*, Math. Biosci., 165 (2000), pp. 27–39.

[4] E. S. Daar and D. D. Richman, *Confronting the emergence of drug-resistant HIV type* 1*: Impact of antiretroviral therapy on individual and population resistance*, AIDS Res. Hum. Retroviruses, 21 (2005), pp. 343–357.

[5] P. De Leenheer and H. L. Smith, *Virus dynamics: A global analysis*, SIAM J. Appl. Math., 63 (2003), pp. 1313–1327.

[6] D. S. Dimitrov, R. L. Willey, H. Sato, L. J. Chang, R. Blumenthal, and M. A. Martin, *Quantitation of human immunodeficiency virus type* 1 *infection kinetics*, J. Virol., 67 (1993), pp. 2182–2190.

[7] N. M. Dixit, M. Markowitz, D. D. Ho, and A. S. Perelson, *Estimates of intracellular delay and average drug efficacy from viral load data of HIV-infected individuals under antiretroviral therapy*, Antivir. Ther., 9 (2004), pp. 237–246.

[8] N. M. Dixit and A. S. Perelson, *Complex patterns of viral load decay under antiretroviral therapy: Influence of pharmacokinetics and intracellular delay*, J. Theoret. Biol., 226 (2004), pp. 95–109.

[9] P. Essunger and A. S. Perelson, *Modeling HIV infection of CD*4^+ *T-cell subpopulations*, J. Theoret. Biol., 170 (1994), pp. 367–391.

[10] A. S. Fauci, *HIV and AIDS:* 20 *years of science*, Nat. Med., 9 (2003), pp. 839–843.

[11] Z. Feng, J. Curtis, and D. J. Minchella, *The influence of drug treatment on the maintenance of schistosome genetic diversity*, J. Math. Biol., 43 (2001), pp. 52–68.

[12] Z. Feng, M. Iannelli, and F. A. Milner, *A two-strain tuberculosis model with age of infection*, SIAM J. Appl. Math., 62 (2002), pp. 1634–1656.

[13] Z. Feng and L. Rong, *The influence of anti-viral drug therapy on the evolution of HIV-*1 *pathogens*, in Disease Evolution: Models, Concepts, and Data Analyses, Z. Feng, U. Dieckmann, and S. A. Levin, eds., AMS, Providence, RI, 2006, pp. 161–179.

[14] G. H. Friedland and A. Williams, *Attaining higher goals in HIV treatment: The central importance of adherence*, AIDS, 13(Suppl.1) (1999), pp. S61–S72.

[15] A. L. Gifford, J. E. Bormann, M. J. Shively, B. C. Wright, D. D. Richman, and S. A. Bozzette, *Predictors of self-reported adherence and plasma HIV concentrations in patients on multidrug antiretroviral regimens*, JAIDS, 23 (2000), pp. 386–395.

[16] M. A. Gilchrist, D. Coombs, and A. S. Perelson, *Optimizing within-host viral fitness: Infected cell lifespan and virion production rate*, J. Theoret. Biol., 229 (2004), pp. 281–288.

[17] G. Gripenberg, S. O. Londen, and O. Staffans, *Volterra Integral and Functional Equations*, Cambridge University Press, New York, 1990.

[18] R. M. Gulick, *New antiretroviral drugs*, Clin. Microbiol. Infect., 9 (2003), pp. 186–193.

[19] A. V. Herz, S. Bonhoeffer, R. M. Anderson, R. M. May, and M. A. Nowak, *Viral dynamics in vivo: Limitations on estimates of intracellular delay and virus decay*, Proc. Natl. Acad. Sci. USA, 93 (1996), pp. 7247–7251.

[20] D. D. Ho, T. Toyoshima, H. Mo, D. J. Kempf, D. Norbeck, C. M. Chen, N. E. Wideburg, S. K. Burt, J. W. Erickson, and M. K. Singh, *Characterization of human immunodeficiency virus type* 1 *variants with increased resistance to a C*2*-symmetric protease inhibitor*, J. Virol., 68 (1994), pp. 2016–2020.

[21] D. D. Ho and Y. Huang, *The HIV-*1 *vaccine race*, Cell, 110 (2002), pp. 135–138.

[22] D. D. Ho, A. U. Neumann, A. S. Perelson, W. Chen, J. M. Leonard, and M. Markowitz, *Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection*, Nature, 373 (1995), pp. 123–126.
[23] T. B. Kepler and A. S. Perelson, *Drug concentration heterogeneity facilitates the evolution of drug resistance*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 11514–11519.
[24] D. E. Kirschner and G. F. Webb, *A model for treatment strategy in the chemotherapy of AIDS*, Bull. Math. Biol., 58 (1996), pp. 367–390.
[25] D. E. Kirschner and G. F. Webb, *A mathematical model of combined drug therapy of HIV infection*, J. Theoret. Med., 1 (1997), pp. 25–34.
[26] D. E. Kirschner and G. F. Webb, *Understanding drug resistance for montherapy treatment of HIV infection*, Bull. Math. Biol., 59 (1997), pp. 763–786.
[27] B. A. Larder and S. D. Kemp, *Multiple mutations in HIV-1 reverse transcriptase confer high-level resistance to zidovudine (AZT)*, Science, 246 (1989), pp. 1155–1158.
[28] B. A. Larder, P. Kellam, and S. D. Kemp, *Convergent combination therapy can select viable multidrug-resistant HIV-1 in vitro*, Nature, 365 (1993), pp. 451–453.
[29] M. Markowitz, M. Louie, A. Hurley, E. Sun, M. Di Mascio, A. S. Perelson, and D. D. Ho, *A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo*, J. Virol., 77 (2003), pp. 5037–5038.
[30] R. K. Miller, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, Menlo Park, CA, 1971.
[31] J. E. Mittler, M. Markowitz, D. D. Ho, and A. S. Perelson, *Improved estimates for HIV-1 clearance rate and intracellular delay*, AIDS, 13 (1999), pp. 1415–1417.
[32] J. E. Mittler, B. Sulzer, A. U. Neumann, and A. S. Perelson, *Influence of delayed viral production on viral dynamics in HIV-1 infected patients*, Math. Biosci., 152 (1998), pp. 143–163.
[33] J. Molto, L. Ruiz, M. Valle, J. Martinez-Picado, A. Bonjoch, I. Bravo, E. Negredo, G. M. Heilek-Sneider, and B. Clotet, *Increased antiretroviral potency by the addition of enfuvirtide to a four-drug regimen in antiretroviral-naive, HIV-infected patients*, Antivir. Ther., 11 (2006), pp. 47–51.
[34] P. W. Nelson, M. A. Gilchrist, D. Coombs, J. M. Hyman, and A. S. Perelson, *An age-structured model of HIV infection that allows for variations in the production rate of viral particles and the death rate of productively infected cells*, Math. Biosci. Eng., 1 (2004), pp. 267–288.
[35] P. W. Nelson, J. D. Murray, and A. S. Perelson, *A model of HIV-1 pathogenesis that includes an intracellular delay*, Math. Biosci., 163 (2000), pp. 201–215.
[36] P. W. Nelson and A. S. Perelson, *Mathematical analysis of delay differential equation models of HIV-1 infection*, Math. Biosci., 179 (2002), pp. 73–94.
[37] M. A. Nowak and R. M. May, *Virus Dynamics: Mathematical Principles of Immunology and Virology*, Oxford University Press, Oxford, UK, 2000.
[38] M. A. Nowak and C. R. Bangham, *Population dynamics of immune responses to persistent viruses*, Science, 272 (1996), pp. 74–79.
[39] M. A. Nowak, S. Bonhoeffer, G. M. Shaw, and R. M. May, *Anti-viral drug treatment: Dynamics of resistance in free virus and infected cell populations*, J. Theoret. Biol., 184 (1997), pp. 203–217.
[40] A. S. Perelson, *Modelling viral and immune system dynamics*, Nature Rev. Immunol., 2 (2002), pp. 28–36.
[41] A. S. Perelson, P. Essunger, Y. Cao, M. Vesanen, A. Hurley, K. Saksela, M. Markowitz, and D. D. Ho, *Decay characteristics of HIV-1-infected compartments during combination therapy*, Nature, 387 (1997), pp. 188–191.
[42] A. S. Perelson, D. E. Kirschner, and R. De Boer, *Dynamics of HIV infection of $CD4^+$ T cells*, Math. Biosci, 114 (1993), pp. 81–125.
[43] A. S. Perelson and P. W. Nelson, *Mathematical analysis of HIV-1 dynamics in vivo*, SIAM Rev., 41 (1999), pp. 3–44.
[44] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho, *HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time*, Science, 271 (1996), pp. 1582–1586.
[45] B. Ramratnam, S. Bonhoeffer, J. Binley, A. Hurley, L. Zhang, J. E. Mittler, M. Markowitz, J. P. Moore, A. S. Perelson, and D. D. Ho, *Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis*, Lancet, 354 (1999), pp. 1782–1785.

[46] D. D. Richman, *The implications of drug resistance for strategies of combination antiviral chemotherapy*, Antiviral Res., 29 (1996), pp. 31–33.
[47] D. D. Richman, *Principles of HIV resistance testing and overview of assay performance characteristics*, Antivir. Ther., 5 (2000), pp. 27–31.
[48] D. D. Richman, *HIV chemotherapy*, Nature, 410 (2001) pp. 995–1001.
[49] D. D. Richman, *Benefits and limitations of testing for resistance to HIV drugs*, J Antimicrob. Chemother., 53 (2004), pp. 555–557.
[50] M. A. Stafford, L. Corey, Y. Cao, E. S. Daar, D. D. Ho, and A. S. Perelson, *Modeling plasma virus concentration during primary HIV infection*, J. Theoret. Biol., 203 (2000), pp. 285–301.
[51] H. R. Thieme, *Persistence under relaxed point-dissipativity (with application to an endemic model)*, SIAM J. Math. Appl., 24 (1993), pp. 407–435.
[52] H. R. Thieme and C. Castillo-Chavez, *How may the infection-age-dependent infectivity affect the dynamics of HIV/AIDS?*, SIAM J. Appl. Math., 53 (1993), pp. 1447–1479.
[53] G. F. Webb, *Theory of Nonlinear Age-Dependent Population Dynamics*, Marcel Dekker, New York, 1985.
[54] L. M. Wein, R. M. D'Amato, and A. S. Perelson, *Mathematical analysis of antiretroviral therapy aimed at HIV-1 eradication or maintenance of low viral loads*, J. Theoret. Biol., 192 (1998), pp. 81–98.
[55] D. Xu, J. Curtis, Z. Feng, and D. J. Minchella, *On the role of schistosome mating structure in the maintenance of drug resistant strains*, Bull. Math. Biol., 67 (2005), pp. 1207–1226.

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 757–776

ANALYSIS OF A MODEL OF THE GLUCOSE-INSULIN REGULATORY SYSTEM WITH TWO DELAYS*

JIAXU LI† AND YANG KUANG‡

Abstract. We continue a recent attempt to better understand the glucose-insulin regulatory system via a mathematical model of delay differential equations with two discrete time delays. With explicit delays, the model is more realistic in physiology, more accurate in mathematics, and more robust in applications. We study this model analytically and perform carefully designed numerical simulations by allowing two parameters to vary. Our analytical and numerical results confirm most current existing physiological observations and reveal more insightful information. The following factors are critical for ensuring the sustained oscillatory regulation and insulin secretion: (1) the time lag for insulin secretion stimulated by glucose and the newly synthesized insulin becoming *"remote insulin"* (Theorem 4.2 (b) and Theorem 5.6); (2) the delayed effect of hepatic glucose production (Theorem 4.2 (c) and Theorem 5.6); (3) moderate insulin clearance rate (Theorem 5.6 and simulations in section 6.4); and (4) nonoverwhelming glucose infusion (simulations in section 6.2, 6.3, and 6.4).

Key words. glucose-insulin regulatory system, insulin secretion, ultradian oscillation, time delay

AMS subject classifications. 92C50, 34C60, 92D25

DOI. 10.1137/050634001

1. Introduction.

Beginning with the pioneering work of Bolie [5] in the 1960s, several attempts at modeling the glucose-insulin regulatory system have been proposed in recent decades [2], [17], [25], [26]. These studies are, at least partially, motivated by the fact that diabetes mellitus is one of the worst diseases in the world due to the large size of the diabetic population, especially among Native Americans [14], as well as severe complications [10] and high health expenses (http://www.diabetes.org). Providing more efficient, effective, and economic treatments is the ultimate goal of these efforts (see [2], [3], [4], [5], [17], [18], [25], [26], and the references therein). The minimal model [3] and its siblings [9], [16], [19] study the insulin sensitivity, while the mathematical models proposed in [2], [5], [17], [25], [26] aim to better understand the glucose-insulin regulatory system.

In the glucose-insulin endocrine metabolic regulatory system, the pancreatic hormone insulin and glucagon are the two key players. Both in-vivo and in-vitro experiments have revealed that the insulin is secreted from the pancreas in oscillatory manners in two time scales. It is widely believed that the rapid pulsatile oscillation is caused by the insulin secretory bursts from the millions of Langerhans islets in hundreds of β-cells in the pancreas at a periodicity of 5–15 minutes [20]. The much slower ultradian oscillation refers to the oscillation of insulin secretion with period in the range of 50–150 minutes [23], [25]. The amplitude of the ultradian oscillation dominates that of the rapid pulsatile oscillation. There exist two time delays in this

*Received by the editors June 19, 2005; accepted for publication (in revised form) November 2, 2006; published electronically March 15, 2007.
http://www.siam.org/journals/siap/67-3/63400.html
†Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287-1804 (jiaxu.li@asu.edu). Current address: Department of Mathematics, University of Louisville, Louisville, KY 40929. The work of this author was partially supported by ASU grant MGIA-2006-08.
‡Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287-1804 (kuang@asu.edu). The work of this author was partially supported by grants DMS-0077790 and DMS/NIGMS-0342388 and ASU grant MGIA-2006-08.

system. One naturally occurring time delay is the time needed for the insulin to release from the β-cells stimulated by elevated glucose concentration and the newly synthesized insulin to cross the endothelial barriers and become the so-called *remote insulin*, now known as interstitial insulin. The remote insulin helps the cells, e.g., muscle and adipose, to uptake glucose. The other time delay refers to the delayed effect of hepatic glucose production. Applying the standard compartment transition technique to mimic the time delays, Sturis et al. [25] formulated a model consisting of six ordinary differential equations (ODEs) that successfully captured some of the basic features (oscillations with periods and amplitudes comparable to experiment observations) of ultradian oscillation. Recently, Li, Kuang, and Mason [17] proposed an alternative model of the glucose-insulin regulatory system consisting of two delay differential equations with two naturally explicit discrete time delays. This two-delay model uses only physiologically meaningful and measurable parameters. It is shown that this two-delay model provides the best overall fit among five plausible model systems with the experimental data given in [2], [17], and [25]. It is also shown [17] that the two-time-delay model is more robust compared to the model proposed in [25]. The authors of [17] concluded that the time delay of insulin responding to glucose stimulation plays a key role in generating the oscillatory behavior of insulin secretion.

This paper attempts to provide a systematical study of the two-delay model of [17] with focuses on analytical studies, bifurcation analysis, and carefully designed numerical simulations. In the following sections, we first introduce the model proposed in [17], then present some preliminary results on positivity, boundedness, and persistence of solutions. Local stability analyses are carried out in details whenever feasible. These analytical results are complemented and confirmed by the bifurcation diagrams produced from our extensive and carefully designed simulations. This paper ends with a discussion section containing a list of observations.

2. The two-delay model. By applying the mass conservation law, the approach used in [27], Li, Kuang, and Mason [17] proposed a glucose-insulin regulatory system model with two explicit time delays based on a set of well-known observations [1], [6], [13], [17], [21], [25], [26], [27]. The model can be expressed by the following word equations.

Glucose change rate = glucose production rate − glucose utilization rate.
Insulin change rate = insulin production rate − insulin removal rate.

Throughout this paper, we use $G(t)$ to represent the plasma glucose concentration and $I(t)$ to represent the plasma insulin concentration at time $t \geq 0$.

In the glucose-insulin endocrine metabolic system, the β-cells, contained in the Langerhans islets in the pancreas, are the only source of insulin production. When the plasma glucose concentration level is elevated, the β-cells secrete insulin after a complex series of cascading physiological processes [1], [17]. The newly synthesized insulin crosses the endothelial barriers to become remote insulin, which readily helps the cells, e.g., muscle and fat cells, to utilize the plasma glucose and convert it to energy [17]. These processes take a total of approximately 5–15 minutes [25], [26]. In the model, this is denoted by $f_1(G(t-\tau_1))$, where $\tau_1 > 0$ represents the time delay of the insulin response to the glucose stimulation and the time needed for the newly synthesized insulin crossing endothelial barrier to become remote insulin.

Insulin is degraded by all insulin sensitive tissues, and the degradation is mediated primarily by the insulin receptor with a smaller contribution from nonspecific

processes. The liver and kidney are the primary sites of portal insulin degradation and peripheral insulin clearance, respectively. Insulin not cleared by the liver and kidneys is ultimately removed mainly by muscle and adipose cells [11], [17]. The function of insulin clearance is to remove and inactivate circulating insulin in order to control insulin action [11]. The degradation is almost linearly proportional to insulin [27]. So the degradation rate is denoted by a constant $d_i > 0$. Since $I(t)$ stands for plasma insulin concentration, it is easier to measure clinically.

In muscle and adipose tissue, insulin facilitates the transport of glucose into cells. The glucose is then metabolized by the cells. This type of glucose consumption is called insulin-dependent glucose utilization. Not all glucose consumption depends on the attendance of insulin. For example, the brain and nerve cells consume the glucose without the aid of insulin. This is referred to as insulin-independent glucose utilization. (See [17] for more details.) Respectively, the insulin-independent and insulin-dependent glucose utilization are represented by $f_2(G(t))$ and $f_3(G(t))f_4(I(t))$.

Glucose enters the circulation in two ways: glucose infusion and hepatic glucose production. Glucose infusion includes meal ingestion, oral glucose intake, continuous enteral nutrition absorption, and constant infusion. Hence the model includes a constant glucose infusion term G_{in} that may model the continuous enteral glucose absorption and constant glucose infusion [17], [25], [27].

Hepatic glucose production is due to glucose dispensation by the liver endogenously. When the plasma glucose concentration level becomes low, the β-cells stop releasing insulin. Instead, the α-cells, also contained in Langerhans islets, start to release glucagon. Glucagon exerts control over pivotal metabolic pathways in the liver and leads the liver to dispense glucose [1]. The liver also converts the previously stored glycogen into glucose. Opposite to the fact that glucagon secretion triggers the liver to dispense glucose, insulin secretion inhibits glucose production by the liver [6], [21]. Thus the hepatic glucose production is primarily controlled by insulin concentration in both inhibitory effect by insulin secretion and recovery effect by insulin suppression. Some time is needed for hepatic glucose production to take significant effect, e.g., half maximal suppression or recovery takes time [17], [21]. However, both the pathways and the length of the delay remain unknown. Nevertheless, this time delay is approximately between a few minutes and a half hour, or even longer [17], [21]. The hepatic glucose production is presented by $f_5(I(t-\tau_2))$, where $\tau_2 > 0$ represents the time taken for a noticeable effect on hepatic glucose production, e.g., half maximal suppression or recovery rate.

Therefore the system [17] can be written as

$$\begin{cases} G'(t) = G_{in} - f_2(G(t)) - f_3(G(t))f_4(I(t)) + f_5(I(t-\tau_2)), \\ I'(t) = f_1(G(t-\tau_1)) - d_i I(t). \end{cases} \tag{2.1}$$

For convenience, the initial conditions of the two-time-delay model (2.1) are assumed to be $I(0) = I_0 > 0$, $G(0) = G_0 > 0, G(t) \equiv G_0$ for all $t \in [-\tau_1, 0]$ and $I(t) \equiv I_0$ for $t \in [-\tau_2, 0]$, $\tau_1, \tau_2 > 0$. In this paper, we assume that the functions f_i, $i = 1, 2, 3, 4, 5$, in model (2.1) satisfy the following conditions:

(i) Notice that β-cells release insulin due to glucose stimulation. We assume that $f_1(x) > 0$ and $f_1'(x) > 0$ for $x > 0$. On the other hand, the capacity of the insulin secretion from β-cells is saturated due to highly increased glucose concentration level, so we assume $\lim_{x\to\infty} f_1(x) = M_1$ and $f_1'(x)$ is bounded by a constant $M_1' > 0$ for $x > 0$. Thus it is reasonable to assume that $f_1(x)$ is in sigmoidal shape. Observing that the insulin can be secreted from the β-cells due to bursting without the glucose stimulation, we assume that $f_1(0) := m_1 > 0$.

(ii) As a term indicating the insulin-independent glucose utilization, it is clear that $f_2(0) = 0$, $f_2(x) > 0$, and $f_2'(x) > 0$ for $x > 0$. On the other hand, the utilization is not unlimited, so we assume that $\lim_{x\to\infty} f_2(x) = M_2$ and there exists a constant M_2' such that $f_2'(x) < M_2'$ for $x > 0$.

(iii) The insulin-dependent glucose utilization $f_3(G(t))f_4(I(t))$ can be depicted as $f_3(0) = 0$, $f_4(0) := m_4 > 0$, $f_3'(x) > 0$, $f_4(x) > 0$, and $f_4'(x) > 0$ for $x > 0$. As suggested by Sturis et al. [25], we also assume that there exist constants $k_3 > 0, M_4 > 0$, and $M_4' > 0$ such that $0 < f_3(x) \le k_3 x$, $\lim_{x\to\infty} f_4(x) = M_4$, and $f_4'(x) < M_4'$ for $x > 0$ and $f_4(x)$ is in sigmoidal shape.

(iv) Low glucose concentration will lead β-cells to stop releasing insulin and α-cells to release glucagon. Thus, when insulin is deficit, liver dispenses glucose caused by glucagon exerting control over pivotal metabolic pathways in the liver, and also converts glycogen into glucose. On the other hand, the liver stops this process when insulin is abundant. Hence we assume $f_5(0) > 0$, $f_5(x) > 0$, and $f_5'(x) < 0$ for $x > 0$, and $\lim_{x\to\infty} f_5(x) = 0$ and f_5 is in inverse sigmoidal shape. Since the amount of glucose converted by the liver is small and the process takes time, we assume $\exists M_5, M_5' > 0$ such that $f_5(x) \le M_5$ and $|f_5'(x)| \le M_5'$ for $x > 0$. We can simply set $M_5 = f_5(0)$.

The shapes of the functions are more important than their forms [13]. Figure 3 of [17] shows the shapes of functions in model (2.1). In section 6, we adopt the functions, (6.1)–(6.5), used in [17], [25], and [26], to perform numerical simulations.

3. Preliminaries. We first present some useful preliminary results of model (2.1). The following fluctuation lemma is elementary and well known [12].

LEMMA 3.1 (fluctuation lemma). *Let $f : \mathbf{R} \to \mathbf{R}$ be a differentiable function. If $l = \liminf_{t\to\infty} f(t) < \limsup_{t\to\infty} f(t) = L$, then there are sequences $\{t_k\} \uparrow \infty$, $\{s_k\} \uparrow \infty$ such that for all k, $f'(t_k) = f'(s_k) = 0, \lim_{k\to\infty} f(t_k) = L$, and $\lim_{k\to\infty} f(s_k) = l$.*

We will apply Lemma 3.1 in the proofs of Proposition 3.2 on solution boundedness and Lemma 3.3 on a set of restrictions for the upper and lower limits of a solution. The proofs are given in Appendices A and B.

PROPOSITION 3.2. *In model (2.1), the following hold:*

(i) *If $\lim_{x\to\infty} f_3(x) > (G_{in} - M_2 + M_5)/m_4$, then model (2.1) has unique positive steady state (G^*, I^*) with $I^* = d_i^{-1} f_1(G^*)$. Furthermore, all solutions exist in $(0, \infty)$, and are positive and bounded.*

(ii) *If $\lim_{x\to\infty} f_3(x) < (G_{in} - M_2)/m_4$, then $\limsup_{t\to\infty} G(t) = \infty$.*

Remark. Condition (i) indicates that the steady state is unique if insulin can help the cells to metabolize enough glucose. Otherwise, if condition (ii) holds, the glucose concentration level will be unbounded.

Remark. If $f_3(x) = k_3 x$, where $k_3 > 0$ is a constant, then

$$\limsup_{t\to\infty} G(t) \le M_G := (G_{in} + M_5)/(m_4 k_3). \tag{3.1}$$

In fact, notice that $m_4 \le f_4(x) \le M_4$ and $f_5(x) \le M_5$ and $f_3(x) = k_3 x$ for $x > 0$. Thus $G'(t) = G_{in} - f_2(G(t)) - f_3(G(t))f_4(I(t)) + f_5(I(t-\tau_2)) \le G_{in} - m_4 k_3 G(t) + M_5$. A standard comparison argument yields (3.1).

Throughout this paper, we assume that condition (i) in Proposition 3.2 holds.

Let $(G(t), I(t))$ be a solution of (2.1). Throughout this paper, we denote

$$\overline{G} = \limsup_{t\to\infty} G(t), \qquad \underline{G} = \liminf_{t\to\infty} G(t), \qquad \overline{I} = \limsup_{t\to\infty} I(t), \qquad \underline{I} = \liminf_{t\to\infty} I(t).$$

Due to Proposition 3.2, it is clear that these limits are finite. Further, we have the following lemma.

LEMMA 3.3. *If* $(G(t), I(t))$ *is a solution of* (2.1)*, then*

$$f_1(\underline{G}) \le d_i \underline{I} \le d_i \bar{I} \le f_1(\bar{G}), \tag{3.2}$$

$$f_2(\bar{G}) + f_3(\bar{G}) f_4(\underline{I}) \le G_{in} + f_5(\underline{I}), \tag{3.3}$$

$$G_{in} + f_5(\bar{I}) \le f_2(\underline{G}) + f_3(\underline{G}) f_4(\bar{I}). \tag{3.4}$$

Remark. Apparently, $\bar{G} = \underline{G}$ implies $\bar{I} = \underline{I}$ due to (3.2). If $\bar{I} = \underline{I}$, then (3.3) and (3.4) together lead to $f_2(\bar{G}) - f_2(\underline{G}) \le f_4(\bar{I})(f_3(\underline{G}) - f_3(\bar{G})) \le 0$. That is, $\bar{G} = \underline{G}$.

PROPOSITION 3.4. *Model* (2.1) *is persistent, that is, the components of all solutions are eventually uniformly bounded from above and away from zero.*

Proof. Notice that $f_2(0) + f_3(0) = 0$ and $f_4(x) < M_4$ for all $x \ge 0$. Then (3.4) implies that $G_{in} \le f_2(\underline{G}) + f_3(\underline{G}) M_4$ for all $t > 0$. Thus $\exists \delta_G > 0, t_G > 0$, such that $G(t) > \delta_G$ for $t > t_G > 0$. Hence $G(t)$ is eventually and uniformly bounded away from zero. Inequality (3.2) implies the same for $I(t)$. The parts of boundedness from above are implied in Proposition 3.2. ∎

4. Local analysis: Case $\tau_1\tau_2 = 0$. We analyze the trivial case $\tau_1\tau_2 = 0$ in this section. The study of the nontrivial case $\tau_1\tau_2 > 0$ will be carried out in the next section.

Clearly the linearized system of model (2.1) about (G^*, I^*) is given by

$$\begin{cases} G'(t) = -AG(t) - BI(t) - CI(t - \tau_2), \\ I'(t) = DG(t - \tau_1) - d_i I_1(t), \end{cases} \tag{4.1}$$

where

$$\begin{cases} A := f_2'(G^*) + f_3'(G^*) f_4(I^*) > 0, \ B := f_3(G^*) f_4'(I^*) > 0, \\ C := -f_5'(I^*) > 0, \ D := f_1'(G^*) > 0. \end{cases} \tag{4.2}$$

The characteristic equation of (4.1) is given by

$$\Delta(\lambda) = \lambda^2 + (A + d_i)\lambda + d_i A + DBe^{-\lambda\tau_1} + DCe^{-\lambda(\tau_1+\tau_2)} = 0. \tag{4.3}$$

Notice that $\Delta(0) = d_i A + DB + DC > 0$. So $\lambda = 0$ is not a solution of the characteristic equation (4.3). Thus, if there is any stability switch of the trivial solution of the linearized system (4.1), there must exist a pair of pure conjugate imaginary roots of the characteristic equation (4.3).

When $\tau_1 = \tau_2 = 0$, the original model (2.1) is an ODE model. The characteristic equation of its linearized equation is given by

$$\Delta(\lambda) = \lambda^2 + (A + d_i)\lambda + d_i A + DB + DC = 0.$$

Then, $A + d_i > 0$ and $d_i A + DB + DC > 0$ imply that (G^*, I^*) is stable.

For the cases of $\tau_1\tau_2 = 0$ and $\tau_1 + \tau_2 > 0$, we need the following lemma, which can be obtained via a standard imaginary root crossing method [8], [15]. The details can be found in [15, pp. 74–77] and [15, Theorem 3.1].

LEMMA 4.1. *Assume* $a, c, d > 0$ *in the following delay differential equation:*

$$x''(t) + ax'(t) + cx(t) + dx(t - \tau) = 0, \ \tau \ge 0. \tag{4.4}$$

Then the number of pairs of pure imaginary roots of the characteristic equation

$$\lambda^2 + a\lambda + c + de^{-\lambda\tau} = 0, \ \tau \ge 0, \tag{4.5}$$

can be zero, one, or two only.

(i) *If* $c > d$ *and* $2c - a^2 < 2\sqrt{c^2 - d^2}$, *then the number of such roots is zero for* $\tau > 0$. *The trivial solution of* (4.4) *is stable for all* $\tau > 0$.

(ii) *If* $c < d$ *or* $d = c$ *and* $2c - a^2 > 0$, *then the number of such roots is one for some* $\tau > 0$. *The trivial solution of* (4.4) *is uniformly asymptotically stable for* $\tau < \tau_0$, *and it becomes unstable for* $\tau > \tau_0$, *where* $\tau_0 > 0$ *is a constant. It undergoes a supercritical Hopf bifurcation at* $\tau = \tau_0$.

(iii) *If* $c > d$ *and* $2c - a^2 > 2\sqrt{c^2 - d^2}$, *then the number of such roots is two for some* $\tau > 0$. *The stability of the trivial solution of* (4.4) *can change (when changing from stable to unstable, the trivial solution undergoes a supercritical Hopf bifurcation) a finite number of times at most as* τ *increases, and eventually it becomes unstable.*

For the case of $\tau_1 > 0$ and $\tau_2 = 0$, the characteristic equation is

$$\Delta(\lambda) = \lambda^2 + (A + d_i)\lambda + d_i A + (DB + DC)e^{-\lambda\tau_1} = 0. \tag{4.6}$$

Notice that, in this case, $2c - a^2 = -A^2 - d_i^2 < 0$ in Lemma 4.1. Then $d_i A > D(B + C)$ implies that the trivial solution of (4.1) is always stable for $\tau_1 > 0$. Also, $d_i A < D(B + C)$ implies that $\exists \tau_{10} > 0$ such that the trivial solution of the linearized system (4.1) is stable when $\tau_1 \in (0, \tau_{10})$ and unstable when $\tau_1 \geq \tau_{10}$.

For the case of $\tau_1 = 0$ and $\tau_2 > 0$, the characteristic equation becomes

$$\Delta(\lambda) = \lambda^2 + (A + d_i)\lambda + (d_i A + DB) + DCe^{-\lambda\tau_2} = 0. \tag{4.7}$$

In Lemma 4.1, $2c - a^2 = 2DB - A^2 - d_i^2$. Thus if $d_i^2 > 2DB - A^2$ and $d_i A > D(C - B)$, the trivial solution of (4.1) is stable for all $\tau_2 > 0$. If $d_i A < D(C - B)$, then the stability of the trivial solution of (4.1) switches from stable to unstable when τ_2 increases through a critical value $\tau_{20} > 0$ and remains unstable for $\tau_2 > \tau_{20}$. If $d_i A > D(C - B)$ and $2DB - A^2 - d_i^2 > 2\sqrt{(d_i A + DB)^2 - D^2C^2}$, then the trivial solution of the linearized system (4.1) has at most a finite number of stability switches and eventually is unstable.

Define

$$H_1(d_i, G_{in}) = D(B + C) - d_i A. \tag{4.8}$$

We summarize the above analysis in the following theorem for model (2.1).

THEOREM 4.2. *Consider model* (2.1).

(a) *If* $\tau_1 = 0$ *and* $\tau_2 = 0$, *then* (G^*, I^*) *is stable.*

(b) *If* $\tau_1 > 0$ *and* $\tau_2 = 0$, *and*

(b.1) *if* $H_1(d_i, G_{in}) < 0$, *then* (G^*, I^*) *is stable for* $\tau_1 > 0$;

(b.2) *if* $H_1(d_i, G_{in}) > 0$, *then* $\exists \tau_{10} > 0$ *such that* (G^*, I^*) *is stable when* $\tau_1 \in (0, \tau_{10})$ *and unstable when* $\tau_1 \geq \tau_{10}$.

(c) *When* $\tau_1 = 0$ *and* $\tau_2 > 0$,

(c.1) *if* $D(C - B) - d_i A < 0$ *and* $d_i^2 > 2DB - A^2$, *then* (G^*, I^*) *is stable;*

(c.2) *if* $D(C - B) - d_i A > 0$, *then* $\exists \tau_{20} > 0$ *such that* (G^*, I^*) *is stable when* $\tau_2 \in (0, \tau_{20})$ *and unstable when* $\tau_1 \geq \tau_{20}$;

(c.3) *if* $D(C - B) - d_i A < 0$ *and* $2DB - A^2 - d_i^2 > 2\sqrt{(d_i A + DB)^2 - D^2C^2}$, *then there are at most a finite number of stability switches and eventually* (G^*, I^*) *is unstable,*

where A, B, C, *and* D *are given in* (4.2).

With the specific functions (6.1)–(6.5) in section 6, Figure 5.1 (right) demonstrates the curve $H_1(d_i, G_{in}) = 0$ in the (G_{in}, d_i)-plane. The curves in Figure 5.1 (right) are

independent of delay τ_1 and τ_2. The steady state is stable when $(G_{in}, d_i) \in R_s$, that is, d_i is small. Our computations show that conditions (b.1) and (c.3) in Theorem 4.2 do not hold. Condition (c.1) holds for some values (G_{in}, d_i), and thus the sustained oscillations would not occur.

When condition (b.2) ($\tau_2 = 0$) or (c.2) ($\tau_1 = 0$) holds, the sustained oscillation takes place if $\tau_1 > \tau_{10}$ or $\tau_2 > \tau_{20}$. Based on the arguments (3.12)–(3.17) from [15, pp. 74–76], we have

$$\tau_{10} = \theta_1/\omega_{1+},$$

where ω_{1+} is the root of (4.6) given by

$$\omega_{1+}^2 = \frac{1}{2}\Big\{ -(A^2 + d_i^2) + \Big[(A^2 - d_i^2)^2 + 4D^2(B + C)^2\Big]^{-\frac{1}{2}}\Big\},$$

and $0 \le \theta_1 < 2\pi$, satisfying

$$\begin{cases} \cos\theta_1 = (\omega_{1+}^2 - d_i A)/(DB + DC), \\ \sin\theta_1 = \omega_{1+}(A + d_i)/(DB + DC). \end{cases}$$

Similarly, we have

$$\tau_{20} = \theta_2/\omega_{2+},$$

where ω_{2+} is the root of (4.7) given by

$$\omega_{2+}^2 = \frac{1}{2}\Big\{2DB - (A^2 + di^2) + \Big[(A^2 - d_i^2)^2 - 4DB(A + d_i)^2 + 4D^2C^2\Big]^{-\frac{1}{2}}\Big\},$$

and $0 \le \theta_1 < 2\pi$, satisfying

$$\begin{cases} \cos\theta_2 = (\omega_{2+}^2 - d_i A - DB)/(DC), \\ \sin\theta_2 = \omega_{2+}(A + d_i)/(DC). \end{cases}$$

When condition (b.2) ($\tau_2 = 0$) holds, our computations show that no sustained oscillation occurs when $\tau_1 < 9$. Similarly, when (c.2) ($\tau_1 = 0$) holds, $\tau_{20} > 12$ for $G_{in} < 0.15$ or $G_{in} > 0.85$, and $d_i = 0.06$. Specifically, if $G_{in} = 1.35, d_i = 0.06$, then the Hopf bifurcation point $\tau_{20} > 40$. These observations, with Theorem 4.2 (a), suggest that both delay τ_1 and τ_2 are critical for sustained oscillations in physiologically meaningful ranges. In addition, notice that condition (b.2) automatically holds if (c.2) holds. This seems to suggest that the role of delay τ_1 is more critical than the role of delay τ_2 to ensure the sustained oscillations of the glucose-insulin regulatory system.

5. Local analysis: Case $\tau_1\tau_2 > 0$. Now assume both $\tau_1 > 0$ and $\tau_2 > 0$. Let $\lambda = \omega i$, $\omega > 0$, be such an eigenvalue in (4.3); then we have

$$\begin{aligned}\Delta(\omega i) = {} & [-\omega^2 + d_i A + DB\cos\omega\tau_1 + DC\cos\omega(\tau_1 + \tau_2)] \\ & + i[(A + d_i)\omega - DB\sin\omega\tau_1 - DC\sin\omega(\tau_1 + \tau_2)] = 0.\end{aligned}$$

That is,

$$\begin{cases} -\omega^2 + d_i A + DB\cos\omega\tau_1 + DC\cos\omega(\tau_1 + \tau_2) = 0, \\ (A + d_i)\omega - DB\sin\omega\tau_1 - DC\sin\omega(\tau_1 + \tau_2) = 0. \end{cases} \tag{5.1}$$

This leads to

$$\omega^4 + (A^2 + d_i^2)\omega^2 + d_i^2 A^2 = D^2(B^2 + C^2 + 2BC\cos\omega\tau_2). \tag{5.2}$$

5.1. Stability of the steady state. We shall consider the stability of the steady state first. From (5.2),

$$\omega^4 + (A^2 + d_i^2)\omega^2 + d_i^2 A^2 \le D^2(B^2 + C^2 + 2BC) = D^2(B + C)^2.$$

It is impossible if $d_i A \ge D(B + C)$. So, by definition of $H_1(G_{in}, d_i)$ in (4.8), we have the following proposition.

PROPOSITION 5.1. *In the linearized system* (4.1)*, when* $\tau_1 > 0$ *and* $\tau_2 > 0$*, if* $H_1(G_{in}, d_i) \le 0$*, then the steady state of the linearized system* (4.1) *is stable.*

Therefore we have the following theorem.

THEOREM 5.2. *In model* (2.1)*, if* $\tau_1 > 0$*,* $\tau_2 > 0$ *and*

$$H_1(d_i, G_{in}) = D(B + C) - d_i A \le 0, \tag{5.3}$$

then the steady state (G^*, I^*) *of system* (2.1) *is stable.*

Remark. When $\tau_1 > 0$, the same condition $H_1(d_i, G_{in}) = D(B + C) - d_i A < 0$ ensures the steady state of system (2.1) to be stable regardless of whether $\tau_2 = 0$ or $\tau_2 > 0$.

5.2. Instability of the steady state. We now study the instability of the steady state (G^*, I^*). We will apply *Rouchè's theorem* [7, pp. 125–126] to identify the case that the characteristic equation (4.3) has roots with positive real part.

ROUCHÈ'S THEOREM. *Given two functions* $f(z)$ *and* $g(z)$ *analytic in a simple connected region* $\mathcal{A} \subset \mathbf{C}$ *with boundary* γ*, a simple loop homotopic to a point in* $\mathcal{A}$*, if* $|f(z)| > |g(z)|$ *on* γ*, then* $f(z)$ *and* $f(z) + g(z)$ *have the same number of zeros in* $\mathcal{A}$*.*

Let

$$S_1 = \left\{ \frac{2m}{2n-1} : m, n \in \mathbf{Z}^+, m, n \ge 1 \right\} \text{ and } S_2 = \left\{ \frac{2m-1}{2n} : m, n \in \mathbf{Z}^+, m, n \ge 1 \right\}.$$

Clearly $\mathbf{Q}^+ = S_1 \cup S_2$ and $S_1 \cap S_2 = \emptyset$. Furthermore, S_1 and S_2 are dense in $\mathbf{Q}^+$, and thus in $\mathbf{R}^+$.

In fact, given $r \in \mathbf{Q}^+ \setminus S_1, \exists p, q \in \mathbf{Z}^+$ such that $r = \frac{2p-1}{2q}$. Thus

$$r_k = \frac{2p - 1 - \frac{2}{2k}}{2q - \frac{1}{2k}} = \frac{(4kp - 2k - 2)/2k}{(4kq - 1)/2k} = \frac{2(2kp - 2k - 1)}{2(2kq) - 1} \in S_1 \quad \forall k = 1, 2, 3, \ldots$$

and $\lim_{k \to \infty} r_k = (2p - 1)/2q = r$. That is, $\overline{S}_1 \supseteq \mathbf{Q}^+$. Similarly, $\overline{S}_2 \supseteq \mathbf{Q}^+$.

PROPOSITION 5.3. *For characteristic equation*

$$\lambda^k + \sum_{j=1}^{k-1} a_j \lambda^j + b + ce^{-\lambda\sigma_1} + de^{-\lambda\sigma_2} = 0, \quad k \ge 2,\ \sigma_1, \sigma_2 > 0, \tag{5.4}$$

where $b, c, d > 0, a_j \in \mathbf{R}, j = 1, 2, 3, \ldots, k$*, if* $b < d - c$ *or* $b < c - d$*, then* $\exists \sigma_{10} > 0$ *and* $\sigma_{20} > 0$ *such that the characteristic equation* (5.4) *has at least one root with positive real part for* $\sigma_1 > \sigma_{10}$ *and* $\sigma_2 > \sigma_{20}$ *provided* $\sigma_1/\sigma_2 \in S_1$ *or* $\sigma_1/\sigma_2 \in S_2$*.*

We need the following lemma to prove Proposition 5.3.

LEMMA 5.4. *For the equation*

$$\epsilon^k z^k + \sum_{j=1}^{k-1} a_j \epsilon^j z^j + b + ce^{-p_1 z} + de^{-p_2 z} = 0, \quad k \ge 2,\ p_1, p_2 > 0,\ z \in \mathbf{C}, \tag{5.5}$$

where $b, c, d > 0, a_j \in \mathbf{R}, j = 1, 2, 3, \ldots, k - 1$*, assume*

(i) $b < d - c$ *and* $p_1/p_2 \in S_1$, *or*
(ii) $b < c - d$ *and* $p_1/p_2 \in S_2$.

Then, $\exists \epsilon_0 > 0$ *such that for all* $\epsilon, 0 < \epsilon < \epsilon_0$, *equation* (5.5) *has at least one root with positive real part.*

The proof of Lemma 5.4 is given in Appendix C. Now we prove Proposition 5.3.

Proof. Assume $b < d - c$ and $\sigma_1/\sigma_2 \in S_1$ (or $b < c - d$ and $\sigma_1/\sigma_2 \in S_2$). In Lemma 5.4, choose p_{10} and p_{20} such that $p_{10}/p_{20} \in S_1$ (or $p_{10}/p_{20} \in S_2$). Suppose ϵ_0 is given by inequality (C.2) in the proof of Lemma 5.4. Let $\sigma_{10} = p_{10}/\epsilon_0$ and $\sigma_{20} = p_{20}/\epsilon_0$. Then given $\sigma_1 > \sigma_{10}$, $\sigma_2 > \sigma_{20}$, and $\sigma_1/\sigma_2 \in S_1$ (or $\sigma_1/\sigma_2 \in S_2$), $\exists \epsilon$, $0 < \epsilon < \epsilon_0$, such that

$$\sigma_1 = p_1/\epsilon > \sigma_{10} \quad \text{and} \quad \sigma_2 = p_2/\epsilon > \sigma_{20}.$$

Let $\lambda = \epsilon z$. Then (5.4) becomes (5.5) in Lemma 5.4 and the conclusion follows. ∎

Remark. In Lemma 5.4, given p_1 and p_2, $p_1/p_2 \in S_1$ in case (i) or $p_1/p_2 \in S_2$ in case (ii), if we carefully choose ϵ_0 in the proof of Lemma 5.4, an estimate of unstable region of σ_1 and σ_2 can be given. For the special case $k = 2$, r_0 and ϵ_0 can be chosen as

$$r_0 = \sqrt{K^2 x_0^2 + q^2 \pi^2} \quad \text{and} \quad \epsilon_0 = \left(\sqrt{a_1^2 + 4\eta_0'} - a_1 \right) / 2r_0.$$

Let $k = 2$ and apply Proposition 5.3 to the linearized system (4.1). We have the following.

PROPOSITION 5.5. *If* $d_i A < D|C - B|$, *then there exist* $\tau_{10} > 0$ *and* $\tau_{20} > 0$ *such that the characteristic equation of system* (4.1) *has at least one root with positive real part if*

(i) $d_i A < D(C - B)$, $\tau_1 > \tau_{10}$, $\tau_1 + \tau_2 > \tau_{20}$, *and* $\tau_1/(\tau_1 + \tau_2) \in S_1$, *or*
(ii) $d_i A < D(B - C)$, $\tau_1 > \tau_{10}$, $\tau_1 + \tau_2 > \tau_{20}$, *and* $\tau_1/(\tau_1 + \tau_2) \in S_2$.

Proof. This is straightforward if in Proposition 5.3 we choose $k = 2, a_1 = A + d_i, b = d_i A, c = DB, d = DC, \sigma_1 = \tau_1$, and $\sigma_2 = \tau_1 + \tau_2$. ∎

Therefore, we have the following theorem.

THEOREM 5.6. *In model* (2.1), *let*

$$H_2(d_i, G_{in}) := D|C - B| - d_i A. \tag{5.6}$$

If $\tau_1 > 0$, $\tau_2 > 0$, *and* $H_2(d_i, G_{in}) > 0$, *then there exist* $\tau_{10} > 0$ *and* $\tau_{20} > 0$ *such that the steady state* (G^*, I^*) *is unstable when* $\tau_1 > \tau_{10}$, $\tau_1 + \tau_2 > \tau_{20}$ *and*

(i) $\tau_1/(\tau_1 + \tau_2) \in S_1$ *and* $d_i A < D(C - B)$, *or*
(ii) $\tau_1/(\tau_1 + \tau_2) \in S_2$ *and* $d_i A < D(B - C)$.

Remark. Using the function (6.1)–(6.5), if $G_{in} = 1.35$ and $d_i = 0.06$, then $H_2(d_i, G_{in}) > 0$ and $H_1(d_i, G_{in}) < 0$. According to Theorem 5.6, the steady state will lose its stability as the delays increase. Let $\tau_1 = 7$ and $\tau_2 = 30$. The simulation result is shown in Figure 5.1 (left). There exists a periodic solution bifurcating from the steady state. This periodic solution can be regarded as the sustained oscillation of the insulin and glucose concentration. The period of the periodic solution is approximately 149 minutes. In each cycle, the glucose concentration peaks about 18 minutes ahead of the insulin concentration peaks. The varying range of glucose concentration is within physiological meaningful scope [70, 109].

Remark. It is clear that $H_2(d_i, G_{in}) \leq H_1(d_i, G_{in})$. When $H_1(d_i, G_{in}) \leq 0$, the steady state of model (2.1) is stable due to Theorem 5.2. On the other hand, when $H_2(d_i, G_{in}) > 0$, according to Theorem 5.6, the steady state is unstable for appropriate delay values given in Theorem 5.6. With specific function (6.1)–(6.5), Figure 5.1

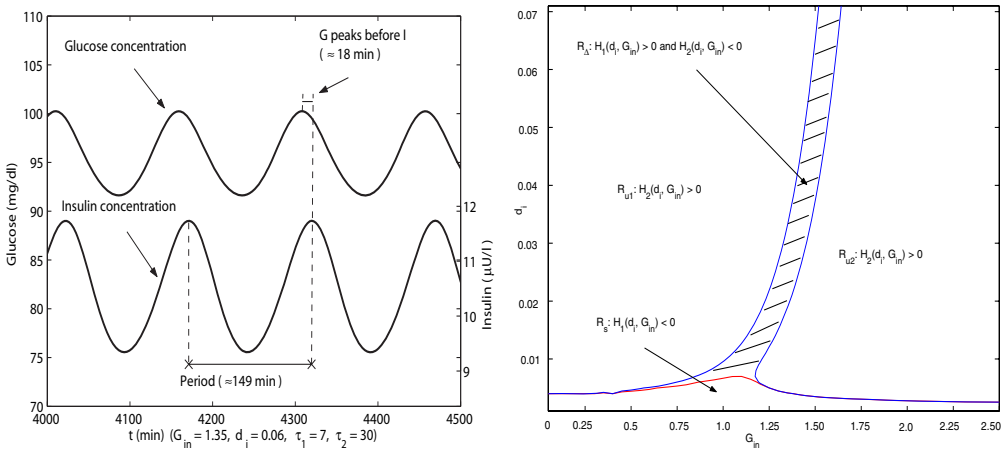


FIG. 5.1. *Left: The periodic solution of model* (2.1) *when* $G_{in} = 1.35, d_i = 0.06, \tau_1 = 7$, *and* $\tau_2 = 30$. *The period is approximately* 149 *minutes and the glucose concentration peaks about* 18 *minutes before the insulin concentration peaks. Right: Regions in the* (G_{in}, d_i)*-plane divided by curves* $H_1(d_i, G_{in}) = 0$ *and* $H_2(d_i, G_{in}) = 0$. *When* $d_i \in R_s$, *the steady state of model* (2.1) *is stable; when* $d_i \in R_{u1} \cup R_{u2}$, *the steady state is unstable.*

(right) shows the delay-independent stable region R_s and delay-dependent unstable region R_{u1} and R_{u2} in the (d_i, G_{in})-plane determined by Theorems 5.2 and 5.6, respectively. The shaded region R_Δ is where $H_1(d_i, G_{in}) > 0$ and $H_2(d_i, G_{in}) \le 0$. The local stability problem of (G^*, I^*) is open when $(d_i, G_{in}) \in R_\Delta$. Our intensive numerical simulations reveal that R_Δ is also a delay-dependent unstable region, that is, with appropriate delay parameters, the steady state is unstable. For example, when $d_i = 0.051$ and $G_{in} = 1.50$, $H_1(0.051, 1.50) = 0.0019$ and $H_2(0.051, 1.50) = -0.00042570$. The steady state is unstable when $\tau_1 \ge 15$ and $\tau_2 \ge 32$. When $d_i = 0.0320$ and $G_{in} = 1.40$, $H_1(0.0320, 1.40) = 0.00099732$ and $H_2(0.051, 1.50) = -0.00026753$. The steady state is unstable when $\tau_1 \ge 18$ and $\tau_2 \ge 36$. Periodic solutions are also observed in these cases.

5.3. Hopf bifurcation. We show below that a local Hopf bifurcation takes place when delay parameter τ_1 or τ_2 varies. It has been shown that when $\tau_1 = 0$ ($\tau_2 = 0$), the steady state of system (2.1) is stable provided that τ_2 (τ_1) is small enough (see Theorem 4.2). To show this system undergoes a unique local Hopf bifurcation at some $\bar\tau_1 > 0$ ($\bar\tau_2 > 0$) as τ_1 (τ_2) increases from 0 and within a physiologically meaningful range, we prove that the characteristic equation (4.3) has a pair of pure conjugate imaginary simple roots at $\tau_1 = \bar\tau_1 > 0$ ($\tau_2 = \bar\tau_2 > 0$) and all such roots cross the imaginary axis from left to right. This indicates that a periodic solution is generated from this stability switch. Our numerical simulations show that the bifurcation is indeed supercritical.

Consider equation

$$\omega^4 + (A^2 + d_i^2)\omega^2 + d_i^2 A^2 = D^2(B^2 + C^2 + 2BC). \tag{5.7}$$

Clearly, (5.7) has a unique positive root $\hat\omega$ when $d_i A < D(B + C)$, where

$$\hat\omega^2 = \frac{1}{2}\Big[-(A^2 + d_i^2) + \sqrt{(A^2 + d_i^2) - 4(d_i^2 A^2 - D^2(B + C)^2)}\Big]. \tag{5.8}$$

Let $g(\omega) = \omega^4 + (A^2 + d_i^2)\omega^2 + d_i^2 A^2 - D^2(B^2 + C^2 + 2BC\cos\omega\tau_2)$. Then (5.2) can be written as $g(\omega) = 0$. If $\omega < \hat{\omega} < \frac{\pi}{2\tau_2}$, then $g(0) = d_i^2 A^2 - D^2(B + C)^2 < 0$ and $g(\hat{\omega}) = 2D^2BC(1 - \cos\omega\tau_2) \geq 0$. Further, $g'(\omega) = 4\omega^3 + 2(A^2 + d_i^2)\omega + 2D^2BC\sin\omega\tau_2 > 0$ for $0 < \omega < \hat{\omega} \leq \frac{\pi}{2\tau_2}$. Therefore we have the following lemma.

LEMMA 5.7. *If $d_iA < D(B + C)$ and $\tau_2 < \frac{\pi}{2\hat{\omega}}$, then (5.2) has a unique root ω_0 with $0 < \omega_0 \leq \hat{\omega}$.*

The following propositions establish sufficient conditions for the existence of Hopf bifurcation when τ_1 or τ_2 varies. We leave the proofs in Appendices D and E.

PROPOSITION 5.8. *If $H_1(G_{in}, d_i) = D(B + C) - d_iA > 0$ and $\tau_1 + \tau_2 < \frac{\pi}{2\hat{\omega}}$, then (4.1) undergoes a Hopf bifurcation when τ_1 increases from 0 to $\frac{\pi}{2\hat{\omega}} - \tau_2$ for given τ_2.*

PROPOSITION 5.9. *If $H_1(G_{in}, d_i) = D(B + C) - d_iA > 0, \tau_1 + \tau_2 < \frac{\pi}{2\hat{\omega}}$, and $\tau_1 < \frac{A+d_i}{DB}$, then (4.1) undergoes a Hopf bifurcation when τ_2 increases from 0 to $\frac{\pi}{2\hat{\omega}} - \tau_1$ for given τ_1.*

Remark. Using the specific functions (6.1)–(6.5) given in section 6, for $(G_{in}, d_i) \in [0, 150] \times [0.001, 0.07]$, approximately, $47.2665 < \frac{\pi}{2\hat{\omega}} < 214.3462$ and $19.6857 < \frac{A+d_i}{DB} < 6361.7$. Thus τ_1 varies within its physiological range under the condition $\tau_1 < \frac{A+d_i}{DB}$. Under the condition $\tau_1 + \tau_2 < \frac{\pi}{2\hat{\omega}}$, both τ_1 and τ_2 are within their physiological meaningful ranges in most situations for $(G_{in}, d_i) \in [0, 150] \times [0.001, 0.07]$. When $\tau_1 + \tau_2$ could be larger than 47.2665, our simulations show that the Hopf bifurcation does exist and it is supercritical.

We summarize the above results in the following theorem.

THEOREM 5.10. *For model (2.1), assume $H_1(G_{in}, d_i) = D(B + C) - d_iA > 0$ and $\tau_1 + \tau_2 < \frac{\pi}{2\hat{\omega}}$.*

(a) *Then there exists a $\bar{\tau}_1 > 0$ such that the steady state (G^*, I^*) is stable when $\tau_1 < \bar{\tau}_1$, and unstable when $\tau_1 \geq \bar{\tau}_1$. The system undergoes a Hopf bifurcation at $\bar{\tau}_1$ and generates a periodic solution.*

(b) *Further, assume $\tau_1 < \frac{A+d_i}{DB}$. Then there exists a $\bar{\tau}_2 > 0$ such that the steady state (G^*, I^*) is stable when $\tau_2 < \bar{\tau}_2$, and unstable when $\tau_2 \geq \bar{\tau}_2$. The system undergoes a Hopf bifurcation at $\bar{\tau}_2$ and generates a periodic solution.*

Remark. With the specific functions (6.1)–(6.5) in the next section, our intensive numerical simulations show that the Hopf bifurcations determined by Theorem 5.10 are supercritical. Moreover, with $G_{in} = 1.35$ and $d_i = 0.06$, $\bar{\tau}_1$ and $\bar{\tau}_2$ approximately satisfy $33.9\bar{\tau}_1 + 17.3\bar{\tau}_2 \approx 36.9$ for $0 \leq \bar{\tau}_1 \leq 20$ and $0 \leq \bar{\tau}_2 \leq 60$.

6. Numerical simulations. In this section, we present numerical analysis on model (2.1) using DDE23 [22] in MATLAB 6.5. We use the same functions $f_i, i = 1, 2, 3, 4, 5$, as [2], [17], [25], and [26] given in (6.1)–(6.5). The parameters, listed in Table 6.1, were generated from experiments [25], [26].

$$f_1(G) = R_m/(1 + \exp((C_1 - G/V_g)/a_1)), \tag{6.1}$$

$$f_2(G) = U_b(1 - \exp(-G/(C_2V_g))), \tag{6.2}$$

$$f_3(G) = G/(C_3V_g), \tag{6.3}$$

$$f_4(I) = U_0 + (U_m - U_0)/(1 + \exp(-\beta\ln(I/C_4(1/V_i + 1/Et_i)))), \tag{6.4}$$

$$f_5(I) = R_g/(1 + \exp(\alpha(I/V_p - C_5))). \tag{6.5}$$

TABLE 6.1
Parameters of the functions in two-time-delay model (2.1).

Parameters	Units	Values	Parameters	Units	Values
V_g	l	10	R_m	$\mu U \min^{-1}$	210
a_1	$\mathrm{mg} \cdot \mathrm{l}^{-1}$	300	C_1	$\mathrm{mg} \cdot \mathrm{l}^{-1}$	2000
U_b	$\mathrm{mg} \cdot \min^{-1}$	72	C_2	$\mathrm{mg} \cdot \mathrm{l}^{-1}$	144
C_3	$\mathrm{mg} \cdot \mathrm{l}^{-1}$	1000	U_0	$\mathrm{mg} \cdot \min^{-1}$	40
U_m	$\mathrm{mg} \cdot \min^{-1}$	940	β		1.77
C_4	$\mu \mathrm{Ul}^{-1}$	80	R_g	$\mathrm{mg} \cdot \min^{-1}$	180
α	$\mathrm{l}\mu \mathrm{U}^{-1}$	0.29	C_5	$\mu \mathrm{Ul}^{-1}$	26

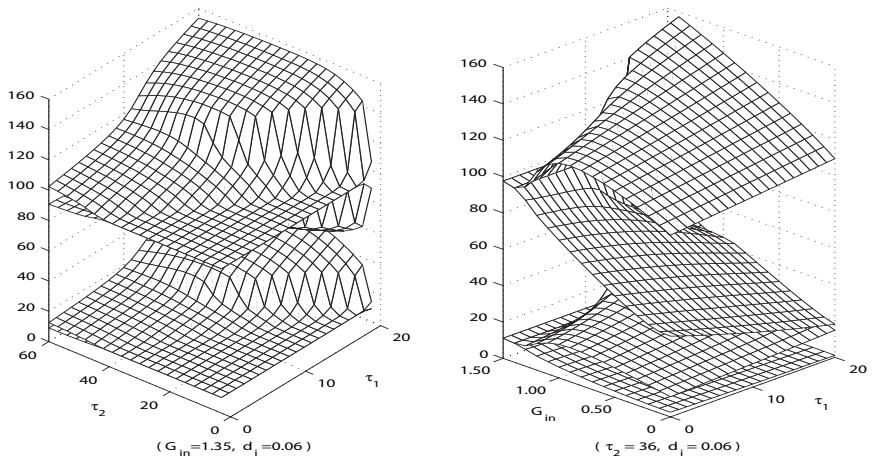


FIG. 6.1. *Limiting values or amplitudes of glucose (top) and insulin (bottom) concentrations when τ_1 and τ_2 vary (left) or G_{in} and τ_1 vary (right).*

The simulations in [17] focused on the bifurcation when a single parameter varies while other parameters are fixed. The detected bifurcation points of the varying parameters can determine when the sustained oscillations occur. In this section, we carry out a sequence of two-parameter bifurcation analyses and depict their quantitative behaviors in three-dimensional meshes or two-dimensional curves formed by transversal points.

For a specific subject, the insulin response time delay, the delayed effect of hepatic glucose production, and the insulin degradation rate are intrinsic. But the exogenous glucose infusion rate is controllable by diet, fasting, and so on. So, in addition to the simulation on the two-delay parameters, we numerically analyze the relationships of the glucose infusion rate G_{in} vs. the insulin response time delay τ_1, the hepatic glucose production time delay τ_2, and the insulin degradation rate d_i, respectively. We end this section by showing the significant impact of the two delays on generating insulin ultradian oscillation.

6.1. Insulin response delay τ_1 vs. hepatic glucose production delay τ_2. We analyzed the relationship between the insulin response delay τ_1 and the hepatic glucose production delay τ_2 while $G_{in} = 1.35$ and $d_i = 0.06$ are fixed. Figure 6.1 (left) shows that a simple curve ($33.9\tau_1 + 17.3\tau_2 \approx 36.9$ for $0 \le \tau_1 \le 20$ and $0 \le \tau_2 \le 60$)

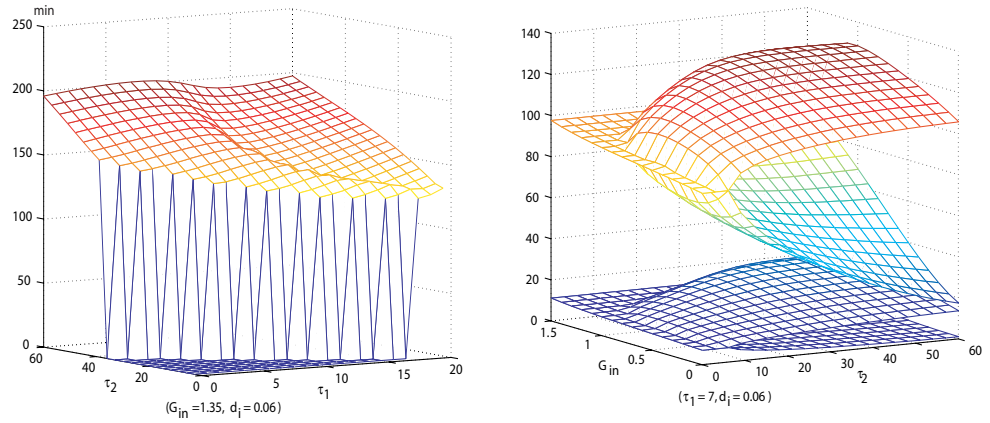


FIG. 6.2. *Left: Mesh of the periods of periodic solutions when* $(\tau_1, \tau_2) \in [0, 20] \times [0, 60]$*, where* $G_{in} = 1.35$ *and* $d_i = 0.06$ *are fixed. Right: Meshes of the amplitudes of glucose and insulin concentrations when* $(G_{in}, \tau_2) \in [0, 1.5] \times [0, 60]$*, where* $\tau_1 = 7$ *and* $d_i = 0.06$ *are fixed.*

divides $[0, 20] \times [0, 60]$ in the (τ_1, τ_2)-plane into two regions. The steady state is stable in one region and unstable in the other. The sustained oscillations occur in the unstable region which requires both $\tau_1 > 0$ and $\tau_2 > 0$ to be sufficiently large. The top and bottom meshes in Figure 6.1 (left) demonstrate the amplitudes of glucose and insulin concentrations, respectively. The periods of periodic solutions are shown in Figure 6.2 (left). According to Figure 6.1 (left) and Figure 6.2 (left), the amplitudes of glucose concentration are between 70 and 109 and the periods of periodic solutions are approximately within 90 and 150 when $\tau_1 \in (5, 15)$ and $\tau_2 \in (25, 50)$. There is a sudden jump of amplitudes of glucose and insulin concentrations when $\tau_1 > 10$ approximately. Also, in such cases, the periods of periodic solutions decrease.

6.2. Glucose infusion rate G_{in} vs. insulin response time delay τ_1. Taking both insulin response delay τ_1 and glucose infusion rate G_{in} as bifurcation parameters, we try to identify the stability regions when $(\tau_1, G_{in}) \in [0, 20] \times [0, 1.5]$. Let $d_i = 0.06$ and $\tau_2 = 36$ be fixed. The computation results are shown in Figure 6.1 (right). The bifurcation point value $\bar{\tau}_1 \approx 1.0429 G_{in} - 1.3740 > 0$ exists for $1.3175 \leq G_{in} \leq 1.50$. The meshes are the amplitudes of glucose (top) and insulin (bottom) concentrations when $(\tau_1, G_{in}) \in [0, 20] \times [0, 1.5]$. It can be seen that a simple curve $(\tau_1 \approx 1.0429 G_{in} - 1.3740 > 0$ for $1.3175 \leq G_{in} \leq 1.50)$ divides the rectangular $[0, 20] \times [0, 1.5]$ in the (τ_1, G_{in})-plane into two regions. The sustained oscillations of model (2.1) occur in the unstable region. The exogenous glucose infusion rate can be larger when τ_1 increases from [5, 15] for sustained regulatory oscillations to occur.

6.3. Glucose infusion rate G_{in} vs. hepatic glucose production delay τ_2. As shown in Figure 6.2 (right), our simulation results indicate that when $\tau_1 = 7$ and $d_i = 0.06$, the rectangular $[0, 60] \times [0, 1.50]$ in the (τ_2, G_{in})-plane is divided by a simple curve into two regions. The steady state of model (2.1) is unstable in the unstable region and the oscillations are sustained. The periods of periodic solutions are in a range of 80 and 195 minutes (not shown). The simple curve is determined by the Hopf bifurcation point values $\bar{\tau}_2(G_{in})$ as G_{in} varies from 0 to 150. The relationship between G_{in} and $\bar{\tau}_2$ is nonlinear. For example, $\bar{\tau}_2 \approx 6.1, 2.8, 6.1, 9, 12, 18, 33$ when

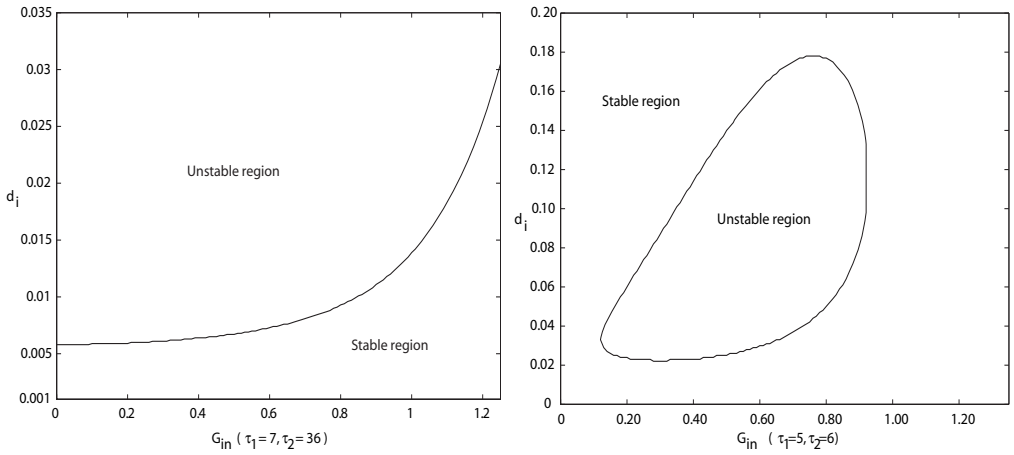


FIG. 6.3. *Left: Insulin degradation rate d_i vs. glucose infusion rate G_{in} while $\tau_1 = 7$ and $\tau_2 = 36$. The steady state of model (2.1) is stable in one region and unstable in the other. Right: Stable and unstable regions of the steady state when $\tau_1 = 5$ and $\tau_2 = 6$.*

$G_{in} = 0, 0.60, 0.80, 1.10, 1.20, 1.30, 1.40$, respectively. Similar to the case of G_{in} vs. τ_1, the exogenous glucose infusion rate can be larger when τ_2 increases from 10 to 60 minutes for sustained regulatory oscillations to occur. When $\tau_2 < 2$, the steady state is stable and no sustained oscillation will occur regardless of what value G_{in} assumes.

6.4. Glucose infusion rate G_{in} vs. insulin degradation rate d_i. Similarly, taking both glucose infusion rate G_{in} and insulin degradation rate d_i as bifurcation parameters while $\tau_1 = 7$ and $\tau_2 = 36$ are fixed, we identified the stability regions in $(G_{in}, d_i) \in [0, 1.35] \times [0.001, 0.20]$. Figure 6.3 shows that a simple curve divides the rectangular $(G_{in}, d_i) \in [0, 1.35] \times [0.001, 0.20]$ into two regions (the figure shows the part of $[0, 1.35] \times [0.001, 0.035]$ only). The steady state of model (2.1) is stable in one region and unstable in the other region. It is clear that larger insulin degradation rate d_i facilitates the oscillatory regulation. However, if $d_i = 1.75$ is large, then no self-sustained oscillation occurs. This suggests that d_i needs to be in moderate range for oscillations to be sustained. Let $\tau_1 = 5$ and $\tau_2 = 6$ be smaller. Figure 6.3 (right) shows that the sustained oscillation occurs in a region surrounded by a closed curve, which requires both d_i and G_{in} to be in moderate ranges. Our simulation shows that the amplitudes of the sustained oscillations are very small ($G \in (80, 100)$ and $I \in (10, 12)$) with periods from 58 to 105 minutes. This shows that when the delays τ_1 and τ_2 are smaller, both the insulin clearance rate and the glucose infusion rate have to be in a moderate range to ensure the oscillatory behavior of insulin secretion.

7. Discussions. In this paper, we studied the glucose-insulin regulatory system model (2.1) analytically and numerically. Compared with the observations obtained in [2], [17], [25], and [26], our work confirms most of the known observations and yields additional insightful information. Using the notation in [17], we refer to the observations in [25] and [26] as [STx] ([ST1]–[ST4]), the observations in [2] as [BGx] ([BG1] and [BG2]), and the observations in [17] as [Ax] ([A1]–[A9]). We conclude this paper with a list of remarks and new observations (denoted by [Bx]).

[B1] Theorem 4.2 reveals that the delays in the glucose-insulin regulatory system are critical for ensuring the sustained oscillations of regulation and

insulin secretions. Particularly, the role of delay of insulin secretion and the newly synthesized insulin becoming remote insulin is more critical than the role of delay of hepatic glucose production.

[B2] If the insulin secretion responds to elevated glucose instantaneously, that is, $\tau_1 = 0$, Theorem 4.2 (c.1) and (c.2) reveal that the insulin degradation rate for sustaining oscillation is likely to be lower than that in the case of $\tau_1 > 0$ and $\tau_2 = 0$ (Theorem 4.2 (b.1) and (b.2)). This suggests that the oscillatory behavior of the glucose-insulin regulation requires the insulin removal rate to be large enough ($H_2(d_i, G_{in}) > 0$) and the delayed effect of hepatic glucose production to be long enough ($\tau_2 > \tau_{20}$).

[B1] and [B2] analytically confirm the numerical observation of [ST3]. It demonstrates that the effort of dividing insulin into two compartments in the model can be and shall be naturally and explicitly replaced by the delay parameter τ_1.

[B3] According to Theorem 4.2 (b.1) for $\tau_1 = 0$, and Theorem 5.2 for $\tau_2 > 0$, the insulin degradation rate d_i has to be "large" enough for sustained oscillatory regulation of the glucose-insulin metabolic system. Here the meaning of "large" refers to the numerical simulation demonstrated in Figure 5.1 (right) that $H_1(d_i, G_{in}) > 0$. This confirms the observation [BG1] in [2].

[B4] When $\tau_1\tau_2 > 0$, Theorem 5.6, Theorem 5.10, and simulations in Figure 5.1 (right) reveal the intrinsic relationship among d_i, G_{in}, τ_1, and τ_2 to secure the oscillatory behavior of the metabolic system. That is, for a subject, the oscillatory regulation occurs if one's insulin degradation rate and the glucose infusion rate satisfy $H_2(d_i, G_{in}) > 0$, and the time delays in the system are long enough ($\tau_1 > \tau_{10}$ and $\tau_1 + \tau_2 > \tau_{20}$). The numerical observations in Figure 6.3 (left) indicate that if the insulin degradation rate is sufficiently small ($H_1(d_i, G_{in}) < 0$), the oscillations cannot be sustained. Small d_i causes the insulin concentration level to remain high in plasma, which prohibits the glucose concentration level to rise. In such cases, the oscillatory regulation does not occur. This provides more insightful information than the general statements in [BG1] and [A7]. On the other hand, Figure 6.3 (right) indicates that both the glucose infusion rate and the insulin clearance rate are sensitive to the delays τ_1 and τ_2. Both rates are required to be in moderate ranges for sustained oscillations when the delayed effects are shorter.

[B5] Figures 6.1 (right), 6.2 (right), and 6.3 (left) show that when the glucose infusion rate is high, the oscillation of insulin secretion is unlikely to persist. This is possibly due to the fact that the β-cells cannot produce and secrete enough insulin to uptake the large amount of glucose infused into plasma. Thus the glucose concentration remains at a high level. The result is that the ultradian oscillations of insulin secretion and the oscillatory regulation of the glucose-insulin metabolic system cannot be sustained. This may help to explain the observed steady state behavior in models of the intravenous glucose tolerance test (IVGTT) where initial glucose infusion values are high [3], [9], [16], [19].

[B6] In the IVGTT, the initial glucose infusion is large. Compared to such large exogenous glucose infusion, the hepatic glucose production is negligible. For this reason, IVGTT models are justified not to include the hepatic glucose production term explicitly (set $f_5 \equiv 0$ thus $\tau_2 = 0$) [3], [9], [16], and [19]. The main goal of these models is to accurately monitor the dynamical behavior of the glucose level, which must return to its basal level after the biphasic insulin secretions have been triggered. Thus the insulin sensitivity can be

tested. The simulations (Figures 1 and 2) in [16] reveal that the delay τ_1 has to be extremely large (> 400 minutes) to produce any sustainable oscillations. Such a huge delay τ_1 clearly falls out of the normal physiological range.

Since we normally eat three meals per day, it is more plausible to consider periodic exogenous glucose infusion. That is, the constant glucose infusion rate G_{in} in model (2.1) shall be replaced by a periodic function $G_{in}(t)$ with a positive period ω between 180 and 300 minutes. Our simulation results reveal that there exists a harmonic solution in such a system. For more details, interested readers can refer to [24].

Appendix A. Proof of Proposition 3.2. For the first part of (i), let

$$H(x) = G_{in} - f_2(x) - f_3(x)f_4(d_i^{-1}f_1(x)) + f_5(d_i^{-1}f_1(x)) = 0, \quad x \geq 0. \tag{A.1}$$

We shall show that (A.1) has a unique root in $[0, \infty)$. Observing that $f_1'(x) > 0$, $f_2'(x) > 0$, $f_4'(x) > 0$, $f_3'(x) > 0$, and $f_5'(x) < 0$, we have $H'(x) < 0$. Notice that $H(0) = G_{in} - f_2(0) - f_3(0)f_4(d_i^{-1}f_1(0)) + f_5(d_i^{-1}f_1(0)) = G_{in} + f_5(d_i^{-1}f_1(0)) > 0$, and

$$\begin{aligned}
\lim_{x\to\infty} H(x) &= G_{in} - \lim_{x\to\infty} f_2(x) - \lim_{x\to\infty} f_3(x)f_4(d_i^{-1} \lim_{x\to\infty} f_1(x)) + f_5(d_i^{-1} \lim_{x\to\infty} f_1(x)) \\
&= G_{in} - M_2 - f_4(d_i^{-1}M_1) \lim_{x\to\infty} f_3(x) + f_5(d_i^{-1}M_1) \\
&< G_{in} - M_2 - m_4 \lim_{x\to\infty} f_3(x) + M_5 < 0.
\end{aligned}$$

In addition, $f_1(x)$ is strictly monotone increasing, so the proof is complete. It is obvious that G^* is the root of (A.1) and $I^* = d_i^{-1}f_1(G^*)$.

For the second part of (i), observing that $|f_i'(x)|$, $i = 1, 2, 3, 4, 5$, are bounded, $f_i(x), i = 2, 3, 4$, and $f_j(x_t), j = 1, 5$, are *Lipschitzian* and completely continuous in $x \geq 0$ and $x_t \in \mathbf{C}([-\max\{\tau_1, \tau_2\}, 0])$, respectively. Then by Theorems 2.1, 2.2, and 2.4 on pp. 19 and 20 in [15], the solution of (2.1) with given initial condition exists and is unique for all $t \geq 0$. If there exists a $t_0 > 0$ such that $G(t_0) = 0$ and $G(t) > 0$ for $0 < t < t_0$, then $G'(t_0) \leq 0$. So

$$\begin{aligned}
0 \geq G'(t_0) &= G_{in} - f_2(G(t_0)) - f_3(G(t_0))f_4(I(t_0)) + f_5(I(t_0 - \tau_2)) \\
&= G_{in} - f_2(0) - f_3(0)f_4(I(t_0)) + f_5(I(t_0 - \tau_2)) \\
&= G_{in} + f_5(I(t - \tau_2)) > 0.
\end{aligned}$$

This contradiction implies that $G(t) > 0$ for all $t > 0$. If $\exists t_0' > 0$ such that $I(t_0') = 0$ and $I(t) > 0$ for all $0 < t < t_0'$, then $I(t_0^{'}) < 0$. Therefore, $0 > I(t_0') = f_1(G(t_0^{'}) - d_i I(t_0' - \tau_1) \geq f_1(G(t_0')) > 0$. This implies that $I(t) > 0$ for all $t > 0$.

Now we show that any given solution $(G(t), I(t))$ of model (2.1) is bounded for $t > 0$. In fact, if $\limsup_{t\to\infty} G(t) = \infty$, there exists a sequence $\{t_n\}_{n=1}^{\infty} \uparrow \infty$ such that $\lim_{n\to\infty} G(t_n) = \infty$ and $G'(t_n) \geq 0$. Thus $0 < G'(t_n) = G_{in} - f_2(G(t_n)) - f_3(G(t_n))f_4(I(t_n)) + f_5(I(t_n - \tau_2)) \leq G_{in} - f_2(G(t_n)) - f_3(G(t_n))m_4 + M_5$, and therefore

$$\begin{aligned}
0 \leq \lim_{n\to\infty} G'(t_n) &\leq G_{in} - \lim_{n\to\infty} f_2(G(t_n)) - m_4 \lim_{n\to\infty} f_3(G(t_n)) + M_5 \\
&\leq G_{in} - M_2 - m_4 \lim_{x\to\infty} f_3(x) + M_5 < 0.
\end{aligned}$$

This contradiction shows that there is an $M_G > 0$ such that $G(t) < M_G$ for all $t > 0$. From the second equation in (2.1), since $|f_1(x)| \leq M_1$, for all $\epsilon > 0$, $I'(t) \leq f_1(M_G +$

$\epsilon) - d_i I(t)$ for sufficiently large $t > 0$. Then we have $\limsup_{t\to\infty} I(t) \le d_i^{-1} f_1(M_G+\epsilon)$. Since $\epsilon > 0$ is arbitrary, $\limsup_{t\to\infty} I(t) \le d_i^{-1} f_1(M_G) := M_I$.

If (ii) is not true, assume $\limsup_{t\to\infty} G(t) = M_G < \infty$. Then $\exists\{t_n\}_{n=1}^{\infty} \uparrow \infty$ such that $G'(t_n) = 0, n = 1, 2, 3, \ldots$, and $\lim_{n\to\infty} G(t_n) = M_G$ according to Lemma 3.1. Thus $G'(t_n) = G_{in} - f_2(G(t_n)) - f_3(G(t_n))f_4(I(t_n)) + f_5(I(t_n - \tau_2)) \ge G_{in} - f_2(G(t_n)) - f_3(G(t_n))m_4$. Let $n \to \infty$; then $0 \ge G_{in} - f_2(M_G) - f_3(M_G)m_4$, that is, $f_3(M_G) \ge (G_{in} - f_2(M_G))/m_4$. On the other hand, $f_3(M_G) \le \lim_{x\to\infty} f_3(x) < (G_{in} - M_2)/m_4 \le (G_{in} - f_2(M_G))/m_4$.

Appendix B. Proof of Lemma 3.3. First we show that (3.2) holds. Due to Lemma 3.1 and part (i) of Proposition 3.2, there exists a sequence $\{t_k\}_{k=1}^{\infty} \uparrow \infty$, such that $I'(t_k) = 0, \lim_{k\to\infty} I(t_k) = \overline{I}$. Thus, $0 = I'(t_k) = f_1(G(t_k - \tau_1)) - d_i I(t_k)$ for all $k = 1, 2, 3, \ldots$. Therefore, $f_1(\overline{G}) - d_i I(t_k) \ge f_1(\overline{G}(t_k - \tau_1)) - d_i I(t_k)$ for $k = 1, 2, 3, \ldots$. Thus, $f_1(\overline{G}) - d_i \overline{I} \ge 0$. On the other hand, there exists a sequence $\{s_k\}_{k=1}^{\infty} \uparrow \infty$ such that $\lim_{k\to\infty} I(s_k) = \underline{I}$ and $I'(s_k) = 0$ for all $k > 0$. Hence, $f_1(\underline{G}) - d_i I(s_k) \le f_1(\underline{G}(s_k - \tau_1)) - d_i I(s_k)$ for $k = 1, 2, 3, \ldots$. Thus, $f_1(\underline{G}) - d_i \underline{I} \le 0$.

Now we show that (3.3) holds. Again, due to Proposition 3.2 and Lemma 3.1, there exists a sequence $\{t_k'\}_{k=1}^{\infty} \uparrow \infty$ as $k \to \infty$ such that $\lim_{k\to\infty} G(t_k') = \overline{G}$ and $0 = G'(t_k') = G_{in} - f_2(G(t_k')) - f_3(G(t_k'))f_4(I(t_k')) + f_5(I(t_k' - \tau_2))$, $k = 1, 2, 3, \ldots$. Notice that $f_5 \downarrow 0$ and f_4 is monotone increasing and bounded from above by M_4; thus $0 = G_{in} - f_2(G(t_k')) - f_3(G(t_k'))f_4(I(t_k')) + f_5(I(t_k' - \tau_2)) \le G_{in} - f_2(G(t_k')) - f_3(G(t_k'))f_4(\underline{I}) + f_5(\underline{I}), k = 1, 2, 3, \ldots$, and thus $G_{in} - f_2(\overline{G}) - f_3(\overline{G})f_4(\underline{I}) + f_5(\underline{I}) \ge 0$.

Similarly we can show that (3.4) is true. According to part (i) of Proposition 3.2 and Lemma 3.1, there exists a sequence $\{s_k'\}_{k=1}^{\infty} \uparrow \infty$ as $k \to \infty$ such that $\lim_{k\to\infty} G(s_k') = \overline{G}$ and $0 = G'(s_k') = G_{in} - f_2(G(s_k')) - f_3(G(s_k'))f_4(I(s_k')) + f_5(I(s_k' - \tau_2)), k = 1, 2, 3, \ldots$. Notice that $f_5 \downarrow 0$ and f_4 is monotone increasing and bounded from above by M_4; thus $0 = G_{in} - f_2(G(s_k')) - f_3(G(s_k'))f_4(I(s_k')) + f_5(I(s_k' - \tau_2)) \ge G_{in} - f_2(G(s_k')) - f_3(G(s_k'))f_4(\overline{I}) + f_5(\overline{I}), k = 1, 2, 3, \ldots$. This leads to $G_{in} - f_2(\underline{G}) - f_3(\underline{G})f_4(\overline{I}) + f_5(\overline{I}) \le 0$.

Appendix C. Proof of Lemma 5.4. Let $f(z) = b + ce^{-p_1 z} + de^{-p_2 z}$. We show that $f(z)$ has a zero with positive real part. Since $p_1/p_2 \in S_1$ in case (i) ($p_1/p_2 \in S_2$ in case (ii)), there exist integers $m, n \ge 1$ such that $\frac{p_1}{p_2} = \frac{2m}{2n-1}$ for case (i), or $\frac{p_1}{p_2} = \frac{2m-1}{2n}$ for case (ii). Let $z = x + q\pi i$, where $q = 2m/p_1 = (2n-1)/p_2$ for case (i) or $q = (2m-1)/p_1 = 2n/p_2$ for case (ii). Then

$$\begin{aligned} f(z) &= b + ce^{-p_1 x}e^{-p_1 q\pi i} + de^{-p_2 x}e^{-p_2 q\pi i} \\ &= b + ce^{-p_1 x}\cos 2m\pi + de^{-p_2 x}\cos(2n-1)\pi \\ &\qquad (= b + ce^{-p_1 x}\cos(2m-1)\pi + de^{-p_2 x}\cos 2n\pi \text{ for case (ii)}) \\ &= b + ce^{-p_1 x} - de^{-p_2 x} \qquad (= b - ce^{-p_1 x} + de^{-p_2 x} \text{ for case (ii)}) \\ &:= H(x). \end{aligned}$$

Notice that $H(0) = b + c - d < 0$ ($H(0) = b - c + d < 0$ for case (ii)) and $\lim_{x\to\infty} H(x) = b > 0$; therefore $H(x)$ has at least one zero $x_0 \in (0, \infty)$. So $f(z)$ has at least one zero $z_0 = x_0 + q\pi i$ with $x_0 > 0$.

We perturb $f(z)$ by $g_\epsilon(z)$ given by

$$g_\epsilon(z) = \epsilon^k z^k + \sum_{j=1}^{k-1} a_j \epsilon^j z^j, \quad \epsilon > 0, \tag{C.1}$$

with small $\epsilon > 0$ and show that $f(z) + g_\epsilon(z)$ has the same number of zeros as $f(z)$ if ϵ is small. To this end, we first construct a simple loop γ homotopic to a point and then show $|f(z)| > |g_\epsilon(z)|$ on γ. Let $z = x$, $x \in (-\infty, \infty)$; then $|f(z)| = b + ce^{-p_1 x} + de^{-p_2 x} > b$. Let $z = x + 2q\pi i$, $x \in (-\infty, \infty)$; then

$$\begin{aligned}|f(z)| &= |b + ce^{-p_1 x} e^{2qp_1\pi i} + de^{-p_2 x} e^{2qp_2\pi i}| \\ &= |b + ce^{-p_1 x} \cos 4m\pi + de^{-p_2 x} \cos 2(2n-1)\pi| \\ &\qquad (= |b + ce^{-p_1 x} \cos 2(2m-1)\pi + de^{-p_2 x} \cos 4n\pi| \text{ for case (ii)}) \\ &= b + ce^{-p_1 x} + de^{-p_2 x} > b.\end{aligned}$$

Let $z = Kx_0 + yi$, $y \in [0, 2q\pi]$, where $K > 1$ such that $b - ce^{-p_1 K x_0} - de^{-p_2 K x_0} > b/2$. Then

$$\begin{aligned}|f(z)| &= |b + ce^{-p_1 K x_0} e^{-p_1 yi} + de^{-p_2 K x_0} e^{-p_2 yi}| \\ &\geq b - ce^{-p_1 K x_0} - de^{-p_2 K x_0} > b/2.\end{aligned}$$

Let $z = yi$, $y \in [0, 2q\pi]$; then

$$\begin{aligned}|f(z)| &= |b + ce^{-p_1 yi} + de^{-p_2 yi}| \geq \begin{cases} d - c - b & \text{for case (i)}, \\ c - d - b & \text{for case (ii)} \end{cases} \\ &:= \eta_0 > 0.\end{aligned}$$

Let $\eta_0' := \min\{\eta_0, b/2\}$. Denote

$$\begin{aligned}\gamma := \{z = x + yi \in \mathbf{C} : z = x \text{ or } z = x \pm 2q\pi i, \quad x \in [0, Kx_0], \\ \text{or } z = yi \text{ or } z = Kx_0 + yi, \quad y \in [0, 2q\pi].\end{aligned}$$

$$\gamma^\circ := \{z = x + yi \in \mathbf{C} : 0 < x < Kx_0, \ -2q\pi < y < 2q\pi\}.$$

Clearly, γ is a simple loop homotopic to the original, $z_0 = x_0 + q\pi i \in \gamma^\circ$ and $|f(z)| > \eta_0'$ on γ. Choose $r_0 > 0$ such that $\gamma \subset \mathcal{A} := \{z \in \mathbf{C} : |z| < r_0\}$. Denote $\partial\mathcal{A} := \{z \in \mathbf{C} : |z| = r_0\}$. Thus for all $z \in \partial\mathcal{A}$, $z = r_0 e^{\theta i}$, $\theta \in [0, 2\pi]$, we have

$$|g_\epsilon(z)| = |\epsilon^k z^k + \sum_{j=1}^{k-1} a_j \epsilon^j z^j| \leq \epsilon^k r_0^k + \sum_{j=1}^{k-1} |a_j| \epsilon^j r_0^j. \tag{C.2}$$

Obviously $\exists \epsilon_0 > 0$ such that for all $\epsilon, 0 < \epsilon < \epsilon_0$, $|g_\epsilon(z)| < \eta_0', z \in \partial\mathcal{A}$. For all $z \in \mathcal{A}, z = re^{\theta i}$; then $r < r_0$, and

$$|g_\epsilon(z)| = |\epsilon^k z^k + \sum_{j=1}^{k-1} a_j \epsilon^j z^j| \leq \epsilon^k r^k + \sum_{j=1}^{k-1} |a_j| \epsilon^j r^j < \epsilon^k r_0^k + \sum_{j=1}^{k-1} |a_j| \epsilon^j r_0^j.$$

Thus $|g_\epsilon(z)| < \eta_0'$ for all $z \in \gamma$. Therefore $|f(z)| > |g_\epsilon(z)|$ on γ. By Rouchè's theorem [7, pp. 125–126], $f(z)$ and $f(z) + g_\epsilon(z)$ have the same number of zeros in γ°. That is, $f(z) + g_\epsilon(z) = 0$ has at least one root $\hat{z}_\epsilon \in \gamma^\circ$.

Appendix D. Proof of Proposition 5.8. We need only show that the conjugate roots of (4.3) cross the imaginary axis from left to right. Assume $\tau_1 + \tau_2 < \frac{\pi}{2\hat{\omega}}$. From (4.3), we have

$$\begin{aligned}&\Big[2\lambda + (A + d_i) - DBe^{-\lambda\tau_1}\tau_1 - DCe^{-\lambda(\tau_1+\tau_2)}(\tau_1 + \tau_2)\Big]\frac{d\lambda}{d\tau_1} \\ &= \Big(DBe^{-\lambda\tau_1} + DCe^{-\lambda(\tau_1+\tau_2)}\Big)\lambda.\end{aligned}$$

If the root $\lambda(\bar{\tau}_1) = i\omega$ is not simple for some $\bar{\tau}_1 > 0$, then $\frac{d\lambda}{d\tau_1}\big|_{\tau_1=\bar{\tau}_1} = 0$. Thus,

$$-DBe^{-i\omega\bar{\tau}_1} - DCe^{-i\omega\bar{\tau}_1+\tau_2}i\omega = 0 \quad \text{and} \quad (B + C\cos\omega\tau_2) - i\sin\omega\tau_2 = 0.$$

This is impossible since $\tau_2 < \frac{\pi}{2\hat{\omega}}$. Therefore,

$$\Big(\frac{d\lambda}{d\tau_1}\Big)^{-1} = \frac{[2\lambda + (A + d_i)]e^{\lambda(\tau_1+\tau_2)} - \tau_2 DC}{(DBe^{\lambda\tau_2} + DC)\lambda} - \frac{\tau_1}{\lambda}.$$

Notice that at $\lambda = i\omega$,

$$\begin{aligned}
&\operatorname{sign}\Big\{\frac{d\mathrm{Re}(\lambda)}{d\tau_1}\Big\} = \operatorname{sign}\Big\{\mathrm{Re}\Big(\Big(\frac{d\lambda}{d\tau_1}\Big)^{-1}\Big)\Big\} \\
&= \operatorname{sign}\Big\{\mathrm{Re}\Big(\frac{(i(A + d_i) - 2\omega)(\cos\omega(\tau_1 + \tau_2) + i\sin\omega(\tau_1 + \tau_2)) - \tau_2 DCi}{-(DB\cos\omega\tau_2 + iDB\sin\omega\tau_2)\omega}\Big)\Big\} \\
&= \operatorname{sign}\Big\{(DB\cos\omega\tau_2 + DC)2\omega\cos\omega(\tau_1 + \tau_2) + DB\sin\omega\tau_2(2\omega\sin\omega(\tau_1 + \tau_2) + \tau_2 DC) \\
&\qquad + DB(A + d_i)\sin\omega\tau_1\Big\} = 1.
\end{aligned}$$

Appendix E. Proof of Proposition 5.9. Similar to the proof of Proposition 5.8 in Appendix D, assume $\tau_1 + \tau_2 < \frac{\pi}{2\hat{\omega}}$. From (4.3), we have

$$\Big[2\lambda + (A + d_i) - \tau_1 DBe^{\lambda\tau_1} - DC(\tau_1 + \tau_2)e^{-\lambda(\tau_1+\tau_2)}\Big]\frac{d\lambda}{d\tau_2} = DC\lambda e^{-\lambda(\tau_1+\tau_2)}.$$

If $\lambda(\bar{\tau}_2) = i\omega$ is a root of (4.3) for some $\bar{\tau}_2 > 0$ with $\omega > 0$, then it must be simple. Otherwise, $\frac{d\lambda}{d\tau_2}\big|_{\tau_2=\bar{\tau}_2} = 0$ and leads to a contradiction, $DC\omega\cos\tau_1 + \bar{\tau}_2 = 0$. We show that if a root of (4.3) crosses the imaginary axis while τ_2 increases, it must cross from left to right. Obviously,

$$\Big(\frac{d\lambda}{d\tau_2}\Big)^{-1} = \frac{(2\lambda + (A + d_i))e^{\lambda(\tau_1+\tau_2)} - \tau_1 DBe^{-\lambda\tau_2}}{DC\lambda} - \frac{\tau_1 + \tau_2}{\lambda}.$$

Thus, at $\lambda = i\omega$,

$$\Big(\frac{d\lambda}{d\tau_2}\Big)^{-1} = \frac{2\omega\cos\omega(\tau_1 + \tau_2) + (A + d_i)\sin\omega(\tau_1 + \tau_2) - \tau_1 DB\sin\omega\tau_2}{DC\omega}.$$

Then, if $\tau_1 DB < A + d_i$,

$$\begin{aligned}
&\operatorname{sign}\Big\{\frac{d\mathrm{Re}(\lambda)}{d\tau_2}\Big\} = \operatorname{sign}\Big\{\mathrm{Re}\Big(\Big(\frac{d\lambda}{d\tau_2}\Big)^{-1}\Big)\Big\} \\
&= \operatorname{sign}\Big\{2\omega\cos\omega(\tau_1 + \tau_2) + (A + d_i)\sin\omega(\tau_1 + \tau_2) - \tau_1 DB\sin\omega\tau_2\Big\} = 1.
\end{aligned}$$

Acknowledgments. We would like to thank the referees and the associate editor for their valuable suggestions that enabled us to improve the presentation of this paper. The first author thanks Prof. Carlos Castillo-Chavez for his encouragement for the analysis of the characteristic equation when $\tau_1\tau_2 > 0$.

REFERENCES

[1] B. Ahrén and G. J. Taborsky Jr., *B-cell function and insulin secretion*, in Ellenberg and Rifkin's Diabetes Mellitus, 6th ed., D. Porte, R. S. Sherwin, and A. Baron, eds., McGraw-Hill Professional, Chapter 4, pp. 43–65.

[2] D. L. Bennette and S. A. Gourley, *Asymptotic properties of a delay differential equation model for the interaction of glucose with the plasma and interstitial insulin*, Appl. Math. Comput., 151 (2004), pp. 189–207.
[3] R. N. Bergman, *Pathogenesis and prediction of diabetes mellitus: Lessons from integrative physiology*, Irving L. Schwartz Lecture, Mount Sinai J. Medicine, 60 (2002), pp. 280–290.
[4] R. N. Bergman, D. T. Finegood, and S. E. Kahn, *The evolution of beta-cell dysfunction and insulin resistance in type* 2 *diabetes*, Eur. J. Clin. Invest., 32 (Suppl. 3) (2002), pp. 35–45.
[5] V. W. Bolie, *Coefficients of normal blood glucose regulation*, J. Appl. Physiol., 16 (1961), pp. 783–788.
[6] A. D. Cherrington, D. Sindelar, D. Edgerton, K. Steiner, and O. P. McGuinness, *Physiological consequences of phasic insulin release in the normal animal*, Diabetes, 51 (Suppl. 1) (2002), pp. S103–S108.
[7] J. B. Conway, *Functions of One Complex Variable*, 2nd ed., Springer-Verlag, New York, 1973; corrected fourth printing, 1986.
[8] K. L. Cooke and P. van den Driessche, *On zeroes of some transcendental equations*, Funkcial. Ekvac., 29 (1986), pp. 77–90.
[9] A. De Gaetano and O. Arino, *Mathematical modeling of the intravenous glucose tolerance test*, J. Math. Biol., 40 (2000), pp. 136–168.
[10] M. Derouich and A. Boutayeb, *The effect of physical exercise on the dynamics of glucose and insulin*, J. Biomechanics, 35 (2002), pp. 911–917.
[11] W. C. Duckworth, R. G. Bennett, and F. G. Hamel, *Insulin degradation: Progress and potential*, Endocr. Rev., 19 (1998), pp. 698–624.
[12] W. M. Hirsch, H. Hanish, and J.-P. Gabriel, *Differential equation model of some parasitic infections: Methods for the study of asymptotic behavior*, Comm. Pure. Appl. Math., 38 (1985), pp. 733–753.
[13] J. Keener and J. Sneyd, *Mathematical Physiology*, Springer-Verlag, New York, 1998.
[14] W. C. Knowler, P. H. Bennett, R. F. Hamman, and M. Miller, *Diabetes incidence and prevalence in Pima Indians: A* 19*-fold greater incidence than in Rochester, Minnesota*, Am. J. Epidemiol., 108 (1978), pp. 497–505.
[15] Y. Kuang, *Delay Differential Equations with Applications in Population Dynamics*, Math. Sci. Eng. 191, Academic Press, Boston, 1993.
[16] J. Li, Y. Kuang, and B. Li, *Analysis of IVGTT glucose-insulin interaction models with time delay*, Discrete Contin. Dyn. Syst. Ser. B, 1 (2001), pp. 103–124.
[17] J. Li, Y. Kuang, and C. Mason, *Modeling the glucose-insulin regulatory system and ultradian insulin secretory oscillations with two time delays*, J. Theoret. Biol., 242 (2006), pp. 722–735.
[18] A. Makroglou, J. Li, and Y. Kuang, *Mathematical models and software tools for the glucose-insulin regulatory system and diabetes: An overview*, Appl. Numer. Math., 56 (2006), pp. 559–573.
[19] A. Mukhopadhyay, A. De Gaetano, and O. Arino, *Modeling the intra-venous glucose tolerance test: A global study for a single-distributed-delay model*, Discrete Contin. Dyn. Syst. Ser. B, 4 (2004), pp. 407–417.
[20] N. Pørksen, M. Hollingdal, C. Juhl, P. Butler, J. D. Veldhuis, and O. Schmitz, *Pulsatile insulin secretion: Detection, regulation, and role in diabetes*, Diabetes, 51 (2002), pp. S245–S254.
[21] R. Prager, P. Wallace, and J. M. Olefsky, *In vivo kinetics of insulin action on peripheral glucose disposal and hepatic glucose output in normal and obese subjects*, J. Clin. Invest., 78 (1986), pp. 472–481.
[22] L. F. Shampine and S. Thompson, *Solving DDEs in MATLAB*, Appl. Numer. Math., 37 (2001), pp. 441–458; http://www.radford.edu/∼thompson.
[23] C. Simon and G. Brandenberger, *Ultradian oscillations of insulin secretion in humans*, Diabetes, 51 (2002), pp. S258–S261.
[24] J. Sturis, *Possible Mechanisms Underlying Slow Oscillations of Human Insulin Secretion*, Ph.D. thesis, The Technical University of Denmark, Lyngby, Denmark, 1991.
[25] J. Sturis, K. S. Polonsky, E. Mosekilde, and E. Van Cauter, *Computer model for mechanisms underlying ultradian oscillations of insulin and glucose*, Am. J. Physiol., 260 (1991), pp. E801–E809.
[26] I. M. Tolic, E. Mosekilde, and J. Sturis, *Modeling the insulin-glucose feedback system: The significance of pulsatile insulin secretion*, J. Theoret. Biol., 207 (2000), pp. 361–375.
[27] B. Topp, K. Promislow, G. De Vries, R. M. Miura, and D. T. Finegood, *A model of β-cell mass, insulin, and glucose kinetics: Pathways to diabetes*, J. Theoret. Biol., 206 (2000), pp. 605–619.

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 777–791

NUMERICAL SIMULATION OF ACOUSTIC TIME REVERSAL MIRRORS*

CHOKRI BEN AMAR†, NABIL GMATI†, CHRISTOPHE HAZARD‡, AND KARIM RAMDANI§

Abstract. We study the time reversal phenomenon in a homogeneous and nondissipative medium containing sound-hard obstacles. We propose two mathematical models of time reversal mirrors in the frequency domain. The first one takes into account the interactions between the mirror and the obstacles. The second one provides an approximation of these interactions. We prove, in both cases, that the time reversal operator T is self-adjoint and compact. The DORT method (French acronym for decomposition of the time reversal operator) is explored numerically. In particular, we show that selective focusing, which is known to occur for small and distant enough scatterers, holds when the wavelength and the size of these scatterers are of the same order of magnitude (medium frequency situation). Moreover, we present the behavior of the eigenvalues of T according to the frequency, and we show their oscillations due to the interactions between the mirror and the obstacles and between the obstacles themselves.

Key words. time reversal, frequency domain, acoustic scattering, selective focusing

AMS subject classifications. 31B10, 35P25, 74J20

DOI. 10.1137/060654542

1. Introduction. During the last decade, time reversal techniques have been extensively studied, in particular for detection, localization, and identification of scatterers in propagative media. In the present paper, we are concerned with one of these techniques, usually referred to as the DORT method (French acronym for decomposition of the time reversal operator). This method was first developed by Prada and Fink [17] in the context of ultrasonics (see [18] for an overview). It consists in determining the invariants of a time reversal process which can be described as follows. A time reversal mirror (TRM), composed of an array of transducers, first emits an incident wave corresponding to a given distribution of signals sent to the transducers. This wave is then scattered by the presence of obstacles in the propagative medium. In a second step, the TRM measures the scattered field and time-reverses the measure, which furnishes a new distribution of signals used to reemit a new incident wave. In short, one cycle of the process corresponds to the succession of steps: emission, scattering, measure, time reversal. The so-called time reversal operator T is obtained by iterating this cycle twice. The DORT method deals with the eigenvalues of T and the associated eigenvectors for a fixed frequency, that is, when time-harmonic waves are considered. In this case, time reversal simply amounts to a phase conjugation. It was shown [17, 19] and confirmed by experiments that for ideally resolved or pointlike and distant enough scatterers with different reflectivities, each eigenvector corresponding to a nonzero eigenvalue of T provides the signals to be sent to the transducers in

*Received by the editors March 17, 2006; accepted for publication (in revised form) November 28, 2006; published electronically March 15, 2007.
http://www.siam.org/journals/siap/67-3/65454.html
†LAMSIN, ENIT, Campus Universitaire, B.P. 37, 1002 Tunis Belvédère, Tunisie (chokri.benamar@lamsin.rnu.tn, nabil.gmati@ipein.rnu.tn).
‡POEMS (CNRS / ENSTA / INRIA), 32, Boulevard Victor, 75739 Paris Cedex 15, France (christophe.hazard@ensta.fr).
§INRIA (Projet CORIDA) and Institut Elie Cartan de Nancy, Université de Nancy I, Vandoeuvre-les-Nancy 54506, France (karim.ramdani@loria.fr).

order to focus on one scatterer. A mathematical justification of these selective focusing properties is given in [8] for a far field approach, i.e., for an ideal TRM which reverses the asymptotic behavior at large distance of the wave scattered by the obstacles (in this case, the time reversal operator is related with the far field operator [13] well known in scattering theory). Other applications of the DORT method, which concern this question of focusing on a selected target, have been developed: acoustic waveguides [10, 15], electromagnetic scattering [22, 14], or propagation in random media [4].

The focusing properties of the eigenvectors of the time reversal operator are known to occur for small enough scatterers, i.e., when the diameters of the scatterers are small compared to the wavelength. Such a situation corresponds to a low frequency case. The object of the present paper is to explore the medium frequency case by a numerical approach, i.e., when the diameters and the wavelength have the same order of magnitude. The model considered here differs from commonly used models in the fact that the TRM is intrusive: instead of an array of pointlike transducers, the TRM consists of a volumic and nonpenetrable object which perturbs the acoustic field. For the sake of simplicity, we consider the usual simplified model of linear electroacoustic transducers (see, e.g., [16]): the inner behavior of the TRM is modeled by a Robin condition on its boundary.

The paper is organized as follows. In section 2, we present a mathematical model of a nonpenetrable intrusive TRM, which is closely related to the active sonar problem dealt with, for instance, in [20]. In this first model, the interactions between the scatterers and the TRM are taken into account, so that we can deal with the case where they are close to each other. Instead of the symmetric matrix obtained for a finite number of pointlike transducers, the time reversal operator then appears, like in the far field approach [8], as an operator acting in an L^2 space representing the finite energy space of possible excitations. The basic properties of this operator, namely self-adjointness and compactness, are proved in section 3. They essentially tell us that its spectrum is that of a symmetric matrix completed by an infinite number of nonsignificant eigenvalues. In section 4, we propose a nonpenetrable intrusive model of a TRM in which the interactions between the obstacles and the TRM are approximated. We briefly show how to adapt the proofs of section 3. Finally, we present some numerical results in section 5. We show that the expected selective focusing properties hold in the medium frequency case. Moreover, we point out the modulations of the eigenvalues of T with respect to the frequency. These oscillations are due to the interactions between the scatterers and the TRM, and between the scatterers themselves.

The main result of this paper, namely the properties of the time reversal operator (Theorem 2.1), holds in many other situations which can be dealt with by similar integral techniques. For instance, here we consider sound-hard obstacles, but we could have chosen a Dirichlet or Robin boundary condition on $\partial\mathcal{O}$ instead of the Neumann condition. Penetrable scatterers, i.e., inhomogeneities of the medium, can also be considered.

2. A model of nonpenetrable intrusive mirror. We consider a homogeneous medium filling the space $\mathbb{R}^n$ ($n = 2$ or 3) and containing a nonpenetrable mirror M and some nonpenetrable obstacles $\mathcal{O}$.

We study the case of an impedance condition on the boundary ∂M of the mirror and a Neumann condition on the boundary $\partial\mathcal{O}$ of the obstacles (sound-hard obstacles). Let $\Omega_M = \mathbb{R}^n\backslash\overline{M}$, $\Omega_{\mathcal{O}} = \mathbb{R}^n\backslash\overline{\mathcal{O}}$, and $\Omega_{M,\mathcal{O}} = \mathbb{R}^n\backslash(\overline{M} \cup \overline{\mathcal{O}})$. We suppose that the

boundary of the mirror is excited by a signal g (proportional to the current which flows through each transducer). So, in the presence of the obstacles, we observe the total field φ_T satisfying the problem

$$\begin{cases} \Delta\varphi_T + k^2\varphi_T = 0 \quad \text{in } \Omega_{M,\mathcal{O}}, \\ \dfrac{\partial\varphi_T}{\partial n} + \mu\varphi_T = g \quad \text{on } \partial M, \\ \dfrac{\partial\varphi_T}{\partial n} = 0 \quad \text{on } \partial\mathcal{O}, \\ \text{RC} \quad \text{at } \infty, \end{cases} \tag{2.1}$$

where n denotes the unit normal vector directed into the interior of the domain $\Omega_{M,\mathcal{O}}$. The wave number k is defined by $k = \omega/c$, where ω is the frequency and c is the speed of sound in the homogeneous medium, μ is a real parameter which represents the inverse of the open-circuit acoustic impedance of the TRM [16], and RC is the outgoing Sommerfeld "radiation condition" which, for φ_T, is

$$\lim_{r\to+\infty} r^{\frac{n-1}{2}} \left(\frac{\partial\varphi_T}{\partial r}(x) - ik\varphi_T(x) \right) = 0, \quad r = |x|, \tag{2.2}$$

where $\partial\varphi_T/\partial r$ denotes the radial derivative of φ_T.

In the absence of obstacles, we should observe an incident field φ_I solution to

$$\begin{cases} \Delta\varphi_I + k^2\varphi_I = 0 \quad \text{in } \Omega_M, \\ \dfrac{\partial\varphi_I}{\partial n} + \mu\varphi_I = g \quad \text{on } \partial M, \\ \text{RC} \quad \text{at } \infty. \end{cases} \tag{2.3}$$

The perturbation due to the presence of the obstacles is the diffracted field $\varphi_D = \varphi_T - \varphi_I$ satisfying the problem

$$\begin{cases} \Delta\varphi_D + k^2\varphi_D = 0 \quad \text{in } \Omega_{M,\mathcal{O}}, \\ \dfrac{\partial\varphi_D}{\partial n} + \mu\varphi_D = 0 \quad \text{on } \partial M, \\ \dfrac{\partial\varphi_D}{\partial n} = h \quad \text{on } \partial\mathcal{O}, \\ \text{RC} \quad \text{at } \infty, \end{cases} \tag{2.4}$$

where $h = -\partial\varphi_I/\partial n$.

We suppose that the signal measured by the mirror is equal to $\varphi_{D_{/\partial M}}$, the value of the diffracted field on ∂M. The measured signal is then conjugated and used to generate the incident and the total fields in the next iteration.

Time reversal operator. Let R denote the operator describing the response of the medium, that is, the three successive steps: emission, diffraction, measure. It is defined by

$$Rg = \varphi_{D_{/\partial M}}.$$

The time reversal operator is obtained by iterating the time reversal process (emission, diffraction, measure, conjugation) twice. Therefore, T is given by

$$Tg = \overline{R\overline{Rg}}, \quad \text{that is,} \quad T = \overline{R}R,$$

where the operator $\overline{R}$ is defined by

$$\overline{R}g = \overline{R\,\overline{g}}.$$

THEOREM 2.1. *T is a self-adjoint positive and compact operator in $L^2(\partial M)$.*

These properties are proved below by an integral approach based on the use of several Green's functions.

3. Proof of Theorem 2.1. This section is devoted to the proof of the following result, from which Theorem 2.1 derives.

PROPOSITION 3.1. *For every $g \in L^2(\partial M)$, the response $Rg \in L^2(\partial M)$ of the medium is given by*

$$(Rg)(x) = \int_{\partial M} G_R(x,y)g(y)d\sigma(y), \tag{3.1}$$

where $G_R \in L^2(\partial M \times \partial M)$ is symmetric, i.e., $G_R(x,y) = G_R(y,x)$.

This proposition shows that R is a Hilbert–Schmidt operator in $L^2(\partial M)$ such that $R^* = \overline{R}$ since

$$(R^*g)(x) = \int_{\partial M} \overline{G_R(y,x)}g(y)d\sigma(y) \quad \text{and} \quad (\overline{R}g)(x) = \int_{\partial M} \overline{G_R(x,y)}g(y)d\sigma(y).$$

Hence $T = R^*R$ is self-adjoint positive and compact in $L^2(\partial M)$. It is actually a Hilbert–Schmidt operator in $L^2(\partial M)$ whose kernel $G \in L^2(\partial M \times \partial M)$ is given by

$$G(x,y) = \int_{\partial M} \overline{G_R(z,x)}G_R(z,y)d\sigma(z).$$

The spectral properties of T follow. On one hand, the eigenvalues of T form a decreasing sequence of positive numbers $(\lambda_n)_{n\in\mathbb{N}^*}$ such that $\sum_{n\in\mathbb{N}^*} \lambda_n^2$ is finite. On the other hand, one can choose an orthonormal basis of $L^2(\partial M)$ composed of eigenvectors of T, and T becomes diagonal in this basis.

Integral representations. To prove Proposition 3.1, first recall that problems (2.1), (2.3), and (2.4) are well-posed [6] in a proper functional framework which is made precise later. Consider then the operators

$$\begin{aligned} S_T &: g \mapsto \varphi_T \quad \text{solution to (2.1)},\\ S_I &: g \mapsto \varphi_I \quad \text{solution to (2.3)},\\ S_D &: h \mapsto \varphi_D \quad \text{solution to (2.4)}, \end{aligned}$$

as well as the Green's functions G_T, G_I, and G_D, which are, respectively, outgoing solutions (in the sense that they satisfy the outgoing radiation condition (2.2)) to

$$\begin{cases} \Delta G_T(x,.) + k^2 G_T(x,.) = \delta_x & \text{in } \Omega_{M,\mathcal{O}},\\ \Theta_M G_T(x,.) = 0 & \text{on } \partial M,\\ \Theta_{\mathcal{O}} G_T(x,.) = 0 & \text{on } \partial\mathcal{O}, \end{cases} \tag{3.2}$$

$$\begin{cases} \Delta G_I(x,.) + k^2 G_I(x,.) = \delta_x & \text{in } \Omega_M, \\ \Theta_M G_I(x,.) = 0 & \text{on } \partial M, \end{cases} \tag{3.3}$$

$$\begin{cases} \Delta G_D(x,.) + k^2 G_D(x,.) = 0 & \text{in } \Omega_{M,\mathcal{O}}, \\ \Theta_M G_D(x,.) = 0 & \text{on } \partial M, \\ \Theta_{\mathcal{O}} G_D(x,.) = -\Theta_{\mathcal{O}} G_I(x,.) & \text{on } \partial\mathcal{O}, \end{cases} \tag{3.4}$$

where δ_x stands for the Dirac measure at point x, $\Theta_M = (\partial/\partial n + \mu)_{/\partial M}$, and $\Theta_{\mathcal{O}} = (\partial/\partial n)_{/\partial\mathcal{O}}$. By construction, we have $G_T = G_I + G_D$.

These functions can be expressed by means of the usual Green's function G_0 of the Helmholtz operator in the free space, i.e., the outgoing solution in $\mathbb{R}^n$ to $\Delta G_0(x,.) + k^2 G_0(x,.) = \delta_x$, which is given by

$$G_0(x,y) = \begin{cases} -\dfrac{e^{ik|x-y|}}{4\pi|x-y|} & \text{if } n = 3, \\ \dfrac{1}{4i} H_0^{(1)}(k|x-y|) & \text{if } n = 2. \end{cases}$$

Indeed, we have

$$\begin{aligned} &G_I(x,.) = G_0(x,.) + \widetilde{G}_I(x,.), \quad \text{where } \widetilde{G}_I(x,.) = -S_I \Theta_M G_0(x,.), \\ &G_D(x,.) = -S_D \Theta_{\mathcal{O}} G_I(x,.). \end{aligned} \tag{3.5}$$

LEMMA 3.2. *Let Ω_i stand for Ω_M if $i = I$, and for $\Omega_{M,\mathcal{O}}$ if $i = T$ or D. Then $\varphi_T = S_T g$, $\varphi_I = S_I g$, and $\varphi_D = -S_D \Theta_{\mathcal{O}} \varphi_I$ are given by*

$$\varphi_i(x) = \int_{\partial M} G_i(x,y) g(y) d\sigma(y) \quad \forall x \in \Omega_i,\ i \in \{T, I, D\}, \tag{3.6}$$

where the kernels G_i are symmetric: $G_i(x,y) = G_i(y,x)$.

Proof. Formulas (3.6) are classical. For the sake of clarity, we recall briefly how to derive them from the usual integral representation [5]

$$\varphi_i(x) = \int_{\partial\Omega_i} \left\{ G_0(x,y) \frac{\partial \varphi_i}{\partial n}(y) - \frac{\partial G_0}{\partial n_y}(x,y) \varphi_i(y) \right\} d\sigma(y) \quad \forall x \in \Omega_i. \tag{3.7}$$

We use the fact that if two functions φ and ψ satisfy the Helmholtz equation either inside a bounded domain Λ, or outside Λ together with the radiation condition (2.2), then we have the reciprocity relation [5]

$$\int_{\partial\Lambda} \left\{ \psi \frac{\partial \varphi}{\partial n} - \frac{\partial \psi}{\partial n} \varphi \right\} d\sigma = 0, \tag{3.8}$$

where the normal derivative can obviously be replaced by $(\partial/\partial n + \mu)$.

For φ_I, we replace G_0 in (3.7) by $G_I - \widetilde{G}_I$, which yields two similar integral terms on ∂M. Thanks to (3.8), the term involving $\widetilde{G}_I$ vanishes. The other one reduces to the single layer potential (3.6) by virtue of the boundary conditions satisfied by φ_I and $G_I(x,.)$ (see (2.3) and (3.3)).

For φ_T, the same idea applies. The integral terms are now set on $\partial M \cup \partial\mathcal{O}$. The term which involves $\widetilde{G}_T = G_T - G_0$ again vanishes by (3.8). Split the other

one, which involves G_T, into two integrals, respectively, on ∂M and $\partial\mathcal{O}$. Thanks to the boundary conditions in (2.1) and (3.2), the former simplifies as above to (3.6), whereas the latter vanishes.

Finally, subtracting the previous representations yields (3.6) for $\varphi_D = \varphi_T - \varphi_I$.

The symmetry of G_I is easily deduced from that of G_0 by proving that the perturbation term $\widetilde{G}_I$ is also symmetric. The integral representation (3.7) of $\widetilde{G}_I(x,.)$ reads

$$\widetilde{G}_I(x,y) = \int_{\partial M} \left\{ \Theta_M G_0(y,z) \widetilde{G}_I(x,z) - G_0(y,z) \Theta_M \widetilde{G}_I(x,z) \right\} d\sigma(z),$$

where the operator Θ_M is understood with respect to z. The boundary conditions satisfied by $\widetilde{G}_I(x,.)$ and $\widetilde{G}_I(y,.)$ then yield

$$\widetilde{G}_I(x,y) = -\int_{\partial M} \Theta_M \widetilde{G}_I(y,z) \widetilde{G}_I(x,z) d\sigma(z) + \int_{\partial M} G_0(y,z) \Theta_M G_0(x,z) d\sigma(z).$$

Thanks to the reciprocity relation (3.8) applied in Ω_M for the first integral, and in M for the second one, we see that both integrals are symmetric for $(x,y) \in \Omega_M \times \Omega_M$, and hence so is $\widetilde{G}_I$.

The symmetry of G_D is proved similarly, and that of G_T follows. ∎

Functional details. It is now clear that (3.1) follows from the integral representation (3.6) of φ_D simply by taking its restriction on ∂M:

$$G_R(x,y) = G_D(x,y) \text{ for } (x,y) \in \partial M \times \partial M.$$

Hence Proposition 3.1 will be proved if we are able to justify that this double restriction actually yields a function of $L^2(\partial M \times \partial M)$. We thus have to make precise the function spaces in which the kernels G_i are defined: the appropriate tool to do so is the notion of tensor product of Hilbert spaces [1].

All the domains considered are assumed to have Lipschitz boundaries (for instance, ∂M and $\partial\mathcal{O}$ may be piecewise smooth). For a bounded domain $\Lambda \subset \mathbb{R}^n$, we denote

$$\mathbf{H}(\Lambda) = \left\{ \varphi \in H^1(\Lambda);\ \Delta\varphi \in L^2(\Lambda) \right\}.$$

Recall that, on one hand, the trace operator $\gamma_{\partial\Lambda}\varphi = \varphi_{/\partial\Lambda}$ is continuous from $\mathbf{H}(\Lambda)$ to $H^{1/2}(\partial\Lambda)$, and, on the other hand, the normal derivative $(\partial\varphi/\partial n)_{/\partial\Lambda}$ is continuous from $\mathbf{H}(\Lambda)$ to $H^{-1/2}(\partial\Lambda)$. Moreover, for all bounded sets $\Lambda_M \subset \Omega_M$ and $\Lambda_{\mathcal{O}} \subset \Omega_{M,\mathcal{O}}$, the operators S_I and S_D are continuous from $H^{-1/2}(\partial M)$ to $\mathbf{H}(\Lambda_M)$ and from $H^{-1/2}(\partial\mathcal{O})$ to $\mathbf{H}(\Lambda_{\mathcal{O}})$ (see [6]).

LEMMA 3.3. *Let $\Lambda_M \subset \Omega_M$ and $\Lambda_{\mathcal{O}} \subset \Omega_{M,\mathcal{O}}$ be two bounded sets such that $\partial M \subset \partial\Lambda_M$, $\partial\mathcal{O} \subset \partial\Lambda_{\mathcal{O}}$, and $\overline{\Lambda_M} \cap \overline{\Lambda_{\mathcal{O}}} = \emptyset$. Then*

$$G_I \in \mathbf{H}(\Lambda_M) \widehat{\otimes} \mathbf{H}(\Lambda_{\mathcal{O}}) \quad \textit{and} \quad G_D \in \mathbf{H}(\Lambda_M) \widehat{\otimes} \mathbf{H}(\Lambda_M).$$

Proof. Formulas (3.5), which involve operators acting on the second variable y, can be rewritten in terms of tensor products of operators as

$$G_I = G_0 - (Id \otimes S_I \Theta_M) G_0,$$
$$G_D = -(Id \otimes S_D \Theta_{\mathcal{O}})\, G_I.$$

Let us first deal with G_I. Thanks to its symmetry, the announced property amounts to showing that $G_I \in \mathbf{H}(\Lambda_{\mathcal{O}})\widehat{\otimes}\mathbf{H}(\Lambda_M)$. This clearly holds for G_0 since it is infinitely differentiable outside the diagonal $x = y$. Moreover, the above-mentioned properties of S_I and traces show that $S_I\Theta_M$ is continuous from $\mathbf{H}(\Lambda_M)$ to $\mathbf{H}(\Lambda_M)$. As a consequence [1], $Id \otimes S_I\Theta_M$ is continuous from $\mathbf{H}(\Lambda_{\mathcal{O}})\widehat{\otimes}\mathbf{H}(\Lambda_M)$ to itself. The conclusion follows.

For G_D, we use the previous result and the fact that $Id \otimes S_D\Theta_{\mathcal{O}}$ is continuous from $\mathbf{H}(\Lambda_M)\widehat{\otimes}\mathbf{H}(\Lambda_{\mathcal{O}})$ to $\mathbf{H}(\Lambda_M)\widehat{\otimes}\mathbf{H}(\Lambda_M)$. □

We finally have to notice that since $\gamma_{\partial M}$ is continuous from $\mathbf{H}(\Lambda_M)$ to $L^2(\partial M) \supset H^{1/2}(\partial M)$, the "double trace" $\gamma_{\partial M} \otimes \gamma_{\partial M}$ is continuous from $\mathbf{H}(\Lambda_M)\widehat{\otimes}\mathbf{H}(\Lambda_M)$ to $L^2(\partial M)\widehat{\otimes}L^2(\partial M) = L^2(\partial M \times \partial M)$. Hence the above lemma yields

$$G_R = (\gamma_{\partial M} \otimes \gamma_{\partial M})G_D \in L^2(\partial M \times \partial M),$$

which is obviously symmetric. This completes the proof of Proposition 3.1 and thus of Theorem 2.1.

4. An approximate model. The model we consider in this section is an approximation of the model introduced in section 2. Although more intricate in its presentation, it leads to a reduction of the computational cost of the time reversal operator, for it separates the respective roles of the TRM and the scatterers. It can be seen as the first step of an iterative method used in the context of multiple scattering problems (see, e.g., [7, 21] and [3, 2] for a rigorous justification of the method and [12] for an overview). The coupled problem of section 2 is solved by considering the successive reflections between the TRM and the scatterers. Here only specular waves, i.e., the first reflections, are taken into account. Comparing this model with that of section 2 will help us in section 5 to understand the influence of multiple scattering between the obstacles and the TRM upon the eigenelements of the time reversal operator.

Considering the same incident wave $\varphi_I = S_I g$ as in section 2, the diffracted field is now approximated near the TRM by a superposition of two waves: $\varphi_D = \varphi_D^{(1)} + \varphi_D^{(2)}$. The first one $\varphi_D^{(1)}$ represents the result of the diffraction of φ_I by the scatterers alone, i.e., the outgoing solution to

$$\begin{cases} \Delta\varphi_D^{(1)} + k^2\varphi_D^{(1)} = 0 & \text{in } \Omega_{\mathcal{O}}, \\ \Theta_{\mathcal{O}}\varphi_D^{(1)} = -\Theta_{\mathcal{O}}\varphi_I & \text{on } \partial\mathcal{O}. \end{cases} \tag{4.1}$$

The second one is the result of the diffraction of the latter by the TRM alone, i.e., the outgoing solution to

$$\begin{cases} \Delta\varphi_D^{(2)} + k^2\varphi_D^{(2)} = 0 & \text{in } \Omega_M, \\ \Theta_M\varphi_D^{(2)} = -\Theta_M\varphi_D^{(1)} & \text{on } \partial M. \end{cases} \tag{4.2}$$

We assume again that the TRM measures the trace of φ_D on ∂M. Hence the response of the medium is now described by the operator

$$Rg = \varphi_{D/\partial M} = (\varphi_D^{(1)} + \varphi_D^{(2)})_{/\partial M}.$$

Theorem 2.1 holds in this case: the time reversal operator $T = \overline{R}R$ is positive, self-adjoint, and compact in $L^2(\partial M)$.

The proof is similar to that of section 3. We simply have to replace the Green's function G_D by $G_D = G_D^{(1)} + G_D^{(2)}$, where $G_D^{(1)}$ and $G_D^{(2)}$ are, respectively, the outgoing solutions to

$$\begin{cases} \Delta G_D^{(1)}(x,.) + k^2 G_D^{(1)}(x,.) = 0 & \text{in } \Omega_{\mathcal{O}}, \\ \Theta_{\mathcal{O}} G_D^{(1)}(x,.) = -\Theta_{\mathcal{O}} G_I(x,.) & \text{on } \partial\mathcal{O} \end{cases}$$

and

$$\begin{cases} \Delta G_D^{(2)}(x,.) + k^2 G_D^{(2)}(x,.) = 0 & \text{in } \Omega_M, \\ \Theta_M G_D^{(2)}(x,.) = -\Theta_M G_D^{(1)}(x,.) & \text{on } \partial M. \end{cases}$$

LEMMA 4.1. *The following integral representation holds:*

$$\forall x \in \Omega_{M,\mathcal{O}}, \quad \varphi_D(x) = \int_{\partial M} G_D(x,y) g(y) d\sigma(y),$$

where G_D is symmetric in $\Omega_{M,\mathcal{O}} \times \Omega_{M,\mathcal{O}}$.

Proof. Contrary to Lemma 3.2, we are not able to give an intrinsic definition of the total field $\varphi_I + \varphi_D$ by means of a problem such as (2.1), which would depend only on the incident field. We thus give a direct proof of the above integral representation, starting from the classical formula (3.7) applied to $\varphi_D^{(1)}$ in $\Omega_{\mathcal{O}}$.

Using (3.5) and (3.8) applied to $\widetilde{G}_I(x,.)$ and $\varphi_D^{(1)}$ in $\Omega_{M,\mathcal{O}}$, formula (3.7) becomes

$$\forall x \in \Omega_{M,\mathcal{O}}, \quad \varphi_D^{(1)}(x) = \int_{\partial M \cup \partial\mathcal{O}} \left\{ -\Theta_\bullet G_I(x,y) \varphi_D^{(1)}(y) + G_I(x,y) \Theta_\bullet \varphi_D^{(1)}(y) \right\} d\sigma(y),$$

where $\Theta_\bullet$ stands for Θ_M or $\Theta_{\mathcal{O}}$. Thanks to the boundary conditions satisfied by $G_I(x,.)$ and $\varphi_D^{(2)}$, the contribution on ∂M is nothing but $S_I \Theta_M \varphi_D^{(1)} = -\varphi_D^{(2)}$ by Lemma 3.2, and so the contribution on $\partial\mathcal{O}$ is exactly $\varphi_D(x)$. We thus have

$$\begin{aligned} \varphi_D(x) &= \int_{\partial\mathcal{O}} \left\{ \Theta_{\mathcal{O}} G_D^{(1)}(x,y) \varphi_D^{(1)}(y) - G_I(x,y) \Theta_{\mathcal{O}} \varphi_I(y) \right\} d\sigma(y) \\ &= \int_{\partial\mathcal{O}} \left\{ G_D^{(1)}(x,y) \Theta_{\mathcal{O}} \varphi_D^{(1)}(y) - \Theta_{\mathcal{O}} G_I(x,y) \varphi_I(y) \right\} d\sigma(y) \\ &= \int_{\partial\mathcal{O}} \left\{ -G_D^{(1)}(x,y) \Theta_{\mathcal{O}} \varphi_I(y) + \Theta_{\mathcal{O}} G_D^{(1)}(x,y) \varphi_I(y) \right\} d\sigma(y) \\ &= \int_{\partial M} \left\{ G_D^{(1)}(x,y) \Theta_M \varphi_I(y) - \Theta_M G_D^{(1)}(x,y) \varphi_I(y) \right\} d\sigma(y). \end{aligned} \tag{4.3}$$

The first and third equalities result from the boundary conditions satisfied by $G_D^{(1)}(x,.)$ and $\varphi_D^{(1)}$. The second one derives from the reciprocity relation (3.8) applied, on one hand, to $G_D^{(1)}(x,.)$ and $\varphi_D^{(1)}$ in $\Omega_{\mathcal{O}}$ and, on the other hand, to $G_I(x,.)$ and φ_I in $\mathcal{O}$. The last one again results from (3.8) applied to $G_D^{(1)}(x,.)$ and φ_I in $\Omega_{M,\mathcal{O}}$. Noticing finally that $\Theta_M G_D^{(2)}(x,.) = -\Theta_M G_D^{(1)}(x,.)$, we have

$$\int_{\partial M} \Theta_M G_D^{(1)}(x,y) \varphi_I(y) d\sigma(y) = -\int_{\partial M} G_D^{(2)}(x,y) \Theta_M \varphi_I(y) d\sigma(y),$$

thanks to (3.8) applied to $G_D^{(2)}(x,.)$ and φ_I in Ω_M. Since $g = \Theta_M \varphi_I$, the integral representation of φ_D follows.

The symmetry of $G_D(x,.)$ is proved by the same argument as in Lemma 3.2. Since $G_D^{(1)}(x,.)$ and $G_D^{(2)}(x,.)$ play the same role as $\varphi_D^{(1)}$ and $\varphi_D^{(2)}$, we obtain for $G_D(x,y)$ a formula similar to (4.3):

$$G_D(x,y) = \int_{\partial\mathcal{O}} \left\{ \Theta_{\mathcal{O}} G_D^{(1)}(y,z) G_D^{(1)}(x,z) - G_I(y,z) \Theta_{\mathcal{O}} G_I(x,z) \right\} d\sigma(z),$$

where both terms are symmetric by (3.8). □

5. Two-dimensional numerical simulation. To solve numerically problems (2.3), (2.4), (4.1), and (4.2), we formulate them in bounded domains to apply a finite element method. We use the so-called coupling method between integral representation and finite elements, which is a nonsingular alternative to the well-known integral equation techniques. This method has been introduced by Jami and Lenoir [9] in hydrodynamics and then extended to many other wave propagation problems.

5.1. Bounded domain formulation. We describe the method only for problem (4.1), but the same technique is also applied for the other ones. We consider a bounded domain Ω' surrounding $\mathcal{O}$ and included in Ω_O (see Figure 5.1), and we introduce the following problem set in the domain Ω':

$$\begin{cases} \Delta\varphi' + k^2\varphi' = 0 & \text{in } \Omega', \\ \dfrac{\partial\varphi'}{\partial n} = h & \text{on } \partial\mathcal{O}, \\ \Theta_\Sigma \varphi' = \Theta_\Sigma \displaystyle\int_{\partial\mathcal{O}} \left\{ G_0(.,y) \frac{\partial\varphi'}{\partial n}(y) - \frac{\partial G_0}{\partial n_y}(.,y)\varphi'(y) \right\} d\sigma(y) & \text{on } \Sigma, \end{cases} \tag{5.1}$$

where $\Sigma = \partial\Omega' \backslash \partial\mathcal{O}$ and $\Theta_\Sigma = (\partial/\partial n + \beta)$, β being an arbitrary complex parameter.

It is clear that if $\varphi_D^{(1)}$ is a solution of (4.1), then $\varphi' = \varphi^{(1)}_{D_{/\Omega'}}$ is a solution of problem (5.1). Similarly, provided $\mathrm{Im}(\beta) \neq 0$, every solution φ' of (5.1) can be uniquely extended to a solution $\varphi_D^{(1)}$ of (4.1) by the integral representation formula on $\partial\mathcal{O}$:

$$\forall x \in \Omega_{\mathcal{O}}, \quad \varphi_D^{(1)}(x) = \int_{\partial\mathcal{O}} \left\{ G_0(x,y) \frac{\partial\varphi'}{\partial n}(y) - \frac{\partial G_0}{\partial n_y}(x,y)\varphi'(y) \right\} d\sigma(y). \tag{5.2}$$

The variational formulation of the problem (5.1) is the following:

$$\begin{cases} \text{Find } \varphi' \in H^1(\Omega') \text{ such that } \forall \psi \in H^1(\Omega'), \text{ we have} \\ \displaystyle\int_{\Omega'} \nabla\varphi'.\overline{\nabla\psi} - k^2 \int_{\Omega'} \varphi'\overline{\psi} - \beta \int_\Sigma \varphi'\overline{\psi} \\ \displaystyle + \int_\Sigma \overline{\psi(x)} \int_{\partial\mathcal{O}} \varphi'(y) \left(\frac{\partial}{\partial n_x} + \beta \right) \frac{\partial G_0}{\partial n_y}(x,y) d\sigma(y) d\sigma(x) \\ \displaystyle = -\int_{\partial\mathcal{O}} h\overline{\psi} + \int_\Sigma \overline{\psi(x)} \int_{\partial\mathcal{O}} h(y) \left(\frac{\partial}{\partial n_x} + \beta \right) G_0(x,y) d\sigma(y) d\sigma(x). \end{cases} \tag{5.3}$$

Finally, we discretize problem (5.3) to obtain a linear system that we solve numerically.

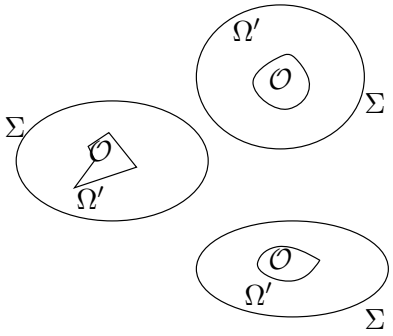


FIG. 5.1. *Bounded domain* Ω'.

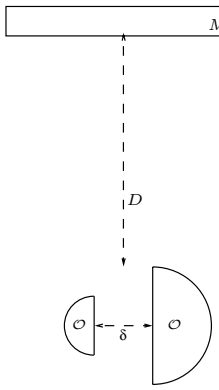


FIG. 5.2. *Geometry of the problem.*

5.2. Numerical results. All the numerical results are obtained by the code MELINA [11]. We consider an oblong mirror of width 8 and height 1 and two half-disk obstacles of diameters 4 and 2 (see Figure 5.2). We denote by D the distance between the mirror and the obstacles. The distance between the scatterers is $\delta = 2$. We investigate two cases: $D = 3$ and $D = 8$. We consider here a Neumann condition on the boundary of the mirror ∂M ($\mu = 0$, that is, the case of a large acoustic impedance of the transducers).

Figure 5.3 (respectively, Figure 5.4) shows the amplitude of the total field corresponding to the emission of the first (respectively, second) eigenvector associated with $\lambda_1 = 0.0499$ if $D = 8$ and $\lambda_1 = 0.2211$ if $D = 3$ (respectively, $\lambda_2 = 0.0191$ if $D = 8$ and $\lambda_2 = 0.0534$ if $D = 3$) in the case of the first model presented in section 2 and where $k = 3.14$ (the wavelength $l_w = 2\pi/k = 2$ is then equal to the distance between the obstacles δ). We observe that the wave is focused on the biggest obstacle (respectively, the smallest). When emitting the third eigenvector associated with $\lambda_3 = 0.0002$ if $D = 8$ and $\lambda_3 = 0.0085$ if $D = 3$, we see in Figure 5.5 that the wave again focuses on the biggest scatterer, although it seems less concentrated in its vicinity than for the first eigenvector. These results essentially show that selective focusing, which is known to occur for small and distant enough scatterers [8], is achieved even though the size of the obstacles, the distance between them, and the wavelength are of the same order.

Figure 5.10 shows the first four eigenvalues of the time reversal operator T ac-

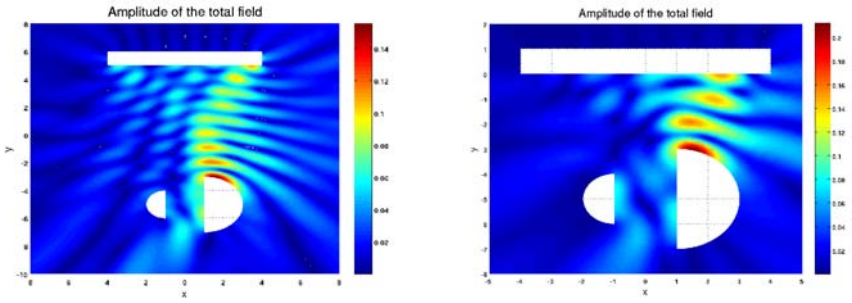


FIG. 5.3. *Emission of the first eigenvector for* $k = 3.14$ *(left:* $D = 8$*, right:* $D = 3$*).*

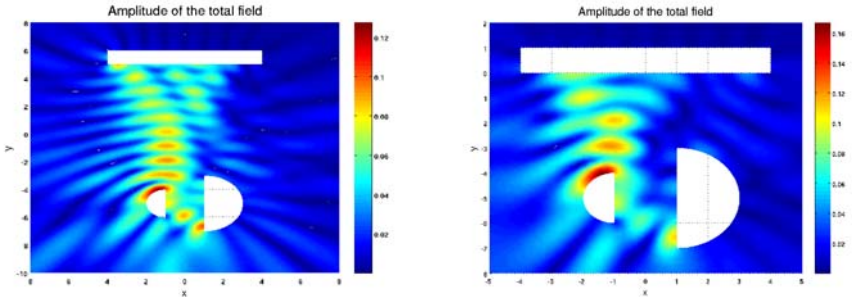


FIG. 5.4. *Emission of the second eigenvector for* $k = 3.14$ *(left:* $D = 8$*, right:* $D = 3$*).*

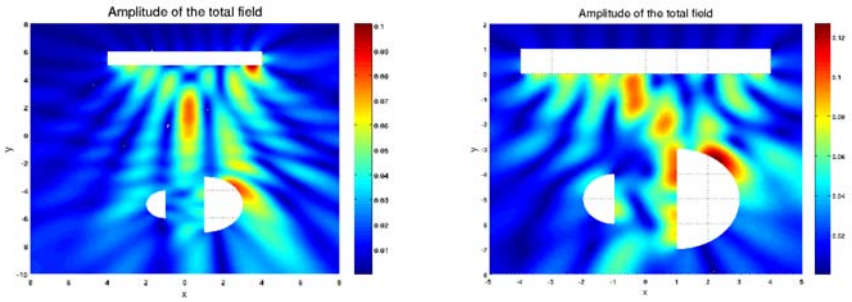


FIG. 5.5. *Emission of the third eigenvector for* $k = 3.14$ *(left:* $D = 8$*, right:* $D = 3$*).*

cording to the wave number k in the case of the first model presented in section 2, where the interactions between the mirror and the obstacles are taken into account and where the distance D between them is, respectively, 8 or 3. Figure 5.11 shows the same results in the case of the second model, where these interactions are approximated by only the first reflections.

We can first notice in these figures that there is only one significant eigenvalue λ_1 at low frequencies, which follows from the fact that the wavelength l_w is large compared to the distance δ between the two obstacles, so that the mirror cannot distinguish between them and see them as only one. For $k = 0.325$ (that is, $l_w \simeq 19$), this is illustrated by Figure 5.6 (respectively, Figure 5.7), which shows the amplitude of the total field corresponding to the emission of the first (respectively, second) eigenvector associated with $\lambda_1 = 0.6994$ if $D = 8$ and $\lambda_1 = 3.2717$ if $D = 3$ (respectively,

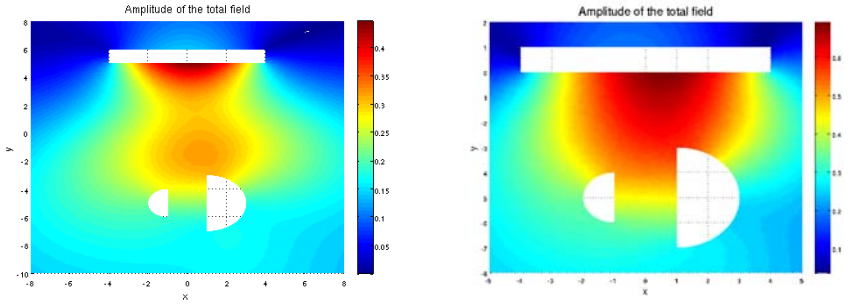


FIG. 5.6. *Emission of the first eigenvector for* $k = 0.325$ *(left:* $D = 8$*, right:* $D = 3$*).*

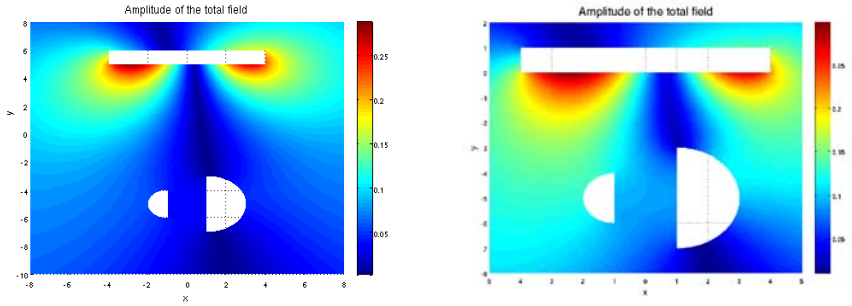


FIG. 5.7. *Emission of the second eigenvector for* $k = 0.325$ *(left:* $D = 8$*, right:* $D = 3$*).*

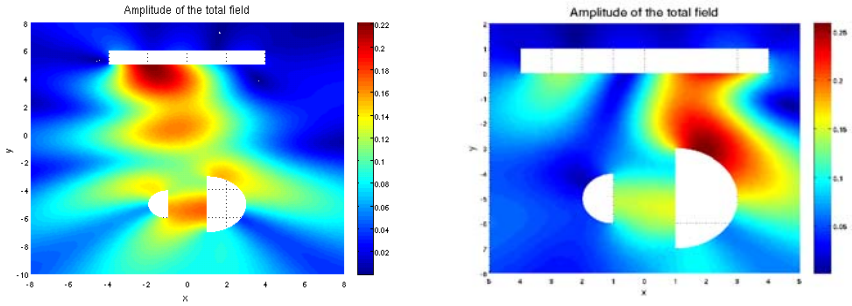


FIG. 5.8. *Emission of the first eigenvector for* $k = 0.875$ *(left:* $D = 8$*, right:* $D = 3$*).*

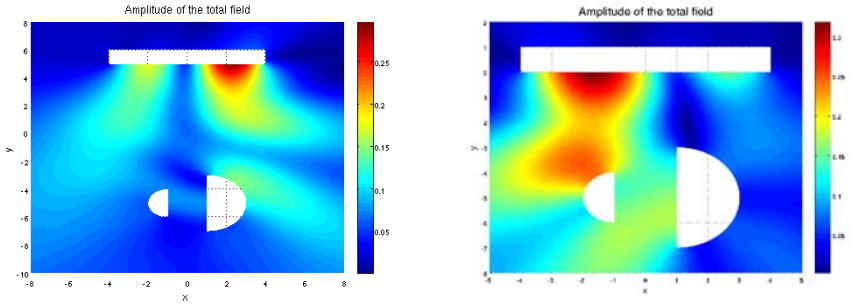


FIG. 5.9. *Emission of the second eigenvector for* $k = 0.875$ *(left:* $D = 8$*, right:* $D = 3$*).*

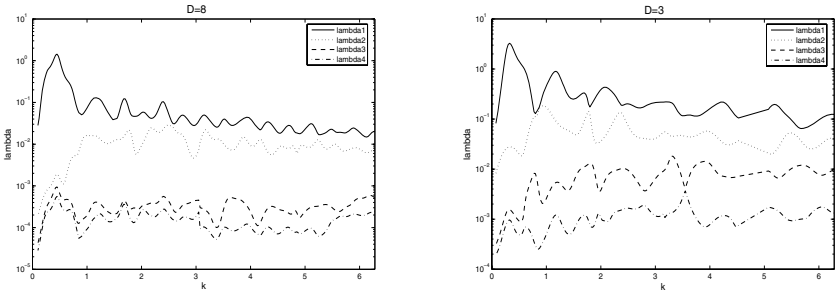


FIG. 5.10. *First model: four largest eigenvalues of T according to k.*

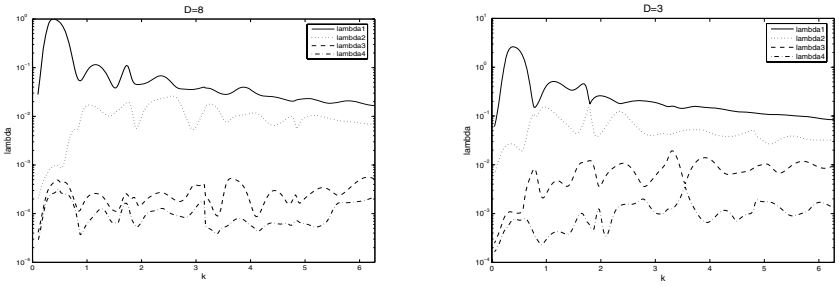


FIG. 5.11. *Second model: four largest eigenvalues of T according to k.*

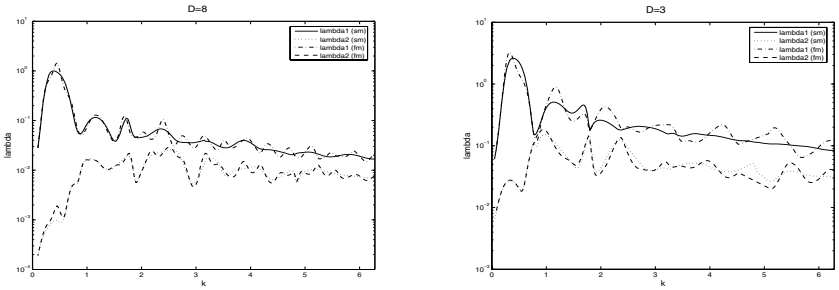


FIG. 5.12. *The first two eigenvalues for the two models.*

$\lambda_2 = 0.0006$ if $D = 8$ and $\lambda_2 = 0.0265$ if $D = 3$) in the case of the first model presented in section 2.

When k increases, Figures 5.10 and 5.11 show that there are two significant eigenvalues: the gap with the third eigenvalue is pronounced when $D = 8$ and becomes smaller when $D = 3$. This confirms a well-known effect: this gap increases when the angular aperture under which the TRM is seen from the obstacles decreases.

We note the presence of oscillations of the first two eigenvalues of T in the case of the first model (Figure 5.10) contrary to the case of the second model (Figure 5.11). To understand this, we show in Figure 5.12 the first two eigenvalues of the two models where, respectively, $D = 3$ and $D = 8$. We remark that the eigenvalues for the first model oscillate around the corresponding eigenvalues for the second model. This can be explained by the fact that the interactions between the mirror and the obstacles can be constructive or destructive according to the distance between the mirror and

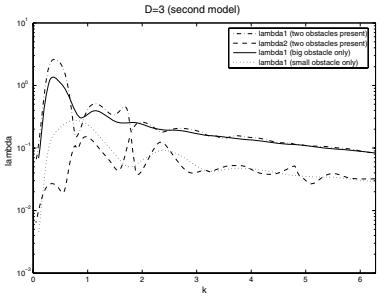


FIG. 5.13. *Analysis of the interactions between obstacles.*

the obstacles. For the greatest eigenvalue, we note the dependence of the period of oscillations on the distance between the mirror and the obstacles: $\Delta k \simeq \pi/D$. More precisely, the interactions between the mirror and the obstacles are constructive for the wave numbers $k_n \simeq n\pi/D$, $n \in \mathbb{N}^*$ (which correspond to local maxima of λ_1), and destructive for $k_n \simeq ((n-1/2)\pi)/D$, $n \in \mathbb{N}^*$. This can be explained by the fact that the wave numbers $n\pi/D$, $n \in \mathbb{N}^*$, represent the eigenvalues of the operator $-\Delta$ with Neumann conditions on the boundaries in the one-dimensional domain $[0, D]$, and the corresponding eigenfunctions are $\varphi_n(y) = \cos(k_n y)$, $y \in [0, D]$.

To study the interactions between the obstacles, we now show in Figure 5.13 the first and second eigenvalues of T for the approximate model and for $D = 3$, together with the first eigenvalue of T for the same model but with a new geometrical configuration in which only the biggest or the smallest obstacle is present. We observe that, at medium frequencies, there is a good coincidence between the first eigenvalue of T corresponding to the case where the two obstacles are present and the one where there is only the biggest obstacle, which explains that the interactions due to the smallest obstacle are negligible. Meanwhile, the second eigenvalue of T corresponding to the case where the two obstacles are present oscillates smoothly around the first eigenvalue where there is only the smallest obstacle, which proves that the interactions due to the biggest obstacle are important.

Figure 5.10 shows that the two greatest eigenvalues become very close near particular values of k. When the time reversal operator has a double eigenvalue, which occurs, for instance, in the case of two identical targets, one generally requires additional information to identify the selective focusing fields within the two-dimensional eigenspace. Here the slight distance between both eigenvalues seems sufficient to identify these fields. Consider, for example, the case $D = 3$ and $k = 0.875$, where $\lambda_1 = 0.1760$ and $\lambda_2 = 0.1637$. The associated eigenvectors generate the fields represented in Figures 5.8 and 5.9 (right). Although both targets are separated by less than one-third of a wavelength, the TRM clearly distinguishes them, contrary to the low frequency situation of Figure 5.6. This effect results from the proximity of the TRM which acts in the near field. Indeed, if the TRM is moved away to $D = 8$ with the same wave number (Figures 5.8 and 5.9 (left)), both targets are seen as a single one, as in the above-mentioned low frequency case.

REFERENCES

[1] J. P. AUBIN, *Analyse Fonctionnelle Appliquée*, Tome II, Presses Universitaires de France, Paris Cedex, France, 1987.

[2] M. Balabane, *Boundary decomposition for Helmholtz and Maxwell equations* 1: *Disjoint subscatterers*, Asymptot. Anal., 38 (2004), pp. 1–10.
[3] M. Balabane and V. Tirel, *Décomposition de domaine pour un calcul hybride de l'équation de Helmholtz*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 281–286.
[4] L. Borcea, G. Papanicolaou, C. Tsogka, and J. Berryman, *Imaging and time reversal in random media*, Inverse Problems, 18 (2002), pp. 1247–1279.
[5] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, 1993.
[6] R. Dautray and J. L. Lions, *Analyse Mathématique et Calcul Numérique pour les Sciences et les Techniques*, Tome I, Masson, Paris, 1985.
[7] E. Domany and O. Entin-Wohlman, *Application of multiple scattering theory to subsurface defects*, J. Appl. Phys., 56 (1984), pp. 137–142.
[8] C. Hazard and K. Ramdani, *Selective acoustic focusing using time-harmonic reversal mirrors*, SIAM J. Appl. Math., 64 (2004), pp. 1057–1076.
[9] A. Jami and M. Lenoir, *A variational formulation for exterior problems in linear hydrodynamics*, Comput. Methods Appl. Mech. Engrg., 16 (1978), pp. 341–359.
[10] J. F. Lingevitch, H. C. Song, and W. A. Kuperman, *Time reversed reverberation focusing in a waveguide*, J. Acoust. Soc. Am., 111 (2002), pp. 2609–2614.
[11] D. Martin, *Documentation MELINA*, http://www.maths.univ-rennes1.fr/∼dmartin/melina/www/homepage.html (2006).
[12] P. A. Martin, *Multiple Scattering: Interaction of Time-Harmonic Waves with N Obstacles*, Cambridge University Press, Cambridge, UK, 2006.
[13] T. D. Mast, A. I. Nachman, and R. C. Waag, *Focusing and imaging using eigenfunctions of the scattering operator*, J. Acoust. Soc. Am., 102 (1997), pp. 715–725.
[14] G. Micolau and M. Saillard, *D.O.R.T method as applied to electromagnetic subsurface sensing*, Radio Sci., 38 (2003), pp. 1038–1049.
[15] N. Mordant, C. Prada, and M. Fink, *Highly resolved detection and selective focusing in a waveguide using the D.O.R.T method*, J. Acoust. Soc. Am., 105 (1999), pp. 2634–2642.
[16] A. D. Pierce, *Acoustics, An Introduction to Its Physical Principle and Applications*, McGraw–Hill, New York, 1981.
[17] C. Prada and M. Fink, *Eigenmodes of the time reversal operator: A solution to selective focusing in multiple-target media*, Wave Motion, 20 (1994), pp. 151–163.
[18] C. Prada, E. Kerbrat, D. Cassereau, and M. Fink, *Time reversal techniques in ultrasonic nondestructive testing of scattering media*, Inverse Problems, 18 (2002), pp. 1761–1773.
[19] C. Prada, S. Manneville, D. Spoliansky, and M. Fink, *Decomposition of the time reversal operator: Detection and selective focusing on two scatterers*, J. Acoust. Soc. Am., 99 (1996), pp. 2067–2076.
[20] H. A. Schenck, *Helmholtz integral formulation of the sonar equations*, J. Acoust. Soc. Am., 79 (1986), pp. 1423–1433.
[21] G. T. Schuster, *A hybrid BIE + Born series modeling scheme: Generalized Born series*, J. Acoust. Soc. Am., 77 (1985), pp. 865–879.
[22] H. Tortel, G. Micolau, and M. Saillard, *Decomposition of the time reversal operator for electromagnetic scattering*, J. Electromagn. Waves Appl., 13 (1999), pp. 687–719.

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 792–809

ON STEP-FUNCTION REACTION KINETICS MODEL IN THE ABSENCE OF MATERIAL DIFFUSION*

DMITRY GOLOVATY†

Abstract. We propose a precise definition of the step-function kinetics suitable for approximating diffuse propagating reaction fronts in one-dimensional gasless-combustion-type models when a Lewis number is large. We investigate this kinetics in the context of free-radical frontal polymerization (FP) in which a monomer-initiator mixture is converted into a polymer via a propagating self-sustaining reaction front. The notion of step-function kinetics has been extensively used in studies of the frontal dynamics both in FP and in combustion problems when the material diffusion is negligible. However, the models have always been effectively reduced to their point-source approximations without defining exactly what the step-function kinetics is for diffuse reaction fronts. We demonstrate numerically that dynamics of diffuse fronts in systems modeled with step-function kinetics and in systems modeled with Arrhenius kinetics are qualitatively the same at time scales at which the bulk reaction ahead of the front can be ignored. We perform stability analysis for the traveling reaction wave and show that the stability threshold is in close agreement with numerical simulations as well as with other existing kinetics approximations. The benefits of using step-function kinetics are two-fold. The reaction dynamics predicted by the step-function kinetics approximates the dynamics predicted by the Arrhenius kinetics over a wider range of system parameters than the point-source approximation. Second, the systems governed by the step-function kinetics can be analyzed both analytically and numerically within the framework of a single model.

Key words. frontal polymerization, gasless combustion, Arrhenius kinetics, traveling wave, reaction-diffusion equations

AMS subject classifications. 35B35, 35R35, 92E20

DOI. 10.1137/050628805

1. Introduction.

1.1. Physical background and existing modeling approaches.

In this paper we give a precise definition of the step-function kinetics in the absence of material diffusion and study it in the context of free-radical *frontal polymerization* (FP).

Frontal polymerization is a process in which a monomer converts into a polymer via a self-propagating localized reaction wave [4], [5]. A typical frontal polymerization experiment is performed in a glass tube filled with reagents. An external heat source, when applied at the top of the tube, initiates a descending front that appears as a moving region of polymer formation. Depending on the choice of reactants and the conditions of the experiment, the front either may or may not propagate with a constant speed. Various nonuniform propagation scenarios can occur, even if it is assumed that the front always remains flat—the situation considered in this paper.

There are several conditions necessary for the existence of the frontal mode. First, the ignition temperature must be high enough to generate and initially sustain the reaction front. Further, the reaction rate must be extremely small at the initial (ambient) temperature but very large at the front temperature. The high reaction

*Received by the editors April 8, 2005; accepted for publication (in revised form) December 4, 2006; published electronically March 15, 2007. This work was supported in part by the NSF grant DMS-0305577.

http://www.siam.org/journals/siap/67-3/62880.html

†Department of Theoretical and Applied Mathematics, University of Akron, Akron, OH 44325 (dmitry@math.uakron.edu).

rate coupled with the exothermicity of the reaction must be sufficient to overcome heat losses into the reactants and product zones [5].

Note that an alternative to frontal polymerization is *bulk polymerization*, in which a mixture of reagents is heated uniformly and polymer formation occurs simultaneously throughout the mixture.

A more extensively studied chemical process with a frontal reaction mechanism is self-propagating high-temperature synthesis (SHS)—a combustion process characterized by a heat release large enough to propagate a combustion front through a powder compact while consuming the reactant powders [6], [13]. The simplest models and front propagation mechanisms for FP and SHS are essentially the same, except for the magnitudes of the model parameters.

Both steady and unsteady front propagation have been observed in FP [18] as well as in SHS [14]. Unsteady front propagation is usually undesirable, as it leads to nonuniform "layered" structure of the final product. One of the goals of our modeling is to determine the range of material parameters within which the stability of a uniformly propagating polymerization front is guaranteed. The analysis of the full model is, however, too complicated because it requires solving a system of coupled nonlinear partial differential equations describing multiple reactions and energy transport. In order to make analytical predictions, numerous simplifications are usually introduced by employing asymptotics in terms of small parameters, considering effective kinetics, etc.

In the presence of an appropriate small nondimensional parameter, the reaction zone can be replaced by a propagating front with the chemical reaction approximated by a heat source attached to the front (point-source kinetics [12]). With the removal of a nonlinear reaction term, the governing equations become significantly simpler, but, because the location of the front is not known a priori, the reduced problem is of a free-boundary type. The approximate problem is easier to study analytically, especially from the point of view of stability analysis for the traveling wave solutions.

Even though sharp-front approximation is not usually derived via a rigorous asymptotic method, it has been shown to be an effective tool to study SHS and FP problems, yielding qualitatively plausible results. On the negative side, the problems with point-source kinetics are difficult to treat numerically [9]; further, for certain regimes of FP, *the main reaction zone does not always remain narrow even in the presence of a small parameter*, resulting in nonnegligible bulk reactions that are ignored by default within point-source kinetics approximation. Indeed, for oscillating reaction waves, the concentration of the monomer does not evolve via the frontal mode alone [3], as the width of the main reaction zone varies periodically by several orders of magnitude, leaving regions of unreacted monomer behind the polymerization front. The monomer in these regions subsequently converts to polymer via bulk (nonfrontal) polymerization.

Another approach that has been successfully applied in a number of combustion and polymerization studies is to introduce simplified distributed kinetics [1], [10], which is usually combined with narrow reaction zone approximation [11], [16], [20]. Within this approach, the Arrhenius temperature dependence is replaced by a step-function with height equal to the value of the Arrhenius function at a solution-dependent characteristic temperature. The exact choice of characteristic temperature is determined by the physics of the problem. Although the kinetics function in this setup is very simple, the strong nonlinearity of the Arrhenius kinetics is preserved by making the characteristic temperature dependent on the solution.

The advantage of the step-function kinetics is that the traveling wave solution can

generally be found analytically. The stability analysis for this solution is, however, very tedious, and researchers have resorted to additional simplifications, in particular via narrow-reaction-zone-type asymptotics that essentially lead back to point-source kinetics.

Typically, the front is postulated to have a width that is determined by a small nondimensional parameter ϵ (cf. (1.11)) [11], [16], [20]. Then the characteristic temperature is set as a limit of the temperature in an outer solution instead of prescribing the explicit formula for the characteristic temperature for a diffuse front and using a rigorous asymptotic procedure. Since the characteristic temperature for the traveling wave solution is indeed the same as the appropriate outer temperature limit at the interface, this approach successfully captures the stability threshold for the fronts propagating with a constant speed. On the other hand, narrow reaction zone approximations of step-function kinetics suffer from the same limitations as those of point-source kinetics.

To our knowledge, *there is no self-consistent distributed step-function kinetics formulation appropriate for modeling of diffuse fronts in the absence of material diffusion.* In this paper, we introduce a version of such kinetics in the context of one-dimensional FP/SHS. The benefits of using step-function kinetics are two-fold. The solution profiles for the (distributed) step-function kinetics closely resemble the solution profiles for the Arrhenius kinetics models over a wider range of system parameters than do the solution profiles for the point-source models. In particular, possible departures from the purely frontal reaction mechanism can be studied by using distributed step-function kinetics but not point-source kinetics. Second, the systems governed by step-function kinetics can be analyzed both analytically and numerically within the framework of a single model.

We perform the stability analysis for the traveling reaction wave in a diffuse-front setting and determine the stability boundary. In order to make the analysis tractable, we assume that the nondimensional parameter ϵ defined in (1.11) is small. This assumption appears in a number of other works [11], [16], [20] as a justification for considering sharp-front asymptotics of the step-function kinetics. The analysis of [3] indicates that there is no clear relationship between the value of ϵ and the width of the reaction zone, especially for pulsating fronts. Hence *we do not consider the reaction zone to be narrow* in our calculations.

Because the algebra involved in handling perturbations of the ground state and the resulting form of the dispersion relation are very complex, we handle some of the symbolic calculations and solve the dispersion relation in Maple. We obtain the stability threshold and show that it is in excellent agreement with numerical predictions and the existing sharp-front approximations. The combination of analytical computations and Maple has a clear advantage over the full numerical simulations in that it does not require numerical solution of a system of partial differential equations. Also the former requires a significantly shorter computational time (minutes versus hours) even for a single simulation run.

The computational costs are considerably lower for the step-function kinetics model than for the point-source kinetics model since the position of the front is not one of the unknowns in the problem. The numerically determined behavior of the front for the step-function kinetics is qualitatively similar to that under the Arrhenius kinetics [7], [18], as it shows a similar hierarchy of dynamics and similar solution features, including those that result from the nonfrontal mode of polymerization. The same spectrum of system behaviors has also been demonstrated for the point-source kinetics [9]; however, all bulk (nonfrontal) reactions are ignored in this setting.

1.2. Mathematical model. Although the mechanism of free-radical polymerization involves three steps—initiation, propagation, and termination—and five reagents—an initiator, an active initiator radical, an active polymer radical, a monomer, and a complete polymer chain [18]—we will make a number of simplifying assumptions that reduce the complexity of the underlying mathematical model. Hence we will assume [16], [18], [19] the following:

- The rates of reactions between the initiator radicals and the monomer and between the polymer radicals and the monomer are the same.
- The rate of change of total radical concentration is much smaller than the rates of their production and consumption.
- The initial concentration of the initiator is so large that it is not appreciably consumed during the polymerization process.
- Material diffusion is negligible compared to thermal diffusion.
- Both reagents and the final product are viscous enough to ignore convective effects and bubble formation.
- The test tube is sufficiently thin with the adiabatic boundary conditions on sidewalls so that the spatial dependence of the solution can be restricted to the axial variable.

Suppose that a test tube containing the monomer-initiator mixture occupies a region $\Omega \in \mathbf{R}^3$, and denote by $M(x,t)$ the monomer concentration and by $T(x,t)$ the temperature of the mixture at the point $x \in \Omega$ and the time $t > 0$. Then the process of free-radical polymerizations can be described [16] by what is known as a single-step effective kinetics model of monomer-to-polymer conversion:

$$\frac{\partial M}{\partial t} = -kMe^{\frac{E}{R_g T_b}\left(1-\frac{T_b}{T}\right)}, \tag{1.1}$$

$$\frac{\partial T}{\partial t} = \operatorname{div}(\kappa \nabla T) + kqMe^{\frac{E}{R_g T_b}\left(1-\frac{T_b}{T}\right)}, \tag{1.2}$$

where κ is a thermal diffusivity of the mixture/final product, k is the effective pre-exponential factor in the Arrhenius kinetics, R_g is the gas constant, E is the effective activation energy, and T_b is a reference temperature that will be specified below. The constant parameter q is $-\frac{\Delta H}{c\rho}$, where ΔH is the reaction enthalpy; c and ρ are the specific heat and the mixture density, respectively.

Throughout this paper we will assume that the test tube is one-dimensional, $\Omega = [-L\,, L]$, and that the thermal diffusivity κ is constant. (We ignore possible dependence of κ on temperature and degree of conversion $1 - M/M_0$.) Then the problem (1.1)–(1.2) reduces to

$$\frac{\partial M}{\partial t} = -kMe^{\frac{E}{R_g T_b}\left(1-\frac{T_b}{T}\right)}, \tag{1.3}$$

$$\frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial x^2} + kqMe^{\frac{E}{R_g T_b}\left(1-\frac{T_b}{T}\right)}. \tag{1.4}$$

We will assume that T and M satisfy the constant initial conditions

$$T(x,0) = T_0, \quad M(x,0) = M_0, \quad x \in [-L\,, L]\,. \tag{1.5}$$

In order to initiate the reaction, heat must be supplied to the system; hence for the first t_0 seconds we will use the following boundary conditions

$$T_x(-L\,, t) = 0\,, \quad M_x(\pm L\,, t) = 0\,, \quad T(L\,, t) = T_b\,, \quad t \in (0\,, t_0)\,. \tag{1.6}$$

During the front propagation regime, we will impose the adiabatic and impenetrability boundary conditions on the temperature and the monomer concentration, respectively, by setting

$$T_x(\pm L\,,t) = 0\,, \quad M_x(\pm L\,,t) = 0\,, \quad t \geq t_0\,. \tag{1.7}$$

Multiplying (1.3) by q, adding the resulting equation to (1.4), integrating with respect to x, applying the adiabatic boundary conditions in (1.7), and setting

$$H := \int_{-L}^{L} (T + qM)\,dx \tag{1.8}$$

yields

$$\frac{dH}{dt} = 0\,, \tag{1.9}$$

expressing conservation of enthalpy in the system when $t > t_0$. Thermodynamics of the problem dictates that the temperature of the reaction products away from the front is given by

$$T_b \;=\; T_0 + qM_0\,, \tag{1.10}$$

where T_0 and M_0 are the initial temperature and concentration, respectively, in (1.5).

We introduce dimensionless parameters

$$\epsilon = \frac{R_g T_b}{E}\,, \quad Z = \frac{qM_0E}{R_g T_b^2}\,, \tag{1.11}$$

$$\tilde{t} = \frac{kt}{Z}\,, \quad \tilde{x} = \sqrt{\frac{k}{Z\kappa}}x\,, \quad \tilde{M} = \frac{M}{M_0}\,, \quad \tilde{T} = \frac{T - T_0}{T_b - T_0}\,.$$

Here T_b is as defined in (1.10), and the Zeldovich number Z is a nondimensionalized activation energy [15] constructed as a ratio of the diffusion temperature scale $T_b - T_0$ to the reaction temperature scale $\frac{R_g T_b^2}{E}$. Also, note that $Z\epsilon < 1$ in order to insure that the initial temperature of the mixture is greater than absolute zero. Then (after dropping tildes) we obtain

$$\frac{\partial M}{\partial t} = -ZM \exp\left(\frac{Z(T-1)}{\epsilon Z(T-1)+1}\right), \tag{1.12}$$

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} + ZM \exp\left(\frac{Z(T-1)}{\epsilon Z(T-1)+1}\right). \tag{1.13}$$

From (1.5)–(1.7), the nondimensional temperature T and concentration M satisfy the following initial and boundary conditions:

$$T(x,0) = 0, \quad M(x,0) = 1, \quad x \in [-l\,,l]\,, \tag{1.14}$$

$$M_x(\pm l\,,t) = 0\,, \quad T_x(-l\,,t) = 0\,, \quad T(l\,,t) = 1\,, \quad t \in (0\,,\tau_0)\,, \tag{1.15}$$

$$M_x(\pm l\,,t) = 0\,, \quad T_x(\pm l\,,t) = 0\,, \quad t \geq \tau_0\,, \tag{1.16}$$

where $l = \sqrt{k/Z\kappa}L$ and $\tau_0 = k\,t_0/Z$.

2. Step-function kinetics. Stability analysis and numerical results.

2.1. Motivation. The essential, Arrhenius-like dependence of the reaction rate on temperature is a common feature of the full Arrhenius kinetics model, the point-source model, and the step-function kinetics model introduced below. As a consequence, all three models demonstrate similar sequences of frontal propagation modes, albeit at different threshold values of the bifurcation parameters. The advantage of the approximate models lies mainly with the simplicity of their analytical treatment.

The relatively simple point-source model is a formal asymptotic reduction of the Arrhenius kinetics model in which the reaction is restricted to a propagating interface between the fresh mixture of reagents and the final product. The rigorous proof of this reduction does not exist. In fact, in [3] we demonstrated for oscillating reaction waves corresponding to large values of the Zeldovich number that the concentration of the monomer does not evolve via the frontal mode alone. Numerical simulations in [3] have shown that the width of the main reaction zone periodically varies in time by several orders of magnitude, and there are always instances when the reaction zone is not narrow. Further, the propagation does not occur purely in the frontal regime, as the "pockets" of unreacted monomer are periodically left behind the reaction front. These pockets then slowly convert into the polymer via a nonfrontal mechanism.

Clearly, the nonfrontal reactions cannot be modeled within the point-source model. The purpose of distributed step-function-kinetics is to model diffuse interfaces and nonlocalized reactions while retaining the relative simplicity of an analytical treatment characteristic of the point-source models. In [3], we showed that the step-function-kinetics model introduced in the present paper reproduces all features of reaction dynamics observed for the Arrhenius kinetics models—including the monomer pockets and the oscillating width of the main reaction zone. That is, step-function kinetics is capable of capturing not only the sequence of bifurcations between the different modes of frontal propagation, but the solution features as well.

From now on, we will assume that ϵ is small; then the system of equations (1.12)–(1.13) reduces to

$$\frac{\partial M}{\partial t} = -ZMe^{Z(T-1)}\,, \tag{2.1}$$

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} + ZMe^{Z(T-1)}\,. \tag{2.2}$$

When the second nondimensional parameter Z is large, this model has been approximated by using the step-function kinetics [2], [16] in a sharp-front limit as $Z \to \infty$. In this limit the model is identical to the sharp-front model of solid combustion with point-source kinetics at the interface considered in [12]. Here we introduce and investigate the behavior and the stability of solutions of the (distributed) step-function kinetics model and compare the results to both Arrhenius kinetics and point-source kinetics. The step-function kinetics can be thought of as an intermediate approximation between point-source kinetics and Arrhenius kinetics in that it yields to relatively straightforward analytical as well as numerical analyses. As is well known, although the systems modeled with Arrhenius kinetics can be studied numerically, the stability analysis for such systems is very difficult, as no analytical expressions are available for traveling wave solutions. On the contrary, for point-source kinetics (sharp-interface approach) the stability analysis is straightforward, while the numerical computations are difficult, as one has to track a free boundary.

A similar model was considered in [1] for a more general case when the Lewis number Le (the ratio of thermal and material diffusivities) is *not* necessarily large. This work has served as a basis for numerous studies in FP and combustion. As will be explained shortly, the straightforward reduction of the treatment in [1] to (2.1)–(2.2) cannot be considered a priori within an asymptotic procedure as $Le \to \infty$, as it requires additional assumptions. These assumptions will be introduced below and studied both analytically and numerically in order to verify their validity.

In the step-function kinetics model, the reaction is assumed to occur in the temperature range $[T_i\,, T_p]$, and the Arrhenius term $Z\,e^{Z(T-1)}$ in (2.1)–(2.2) is replaced by the step-function

$$K(T) := \begin{cases} 0\,, & T < T_i\,, \\ A(T_p)\,, & T \geq T_i\,. \end{cases} \tag{2.3}$$

Here T_p is the burnout temperature of the mixture immediately upon the completion of the reaction (or, analogously, the temperature at the product end of the reaction zone), and T_i is the ignition temperature. The temperature T_p is, generally, the highest temperature of the mixture, and, unless the front is a steadily propagating wave, both temperatures T_p and T_i as well as the kinetic function $K(T)$ depend on time.

The step-function-type and Arrhenius-type source terms have two important common features—the reactions at both the low-temperature and the product zones are either absent or negligible and the speed of propagation of the reaction wave is temperature-dependent. By selecting a proper height $A(T_p)$ and a jump location T_i in the definition of the step-function, we can obtain the "best" qualitative match between the properties of solutions of the Arrhenius and step-function kinetics models.

Note that the dependence of A on T_p can be arbitrary, and T_i and T_p are essentially "fitting" parameters—we claim neither that the step-function approximation of Arrhenius kinetics is mathematically rigorous nor that step-function kinetics can be obtained from Arrhenius kinetics as a result of a limiting procedure. Rather, we observe that the behavior of solutions of two respective systems is in a qualitative agreement over a wide range of the system characteristics.

Even though the temperatures T_i and T_p have a clear physical interpretation as the temperatures at which the (Arrhenius-type) reaction initiates and completes, respectively, the values of T_i and T_p are not unique, since the monomer concentration in the Arrhenius model is a continuous nonvanishing function of space and time. Instead, we identify the valid ranges for the ignition and burnout temperatures and show that the results are not sensitive to the exact choices of T_i and T_p.

The downside of this approach is that it can lead to a number of different step-function-type kinetics models. On the other hand, the flexibility of choosing A, T_i, and T_p allows for different step-function kinetics for systems that occupy well-separated regions in the parameter space. In the present paper, we concentrate on the case when $\epsilon \ll 1/Z \ll 1$—and the reaction is usually understood as being purely frontal. For this relationship between the nondimensional parameters, it is our experience that, as long as T_i and T_p are "reasonable" (e.g., the concentration of the monomer at the burnout temperature is small, but not too small), the solutions remain qualitatively similar.

2.2. Formulation and traveling wave solution. Here we follow formal arguments along the lines of [21] to examine the relationship between the temperatures

T_i and T_p and to establish the expression for A. Then we discuss the proper choice of T_p.

The main idea is to preserve the dependence between the temperature and the speed of the reaction front when replacing the full kinetics by an approximate kinetics. Formally, for Arrhenius kinetics, the fast reaction occurs in a narrow temperature interval $[T_p - 1/Z\,, T_p]$, where $1/Z \ll 1$ [21]; therefore we set

$$T_i = T_p - \frac{1}{Z}\,.$$

The energy equation for the steady reaction wave propagation (in front-attached coordinates, $x \to x - vt$) is

$$-vT_x = (T_x)_x + M\,\Phi(T)\,, \tag{2.4}$$

where v is the velocity of the front and $\Phi(T)$ is the rate of heat release due to the reaction. The heat released is used to heat up the reacting mixture and to preheat the fresh reagents to the reaction temperature; the terms responsible for these processes are $-\,v\,T_x$ and $(T_x)_x$, respectively. Since $1/Z \ll 1$, we can neglect the right-hand side (RHS) of (2.4) to obtain that

$$(T_x)_x + M\,\Phi(T) = 0\,,\ \text{where } T_p - \frac{1}{Z} < T < T_p\,,$$

in the main reaction zone. Then, multiplying by T_x and integrating, we have

$$\left|T_x^-\right| \approx \left(2\int_{T_p-\frac{1}{Z}}^{T_p} M\,\Phi(T)\,dT\right)^{\frac{1}{2}}\,,$$

where $|T_x^-| = vQ$ is the net flux from the main reaction zone toward the fresh mixture of reagents and Q is the (nondimensional) heat generated by the reaction per unit volume. Here we assume that the heat flux toward the (cold) mixture of reagents is significantly larger than the flux toward the products of the reaction. Further, we assume that both M and T are monotone in the main reaction zone so that we can interpret M as a function of T in the integral above. We conclude that

$$Q\,v \approx \left(2\int_{T_p-\frac{1}{Z}}^{T_p} M\,\Phi(T)\,dT\right)^{\frac{1}{2}}\,, \tag{2.5}$$

where the integral in temperature is taken over the whole temperature region where the reaction is not negligible [21]. Since the temperature zone $[T_p - 1/Z\,, T_p]$ is narrow, we approximate the Arrhenius term $\Phi(T)$ by its maximum value $\Phi(T) \approx \Phi(T_p) = Ze^{Z(T_p-1)}$. Then (cf. [8])

$$Q\,v \approx \left(2\int_{T_p-\frac{1}{Z}}^{T_p} M\,Z\,e^{Z(T_p-1)}\right)^{\frac{1}{2}}\,. \tag{2.6}$$

If we assume that $\Phi(T) = K(T)$ in (2.5), where $K(T)$ is given by (2.3), we obtain an expression of the same form as (2.6) once we set

$$A(T_p) = Z\,e^{Z(T_p-1)}\,. \tag{2.7}$$

Hence, given the similar profiles for M, the speed of front propagation is approximately the same for both Arrhenius and step-function kinetics with A given by (2.7).

Note that, when Z is large, the integral value of the kinetic function over the interval from the initial to the burnout temperature is approximately the same for both step-function kinetics and Arrhenius kinetics. Indeed,

$$\int_0^{T_p} K(T)\,dT = e^{Z(T_p-1)}\,,$$

while

$$\int_0^{T_p} Z\,e^{Z(T-1)}\,dT = e^{Z(T_p-1)} - e^{-Z}\,.$$

Now, to fix ideas, we need a rigorous definition of the burnout temperature T_p. In [1] this quantity is defined as the temperature at the end of the reaction zone on the boundary $x_b(t)$ separating the reacting mixture and the products of the reaction. Since the reaction is negligible in the products zone, it was assumed in [1] that the concentration of the reagent vanishes at $x_b(t)$ and, hence, $T_p(t) = T(x_b(t), t)$. Both T_p and x_b are unknown and are determined as a part of the solution procedure.

The spatial domain in [1] was separated into three zones: preheating zone, reaction zone, and products zone. When traveling-wave solutions are sought subject to boundary conditions at infinity, the problem within each zone consists of two second order ordinary differential equations with solutions that are "glued" together so that the temperature, the reagent concentration, and their derivatives are continuous across all interfaces.

Consider now the same model when the inverse of the Lewis number is equal to zero. Because monomer diffusion is neglected, the concentration equation becomes a first order ordinary differential equation, reducing the number of boundary conditions within each zone—either we have to consider a singularly perturbed problem in terms of the Lewis number, or we can no longer define x_b as a point at which the concentration of the monomer will vanish. Indeed, either it will cause the concentration to vanish everywhere in the domain, or $x_b \to \infty$ as $Le \to \infty$.

As has been already noted, the step-function kinetics approach has been widely applied in a number polymerization studies. Step-function kinetics was used to investigate polymerization waves for two-species [11] and four-species models [19] as well as for two-step and one-step polymerization models in the nonadiabatic case [10]. The problem of defining the temperature of the reaction cutoff has generally been circumvented by assuming that the reaction zone remains narrow. In [10], for example, it was assumed that the heat losses lead to a temperature profile with a maximum T_m within the reaction zone. The reaction is then cut off below the ignition temperature T_f, which itself depends on T_m. The value of T_m is not known in advance and has to be determined as a part of the solution procedure. Even in the simplest case, the solution to this problem is quite complicated and requires additional assumptions; in [10] it was assumed that the reaction zone had a small width of order $0 < \epsilon \ll 1$.

Subsequently, the step-function kinetics was used in conjunction with the narrow reaction zone assumption to study the stability of a uniformly propagating front in various polymerization models in the presence [20] and in the absence [16] of heat losses. The width of the front in both instances was assumed to be of order ϵ, and T_f was set to be equal to $T_p(1-\epsilon)$. Then $T_f = T_p$ in the limit $\epsilon \to 0$, thus reducing the reaction term to a jump condition on the gradient of the temperature on the interface.

Here, instead of either considering asymptotics in terms of the Lewis number or assuming that the reaction zone is narrow, we will set $T_p(t) = T(x_b(t), t)$, where $x_b(t)$ is a point at which the monomer concentration falls below a small prescribed threshold value $M(x_b(t), t) = \beta > 0$.

Given this definition of the burnout temperature, $K(T)$ is actually a functional $K[M, T]$, and the dependence of K on T is nonlocal. The methodology for choosing the parameter β will be discussed below. Essentially, β can be interpreted as the concentration of monomer in the transitional layer between the inner region adjacent to the reaction front and the outer product region. Since the temperature in this layer is approximately constant in spatial variables, the exact value of β should not have a significant influence on the predictions of our model. This is indeed the case, as will be demonstrated below. The range of reasonable values of β appears to be "robust" in the sense that it is independent of the other parameters of the problem. The magnitude of β is of order $10^{-2} \approx 1/Z^2 \ll \frac{1}{Z}$ for the parameter regimes that we have considered.

The formal justification for our choice of $x_b(t)$ is that the burnout temperature T_p is defined as the maximum temperature in the main reaction zone that is reached at the point of full monomer conversion. Due to the exponential character of the model, the concentration of monomer never vanishes, and we have to choose a threshold concentration below which we can assume that the reaction has terminated. If we set this threshold too low, then the temperature at $x_b(t)$ will be measured far away from the front and deep in the product zone. In this case, $T_p \approx const = 1$ and the dependence of the front velocity on temperature will be lost. On the other hand, if the threshold concentration of the monomer is set to be too high, then T_p will be too low and the front will fail to propagate.

Similar to, e.g., [16], we imposed the condition that $\epsilon \ll 1$. Although, typically $1/Z$ and ϵ are roughly of the same order of magnitude—$1/Z\epsilon \approx 3$ in this paper—*the width of the main reaction zone is not determined by* ϵ alone and does not have to be small.

Given our choice of $K(T)$ in (2.3), the step-function approximation of the model (2.1)–(2.2) is

$$\frac{\partial M}{\partial t} = -MK(T)\,, \tag{2.8}$$

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} + MK(T)\,. \tag{2.9}$$

Since we assume that the reaction begins when the temperature reaches the threshold value of $T_p - \frac{1}{Z}$, we will associate with this temperature the position $\psi(t)$ of the *reaction front* by defining $\psi(t)$ implicitly through the relation

$$T(\psi(t), t) = T_p(t) - \frac{1}{Z}\,.$$

In order to obtain traveling-wave solutions and study their stability, we introduce the front-attached spatial coordinate $y = x - \psi(t)$. Then the system of equations (2.8)–(2.9) can be written as

$$\frac{\partial M}{\partial t} - \psi'(t)\frac{\partial M}{\partial y} = -ZMe^{Z(T_p-1)}\eta(y)\,, \tag{2.10}$$

$$\frac{\partial T}{\partial t} - \psi'(t)\frac{\partial T}{\partial y} = \frac{\partial^2 T}{\partial y^2} + ZMe^{Z(T_p-1)}\eta(y)\,, \tag{2.11}$$

where

$$\eta(y) = \begin{cases} 0\,, & \text{if } y < 0\,, \\ 1\,, & \text{if } y \geq 0\,, \end{cases}$$

is the Heaviside function.

We will assume that $y \in \mathbf{R}$ and that the following boundary conditions at infinity (cf. dimensional boundary conditions (1.5)–(1.7) on a finite domain) are satisfied:

$$T(-\infty, t) = 0\,, \qquad T_y(\infty, t) = 0\,, \tag{2.12}$$

$$M(-\infty, t) = 1\,, \qquad M(\infty, t) = 0\,. \tag{2.13}$$

These must be supplemented by the conditions on M and T when $y = 0$. By the definition of $\psi(t)$, we immediately have that

$$T(0, t) = T_p(t) - \frac{1}{Z}\,. \tag{2.14}$$

We will require that both monomer concentration and the derivative of the temperature are continuous across the polymerization front, that is,

$$[T_y]_{y=0} = [M]_{y=0} = 0\,, \tag{2.15}$$

where $[f]_{y=a} = f(a^+) - f(a^-)$ denotes the jump of the function f at $y = a$.

Further, by the definition of T_p, an additional condition

$$T\left(x_b(t)\,, t\right) = T_p\,, \quad \text{where } M\left(x_b(t), t\right) = \beta\,, \tag{2.16}$$

must be satisfied. Here we assume that the monomer concentration is a monotone function of x for all $t > 0$. In general, as has been pointed out in [3], this assumption may not be correct for a nonuniformly propagating reacting front. This complication can be easily circumvented, however, by assuming that $x_b(t)$ is the leftmost point in the reaction zone satisfying the condition $M\left(x_b(t), t\right) = \beta$.

First, we seek traveling-wave solutions of (2.10)–(2.11) propagating in the negative y-direction. We set $\psi'(t) = -v$, where v is a positive constant, and suppose that $\frac{\partial M}{\partial t} \equiv \frac{\partial T}{\partial t} \equiv 0$. Then (2.10)–(2.11) reduce to the system of the ordinary differential equations

$$v\frac{d\bar{M}}{dy} = -Z\bar{M}e^{Z(T_p-1)}\eta(y)\,, \tag{2.17}$$

$$v\frac{d\bar{T}}{dy} = \frac{d^2\bar{T}}{dy^2} + Z\bar{M}e^{Z(T_p-1)}\eta(y)\,, \tag{2.18}$$

where $\bar{M}$ and $\bar{T}$ are time-independent solutions of (2.10)–(2.11).

Fix Z and β. Denote the temperature at the reaction front, as in (2.14), by

$$T_f := \bar{T}(0) = T_p - \frac{1}{Z}\,, \tag{2.19}$$

and set

$$A := Z\,e^{Z(T_p-1)} = Z\,e^{Z(T_f-1)+1}\,, \tag{2.20}$$

to be the strength of the kinetics term. The problem (2.17)–(2.18), (2.12)–(2.16) admits the following set of solutions:

$$\bar{M}(y) = \begin{cases} 1\,, & y < 0\,, \\ e^{-\frac{A\,y}{v}}\,, & y \geq 0\,, \end{cases} \qquad \bar{T}(y) = \begin{cases} \dfrac{(1-\beta)Z-1}{(1-\beta)Z} e^{v\,y}\,, & y < 0\,, \\ 1 - \dfrac{1}{(1-\beta)Z} e^{-\frac{A\,y}{v}}\,, & y \geq 0\,, \end{cases} \tag{2.21}$$

where

$$A = Ze^{-\frac{\beta}{1-\beta}}\,, \quad v = \sqrt{\frac{Ze^{-\frac{\beta}{1-\beta}}}{Z(1-\beta)-1}}\,, \quad \bar{T}_f = 1 - \frac{1}{(1-\beta)Z}\,. \tag{2.22}$$

Furthermore, because of the constraint (2.16), the reaction zone extends from $y = 0$ to $y = \bar{y}_b$, where

$$\bar{y}_b := -\frac{v \ln \beta}{A} = -\ln \beta \sqrt{\frac{e^{\frac{\beta}{1-\beta}}}{Z\left(Z(1-\beta)-1\right)}}\,, \tag{2.23}$$

so that

$$\bar{M}\left(\bar{y}_b\right) = \beta\,, \qquad \bar{T}\left(\bar{y}_b\right) = \bar{T}_f + \frac{1}{Z}\,. \tag{2.24}$$

Our choice of β is dictated by the requirement that the reaction zone have a width of order Z^{-1}. Then $\ln \beta$ should be an order-$O(1)$ quantity, independent of Z. As we demonstrate below, choosing β in this range leads to solution behavior that closely resembles what is observed for the Arrhenius kinetics and its point-source approximation.

2.3. Stability analysis. Next, we consider the following perturbations of the base state (2.21)–(2.23):

$$\begin{aligned} M(y\,,t) &= \bar{M}(y) + \delta\, e^{\omega\, t} \mu(y)\,, \\ T(y\,,t) &= \bar{T}(y) + \delta\, e^{\omega\, t} \tau(y)\,, \\ T_f(t) &= \bar{T}_f + \delta\, e^{\omega\, t} \xi\,, \\ \psi(t) &= -vt + \delta\, e^{\omega\, t}\,, \\ y_b(t) &= \bar{y}_b + \delta\, e^{\omega\, t} \zeta\,, \end{aligned}$$

where $0 < \delta \ll 1$ and $\omega \in \mathbf{C}$.

First, we linearize the condition (2.16). To the first order in δ we have

$$\begin{aligned} \beta = M\left(y_b(t)\,,t\right) &= \bar{M}\left(y_b(t)\right) + \delta\, e^{\omega\, t} \mu\left(y_b(t)\right) \\ &\sim \bar{M}\left(\bar{y}_b\right) + \delta\, e^{\omega\, t} \left(\bar{M}'(\bar{y}_b)\zeta + \mu\left(\bar{y}_b\right)\right) \end{aligned}$$

and

$$\begin{aligned} T_f(t) = T\left(y_b(t)\,,t\right) &= \bar{T}\left(y_b(t)\right) + \delta\, e^{\omega\, t} \tau\left(y_b(t)\right) \\ &\sim \bar{T}\left(\bar{y}_b\right) + \delta\, e^{\omega\, t} \left(\bar{T}'(\bar{y}_b)\zeta + \tau\left(\bar{y}_b\right)\right). \end{aligned}$$

Then, using the perturbation of $T_f(t)$ and (2.23), we obtain

$$\bar{M}'(\bar{y}_b)\zeta + \mu\,(\bar{y}_b) = 0\,, \qquad \bar{T}'(\bar{y}_b)\zeta + \tau\,(\bar{y}_b) = \xi\,. \tag{2.25}$$

It follows from (2.21), (2.24), and (2.25) that

$$\xi = \tau\,(\bar{y}_b) - \frac{\bar{T}'(\bar{y}_b)}{\bar{M}'(\bar{y}_b)}\mu\,(\bar{y}_b) = \tau\,(\bar{y}_b) + \frac{1}{(1-\beta)Z}\mu\,(\bar{y}_b)\,. \tag{2.26}$$

Linearization of (2.10)–(2.15) yields the following problems:

$$\begin{cases} v\mu' + \omega\mu = \omega\bar{M}'\,, & y < 0\,, \\ v\mu' + (\omega + Z)\,\mu = \omega\bar{M}' - Z^2\xi\bar{M}\,, & y \geq 0\,, \end{cases} \tag{2.27}$$

and

$$\begin{cases} \tau'' - v\tau' - \omega\tau = -\omega\bar{T}'\,, & y < 0\,, \\ \tau'' - v\tau' - \omega\tau = -\omega\bar{T}' - Z\left(\mu + Z\xi\bar{M}\right)\,, & y \geq 0\,, \end{cases} \tag{2.28}$$

subject to the boundary conditions

$$\tau(-\infty) = \tau'(\infty) = [\tau']_{y=0} = 0\,, \qquad \tau(0) = \xi\,, \tag{2.29}$$

$$\mu(-\infty) = \mu(\infty) = [\mu]_{y=0} = 0\,, \tag{2.30}$$

and (2.26).

As has already been pointed out, $\beta \ll \frac{1}{Z}$; however, the value of $\ln\beta$ that enters into $\bar{y}_b$ is large. We simplify the computations by keeping $\ln\beta$ as a parameter in our calculations, and otherwise restrict our analysis to the $O(1)$-approximation of (2.26)–(2.30) in β by setting $\beta = 0$. Then the base state can be written as

$$\bar{M}(y) = \begin{cases} 1\,, & y < 0\,, \\ e^{-\frac{Z\,y}{v}}\,, & y \geq 0\,, \end{cases} \qquad \bar{T}(y) = \begin{cases} \dfrac{1}{v^2}e^{v\,y}\,, & y < 0\,, \\ 1 - \dfrac{1}{Z}e^{-\frac{Z\,y}{v}}\,, & y \geq 0\,, \end{cases} \tag{2.31}$$

where

$$v = \sqrt{\frac{Z}{Z-1}}\,, \quad \bar{T}_f = 1 - \frac{1}{Z}\,, \quad \bar{y}_b = -\frac{\ln\beta}{\sqrt{Z(Z-1)}}\,. \tag{2.32}$$

The solution to (2.26)–(2.30) and the dispersion relation satisfied by the parameters Z and ω is very complicated, in particular due to the coupling (2.26), and was obtained using the Maple computer algebra system. In order to find the stability boundary, the real part of ω was set equal to zero, $\omega = i\phi$, and the resulting system of equations was solved in Maple for Z and ϕ.

The dependence on $\ln\beta$ of the critical value of the Zeldovich number Z and the dimensional period of front velocity oscillations $\lambda = \frac{2\pi Z}{k\phi}$ at the critical Z are shown in Figure 2.1. Note that the stability boundary is in remarkable agreement with the sharp-front stability boundary obtained in [12] ($Z = 2\left(2 + \sqrt{5}\right) \approx 8.47$) when the reaction kinetics is approximated by the point-source on the front. The same critical value of the Zeldovich number Z as in [12] has been obtained by using the sharp-front limit $Z \to \infty$ of the model with step-function kinetics. This result may be related to the fact that the reductions of Arrhenius kinetics to point-source kinetics and to

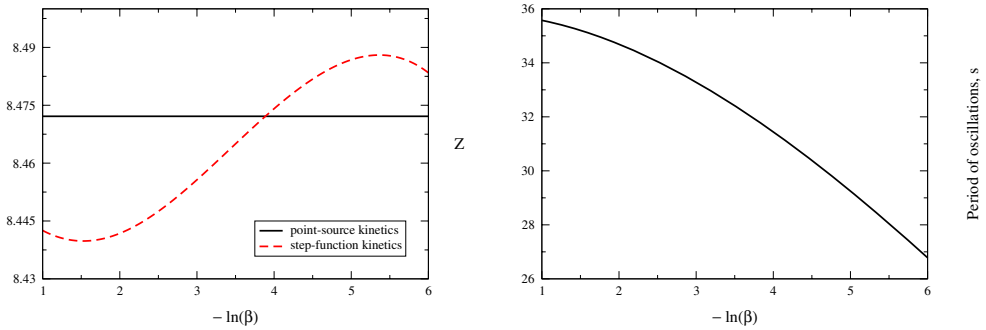


FIG. 2.1. *Dependence of the critical value of the Zeldovich number Z and the dimensional period of velocity oscillations λ on β.*

step-function kinetics are based on similar arguments that enforce the same speed of front propagation for both approximate kinetics and Arrhenius kinetics.

Analysis analogous to that in [16] leads to the same value of the stability boundary as in point-source kinetics, once the jump condition for the balance of heat on the front is imposed in place of the similar one-sided condition employed in [16]. The latter condition was derived on the basis of generalized matched asymptotic expansions [15] and leads to a stability threshold of $Z = 6$. This value disagrees with predictions of other models.

The value of the artificial parameter β can be "tuned" so that the stability threshold coincides with that obtained in [12]. Further, the dependence of the period of oscillations on β is much stronger than that for the critical Zeldovich number.

Although the stability boundaries for the approximate models discussed here are essentially the same ($Z \approx 8.47$), the stability boundary for the model with full Arrhenius kinetics [17] and small ϵ is slightly higher ($Z \approx 9.1$).

2.4. Numerical results. To verify these conclusions, we conducted numerical experiments with the model (2.8)–(2.9), with the Arrhenius kinetic function replaced by the step function.

The system of equations was solved numerically using a finite difference method with semi-implicit time integration. The physical model has no-flow homogeneous Neumann boundary conditions at both ends of the domain. For some parameter combinations, however, we found it necessary to apply Dirichlet boundary condition $T(L, t) = T_b$, where T_b is defined in (1.10), at the ignition end of the domain for a short period of time to initiate the reaction, and then switch to the homogeneous Neumann boundary condition. Numerical experiments have demonstrated that the long-term behavior of the reaction-diffusion equation system studied in this paper is not affected by the application of the Dirichlet boundary condition during the initiation stage.

At each time step, the location of the reaction front was defined as the first grid point, going from left to right, at which the concentration of the monomer drops below 50% of the initial value. The average velocity of the front was calculated by

$$v = a\frac{\Delta x}{\Delta t},$$

where Δx is the distance between grid points, Δt is the size of the time step, and

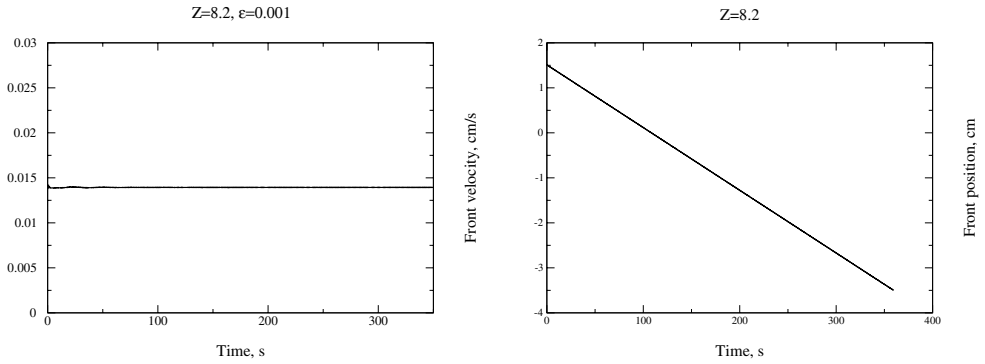


FIG. 2.2. *Reaction front velocity and position when* $Z = 8.2$ *and* $\epsilon = 10^{-3}$.

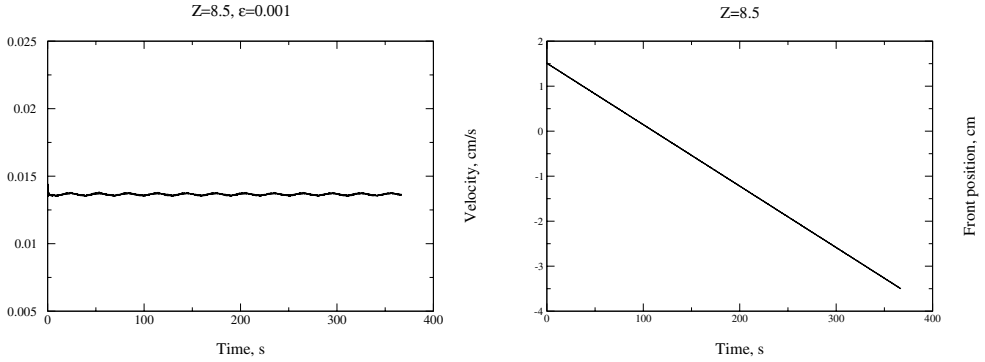


FIG. 2.3. *Reaction front velocity and position when* $Z = 8.5$ *and* $\epsilon = 10^{-3}$.

a is the number of grid intervals that the front travels through in Δt seconds. Note that it may take multiple, say m, time steps for the front to travel through one grid interval. In that case, we have $a = \frac{1}{m}$.

Since we do not use an adaptive scheme, we used uniform grid refinement technique, which clearly indicated numerical convergence and demonstrated that all sharp features are resolved and grid-independent.

We will assume that the parameters

$$q = 33.24 \text{ KL/mol}\,, \quad \kappa = 0.0014 \text{ cm}^2/\text{s}\,, \quad k = 1 \text{ s}^{-1}\,, \quad T_b = 500 \text{ K}$$

are fixed; then the state of the system is completely determined once the values of Z and ϵ are specified. The length of the spatial domain (test tube) in our computations varies from 6 cm to 10 cm, depending on the characteristic time scale of the process of interest.

Using this choice of system parameters, we varied Z while keeping $\epsilon = 1 \cdot \text{E-3}$ and $\beta = 2 \cdot \text{E-2}$ fixed. Since $\ln 0.02 \approx -3.92$, the analytical stability threshold is almost the same (Figure 2.1) as the one predicted by the stability analysis for the point-source kinetics [12].

Our simulations predict that the stability threshold is $Z \approx 8.5$. The typical velocity and the front position profiles are presented in Figures 2.2–2.5. The front

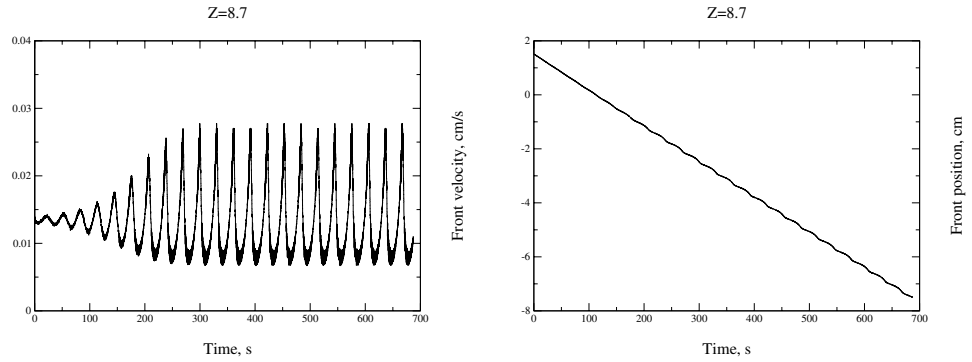


FIG. 2.4. *Reaction front velocity and position when* $Z = 8.7$ *and* $\epsilon = 10^{-3}$.

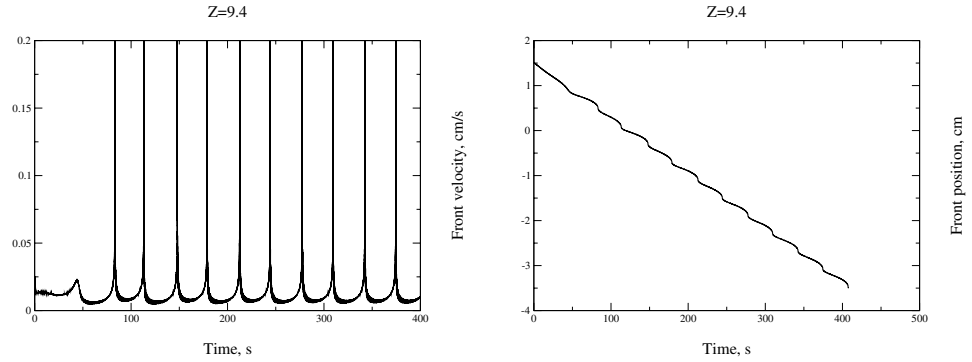


FIG. 2.5. *Reaction front velocity and position when* $Z = 9.4$ *and* $\epsilon = 10^{-3}$.

propagates with the constant velocity when $Z = 8.2$; the velocity oscillations appear when $Z = 8.5$ and become more pronounced once Z is increased ($Z = 8.7$). Further increase in Zeldovich number shows that the behavior of the system is similar to the behavior of a system reacting via Arrhenius kinetics—for instance, the "period doubling" can be observed for $Z = 9.4$.

Another similarity with Arrhenius kinetics was demonstrated in [3], where we showed that pulsating fronts in systems governed both by Arrhenius kinetics and by the step-function kinetics evolve via a combination of bulk and frontal modes. There we showed that, in fact, combustion can be considered to be purely frontal only for uniformly propagating reaction waves in systems governed by kinetics that ignore low-temperature bulk reactions. That is, for Zeldovich numbers beyond the first critical threshold (pulsating mode) there *always* exists a nonfrontal component of the dynamics. In particular, we showed that the monomer profile periodically becomes nonmonotone in regions experiencing high front acceleration, and pockets of unreacted monomer form behind the rapidly advancing front. These pockets later disappear via bulk polymerization. Note that both bulk contribution and the related solution features are *always absent* from the sharp-front–based models, since the reaction is limited to the front.

The numerically determined period of velocity oscillations for $Z = 8.5$ near the threshold of instability is approximately $\lambda \approx 31$ s. This value is almost identical

to the value obtained through the stability analysis (Figure 2.1). We conclude that the predictions of the stability analysis are in very close agreement with numerical simulations.

3. Conclusions. In the context of FP, we introduced a precise definition of distributed step-function kinetics without resorting to a sharp-front approximation. This kinetics is appropriate for simulating the behavior of one-dimensional large-Lewis-number reaction systems governed by Arrhenius kinetics. Among the interesting features of the distributed step-function kinetics is the numerical and analytical tractability of the corresponding model that takes into account possible bulk reactions behind the advancing front.

We demonstrated numerically that dynamics of fronts in systems modeled with distributed step-function kinetics and in systems modeled with Arrhenius kinetics are qualitatively the same for the time scales at which bulk reactions ahead of the front can be neglected. Further, we showed that the stability threshold of the traveling wave solution for the step-function kinetics is in excellent agreement with its numerically determined value as well as with other existing kinetics approximations.

Acknowledgments. The author would like to express his gratitude to L. K. Gross, V. A. Volpert, and J. Zhu for valuable discussions.

REFERENCES

[1] A. P. Aldushin and S. G. Kasparyan, *Thermodiffusional Instability of a Stationary Flame Wave*, Technical report, Institute of Chemical Physics, Chernogolovka, Russia, 1978.

[2] A. P. Aldushin and S. G. Kasparyan, *Thermodiffusional instability of a combustion front*, Sov. Phys. Dokl., 24 (1979), pp. 29–31.

[3] S. A. Cardarelli, D. Golovaty, L. K. Gross, V. T. Gyrya, and J. Zhu, *A numerical study of one-step models of polymerization: Frontal versus bulk mode*, Phys. D, 206 (2005), pp. 145–165.

[4] K. M. Chechilo, R. Y. A. Khvilivitskii, and N. S. Enikolopyan, *Phenomenon of polymerisation reaction spreading*, Dokl. Akad. Nauk SSSR, 205 (1972), pp. 1180–1181.

[5] Y. Chekanov, D. Arrington, G. Brust, and J. A. Pojman, *Frontal curing of epoxy resins: Comparison of mechanical and thermal properties to batch-cured materials*, J. Appl. Polymer Sci., 66 (1997), pp. 1209–1216.

[6] Y. Choi, J. K. Lee, and M. E. Mullins, *Densification process of TiC_x-Ni composites formed by self-propagating high-temperature synthesis reaction*, J. Materials Sci., 32 (1997), pp. 1717–1724.

[7] P. Dimitriou, J. Puszynski, and V. Hlavacek, *On the dynamics of equations describing gasless combustion in condensed systems*, Combust. Sci. Tech., 68 (1989), pp. 101–111.

[8] D. A. Frank-Kamenetskii, *Diffusion and Heat Exchange in Chemical Kinetics*, Princeton University Press, Princeton, NJ, 1955.

[9] M. Frankel, V. Roytburd, and G. Sivashinsky, *Complex dynamics generated by a sharp interface model of self-propagating high-temperature synthesis*, Combust. Theory Model., 2 (1998), pp. 1–18.

[10] P. M. Goldfeder and V. A. Volpert, *Nonadiabatic frontal polymerization*, J. Engrg. Math., 34 (1998), pp. 301–318.

[11] P. M. Goldfeder, V. A. Volpert, V. M. Ilyashenko, A. M. Khan, J. A. Pojman, and S. E. Solovyov, *Mathematical modeling of free-radical polymerization fronts*, J. Phys. Chem., 101 (1997), pp. 3474–3482.

[12] B. J. Matkowsky and G. Sivashinsky, *Propagation of a pulsating front in solid fuel combustion*, SIAM J. Appl. Math., 35 (1978), pp. 465–478.

[13] A. G. Merzhanov, A. K. Filonenko, and I. P. Borovinskaya, *New phenomena in combustion of condensed systems*, Dokl. Akad. Nauk SSSR, 208 (1973), pp. 122–125.

[14] A. G. Merzhanov and B. I. Khaikin, *Theory of combustion waves in homogeneous media*, Proc. Energy Combust. Sci., 14 (1988), pp. 1–98.

[15] D. A. Schult, *Matched asymptotic expansions and the closure problem for combustion waves*, SIAM J. Appl. Math., 60 (1999), pp. 136–155.

[16] D. A. Schult and V. A. Volpert, *Linear stability analysis of thermal free radical polymerization waves*, Int. J. Self-Propagating High-Temp. Synthesis, 8 (1999), pp. 417–440.

[17] K. G. Shkadinskii, B. I. Khaikin, and A. G. Merzhanov, *Propagation of a pulsating exothermic reaction front in the condensed phase*, Fizika Goreniya i Vzryva, 1 (1971), pp. 19–28 (English translation in Combust. Expl. Shock Waves, 7 (1971), pp. 15–22).

[18] S. E. Solovyov, V. M. Ilyashenko, and J. A. Pojman, *Numerical modeling of self-propagating polymerization fronts: The role of kinetics on front stability*, Chaos, 7 (1997), pp. 331–340.

[19] C. A. Spade and V. A. Volpert, *On the steady-state approximation in thermal free radical frontal polymerization*, Chem. Engrg. Sci., 55 (2000), pp. 641–654.

[20] C. A. Spade and V. A. Volpert, *Linear stability analysis of no-adiabatic free-radical polymerization waves*, Combust. Theory Model., 5 (2001), pp. 21–39.

[21] Ya. B. Zel'dovich, *On theory of burning of the gunpowder and explosives*, J. Exper. Theoret. Phys., 12 (1942), pp. 498–524 (in Russian).

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 810–836

STATICS OF POINT JOSEPHSON JUNCTIONS IN A MICROSTRIP LINE*

J. G. CAPUTO†‡ AND L. LOUKITCH†

Abstract. We model the static behavior of point Josephson junctions in a microstrip line using a one-dimensional linear differential equation with delta distributed sine nonlinearities. We analyze the maximum current γ_{max} crossing the microstrip for a given magnetic field H. In particular, we establish its periodicity and analyze how it is affected by the geometry, length, type of current feed, position, and area of the junctions. For the common situation of small currents, we show that γ_{max} can be obtained by a simple formula, the magnetic approximation. This model is in excellent agreement with measurements obtained for real devices.

Key words. Josephson junctions, sine Gordon equation, Dirac delta function, optimization

AMS subject classifications. 35Qxx, 35Jxx, 46Fxx

DOI. 10.1137/060664811

1. Introduction.

The coupling of two low T_c superconductors across a thin oxide layer is described by the Josephson equations [13]

$$V = \Phi_0 \frac{d\phi}{dt}, \quad I = sJ_c \sin(\phi), \tag{1.1}$$

where V and I are, respectively, the voltage and current across the barrier; s is the contact surface; J_c is the critical current density; and $\Phi_0 = \hbar/2e$ is the reduced quantum flux. These two Josephson relations together with Maxwell's equations imply the modulation of DC current by an external magnetic field in the static regime and the conversion of AC current into microwave radiation [1, 14]. Other applications include rapid single flux quantum logic electronics [14] and microwave signal mixers used in integrated receivers for radio-astronomy [17, 4]. In all these systems there is a characteristic length which reduces to the Josephson length, λ_J, the ratio of the electromagnetic flux to the quantum flux Φ_0 for standard junctions.

For many applications and in order to protect the junction, Josephson junctions are embedded in a so-called microstrip line which is the capacitor made by the overlap of the two superconducting layers. This is the "window geometry" where the phase difference between the top and bottom layer satisfies an inhomogeneous two-dimensional (2D) damped driven sine Gordon equation [9] resulting from Maxwell's equations and the Josephson constitutive relations (1.1). The damping is due to the normal electrons, and the driving is through the boundary conditions with an external current or magnetic field applied to the device.

Even in the static regime ($V \equiv 0$) the 2D problem is complicated because of the multiplicity of solutions due to the sine term. However, flux penetration occurs along the direction of the magnetic field, so one direction dominates the other. A quantity measured by experimentalists is the maximum (static) current $I_{\max}(H)$, which can cross the device for a given magnetic field H. This gives information

*Received by the editors July 11, 2006; accepted for publication (in revised form) November 1, 2006; published electronically March 22, 2007.
http://www.siam.org/journals/siap/67-3/66481.html

†Laboratoire de Mathématiques, Institut de Sciences Appliquees, B.P. 8, 76131 Mont-Saint-Aignan Cedex, France (caputo@insa-rouen.fr, loukitch@insa-rouen.fr).
‡Laboratoire de Physique theorique et modelisation, Université de Cergy-Pontoise and C.N.R.S.

on the quality of the junctions. An important issue is how defects in the coupling will affect this maximum current. In particular, high T_c superconductors can be described as Josephson junctions where the critical current density is a rapidly varying function of the position, due to grain boundaries. Fehrenbacher, Geshkenbein, and Blatter [10] calculated $I_{\max}(H)$ for such disordered long Josephson junctions and for a periodic array of defects. Experiments were also done by Itzler and Tinkham on large 2D disordered junctions [11, 12]. However, the overall picture is complex, and it is difficult to obtain geometric information on the junction from the curve $I_{\max}(H)$. The analysis of such a 2D problem [8] provided bounds on the gradient of the solution that were independent of the area of the junctions, so that little information could be obtained on $I_{\max}(H)$. However, the study [8] proved the existence of solutions and the convergence of the Picard iteration to obtain them.

Small junctions of length $w_i < \lambda_J$ are easier to study and lead to the well-known $I_{\max}(H) = \sin(Hw_i)/H$ [1]. Two such junctions are commonly associated to form a superconducting quantum interference device (SQUID), now routinely used to measure magnetic fields. More junctions can be used to form arrays [18] that can bear more critical current and are more flexible than a long junction because the area of the junction components and their position can be varied. When the junctions are closer than λ_J, such arrays behave as a long junction and could be used as microwave generators. Almost all models are discrete lumped models where the effect of the space between junctions is neglected. In particular, the interaction of the junctions through this passive region has always been neglected. This makes it difficult to describe junctions of different areas, placed nonuniformly in the microstrip. This is why up to now mostly equidistant and identical junctions have been considered. To overcome these difficulties we recently introduced a continuous/discrete model that preserves the continuity of the phase and its normal gradient across the junction interface and where the phase is assumed constant in the junctions. The 1D dynamics [6] of one junction in a cavity revealed that the junction could stop waves across the cavity or enhance them throughout. In [5] this model was used to calculate $I_{\max}(H)$ for a misaligned SQUID in a 2D cavity. Nonuniform arrays of junctions that are generalized SQUIDs have been produced and analyzed in particular by Salez and coworkers at the Observatory of Paris [17, 4], and our analysis is in excellent agreement with the measured $I_{\max}(H)$.

In this article we will concentrate on the 1D static problem. We justify our continuous/discrete model and show that it allows an in-depth analysis that was out of reach in the general 2D case. In particular, we will show the properties of $I_{\max}(H)$, its periodicity, its regularity, the relation between different types of current feeds, and how it is affected by the position of the array in the microstrip. In addition, we introduce and justify the "magnetic approximation" in which many details of $I_{\max}(H)$ can be controlled. Specifically in section 2 we introduce our model, and we give preliminary analytical results in section 3. Section 4 details the intrinsic properties of the maximal current as a function of the magnetic field: its periodicity, the relation between the inline and overlap current feed, and the simple magnetic approximation. Section 5 introduces two numerical ways to solve the problem. In section 6, we study a SQUID and examine the effect of a little difference between the junction parameters, and we compare this to the experiment. Section 7 deals with devices with more junctions; there we analyze the effect of separating one junction from the others and show the agreement with the experimental results. After concluding in section 8, we give details of the proofs in section 9.

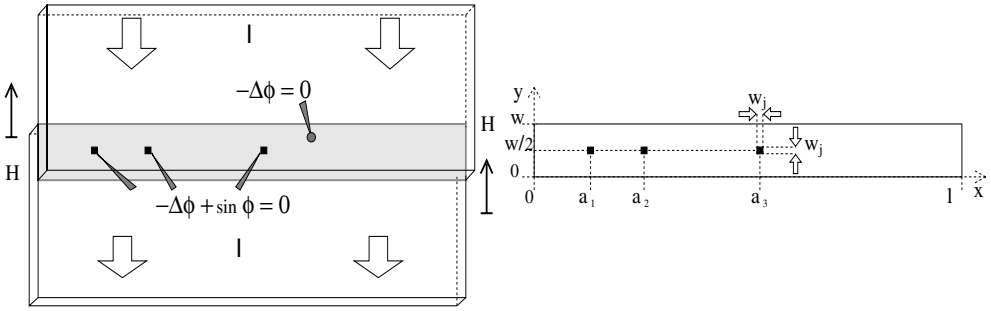


FIG. 2.1. *The left panel shows the top view of a superconducting microstrip line containing three Josephson junctions; H, I, and ϕ are respectively the applied magnetic field, current, and phase difference between the two superconducting layers. Here the current feed is of the overlap type. The phase difference ϕ between the two superconducting layers satisfies $-\Delta\phi = 0$ in the linear part and $-\Delta\phi + \sin(\phi) = 0$ in the Josephson junctions. The right panel shows the associated $2D$ domain of size $l \times w$ containing $n = 3$ junctions placed at the positions $y = w/2$ and $x = a_i$, $i = 1, n$.*

2. The model. The device we model, shown in Figure 2.1, is a so-called microstrip cavity (the grey area in Figure 2.1) between two superconducting layers. Inside this microstrip there are regions where the oxide layer is very thin (~ 10 Angstrom), enabling Josephson coupling between the top and bottom superconductors. The dimensions of the microstrip are about 100 μm in length and 20 μm in width. The phase difference between the top and bottom superconducting layers obeys in the static regime the following semilinear elliptic partial differential equation [9]:

$$-\Delta\varphi + g(x,y)\sin\varphi = 0, \tag{2.1}$$

where $g(x,y)$ is 1 in the Josephson junctions and 0 outside. This formulation guarantees the continuity of the normal gradient of φ, the electrical current on the junction interface. The unit of space is the Josephson length λ_J, the ratio of the flux formed with the critical current density and the surface inductance to the flux quantum Φ_0.

The boundary conditions representing an external current input I or an applied magnetic field H (along the y axis) are

$$\begin{aligned} \left.\frac{\partial\varphi}{\partial y}\right|_{y=0} &= -\frac{I}{2l}\nu, \qquad \left.\frac{\partial\varphi}{\partial y}\right|_{y=w} = \frac{I}{2l}\nu, \\ \left.\frac{\partial\varphi}{\partial x}\right|_{x=0} &= H - \frac{I}{2w}(1-\nu), \quad \left.\frac{\partial\varphi}{\partial x}\right|_{x=l} = H + \frac{I}{2w}(1-\nu), \end{aligned} \tag{2.2}$$

where $0 \le \nu \le 1$ gives the type of current feed. The case $\nu = 1$ shown in Figure 2.1, where the current is applied only to the long boundaries $y = 0, w$, is called overlap feed, while $\nu = 0$ corresponds to the inline feed.

We consider long and narrow strips containing a few small junctions of size $w_j \times w_j$ placed on the line $y = w/2$ and centered on $x = a_i$, $i = 1, n$, as shown in Figure 2.1. We then search for φ in the form

$$\varphi(x,y) = \frac{\nu I}{2L}\left(y - \frac{\omega}{2}\right)^2 + \sum_{n=0}^{+\infty} \phi_n(x)\cos\left(\frac{n\pi y}{w}\right), \tag{2.3}$$

where the first term takes care of the y boundary condition. For narrow strips $w < \pi$, only the first transverse mode needs to be taken into account [7, 2] because the

curvature of φ due to current remains small. Inserting (2.3) into (2.1) and projecting on the zero mode, we obtain the following equation for ϕ_0, where the 0's have been dropped for simplicity:

$$-\phi'' + g\left(x, \frac{w}{2}\right)\frac{w_j}{w}\sin\phi = \nu\frac{\gamma}{l}, \tag{2.4}$$

where $\gamma = I/w$ and the boundary conditions are $\phi'(0) = H - (1-\nu)\gamma/2$ and $\phi'(l) = H + (1-\nu)\gamma/2$.

As the area of the junction is reduced, the total Josephson current is reduced and tends to zero. To describe small junctions where the phase variation can be neglected but that can carry a significant current, we introduce the following function g_h,

$$g_h(x) = \frac{w_j}{2h} \quad \text{for } a_i - h < x < a_i + h, \quad g_h(x) = 0 \quad \text{elsewhere,} \tag{2.5}$$

where $i = 1, \dots n$. In the limit $h \to 0$ we obtain our final delta function model [6],

$$-\phi'' + \sum_{i=1}^{n} d_i\delta(x - a_i)\sin\phi = \nu j, \tag{2.6}$$

where

$$d_i = \frac{w_j^2}{w}, \quad j = \frac{\gamma}{l}, \tag{2.7}$$

and the boundary conditions are

$$\phi'(0) = H - \frac{(1-\nu)\gamma}{2}, \quad \phi'(l) = H + \frac{(1-\nu)\gamma}{2}. \tag{2.8}$$

This is our continuous/discrete 1D model of a parallel array of many point Josephson junctions embedded in a microstrip cavity. It preserves the spatial degrees of freedom in the linear cavity and the matching conditions at the junction interfaces.

3. General properties. The delta function seems to be a theoretical way to approach the problem. Nevertheless we will show that it provides an excellent agreement with experiments, in addition to allowing analytical solutions. We have the following properties:

1. Integrating twice, (2.6) shows that the solution ϕ is continuous at the junctions $x = a_i$, $i = 1, \dots, n$.

2. Let ϕ be a solution of (2.6); then $\phi + 2k\pi$ is also a solution.

3. Almost everywhere, $-\phi''(x) = \nu\gamma/l$, so that outside the junctions, ϕ is a second degree polynomial by parts,

$$\phi(x) = -\frac{\nu j}{2}x^2 + B_i x + C_i \qquad \forall x \in]a_i, a_{i+1}[. \tag{3.1}$$

4. At each junction ($x = a_i$), ϕ' is not defined, but choosing $\epsilon_1 > 0$ and $\epsilon_2 > 0$, we get

$$\lim_{\epsilon_1\to 0}\lim_{\epsilon_2\to 0}\int_{a_i-\epsilon_1}^{a_i+\epsilon_2}\phi''(x)dx = \int_{a_i^-}^{a_i^+}\phi''(x)dx = [\phi'(x)]_{a_i^-}^{a_i^+}.$$

Since the phase is continuous at the junction $x = a_i$, we obtain

$$[\phi'(x)]_{a_i^-}^{a_i^+} = d_i\sin(\phi_i), \tag{3.2}$$

with $\phi_i \equiv \phi(a_i)$.

5. Integrating (2.6) over the whole domain,

$$[\phi']_0^l = \int_0^l \phi'' dx = \sum_{i=1}^{n} d_i \sin(\phi_i) - \nu\gamma,$$

and taking into account the boundary conditions, we obtain

$$\gamma = \sum_{i=1}^{n} d_i \sin(\phi_i), \tag{3.3}$$

which indicates the conservation of current. Note that the total current is equal to the sum of the jumps of ϕ'.

3.1. The solution as a piecewise polynomial. Let ϕ be a solution of (2.6) and $\phi_1 = \phi(a_1)$. From remark (3.1), ϕ is a polynomial by parts. We define $P_{i+1}(x)$ as the second degree polynomial such that $P_{i+1}(x) = \phi(x)$ for $a_i \le x \le a_{i+1}$. Using the left boundary condition, we can specify ϕ on $[0, a_1]$:

$$P_1(x) = -\frac{\nu j}{2}\left(x^2 - a_1^2\right) + \left(H - \frac{1-\nu}{2}\gamma\right)(x - a_1) + \phi_1. \tag{3.4}$$

At the junctions, (3.2) tells us that $\forall k \in \{1, \dots, n\}$,

$$P'_{k+1}(a_k) - P'_k(a_k) = d_k \sin(P_k(a_k)). \tag{3.5}$$

Considering that $\phi'' = -\nu j$ on each interval, the previous relation, and the continuity of the phase at the junction, we can give a first expression for P_{k+1},

$$P_{k+1}(x) = -\frac{\nu j}{2}(x - a_k)^2 + \left[P'_k(a_k) + d_k \sin P_k(a_k)\right](x - a_k) + P_k(a_k). \tag{3.6}$$

Notice that $P_{k+1}(x)$ depends on $P_k(x)$, ν, j, and H. The parameters ν and l are fixed by the geometry of the device. So by recurrence we see that ϕ is entirely determined by the values of ϕ_1, γ, and H.

From (3.5) we can obtain another expression for P_{k+1},

$$P_{k+1}(x) - P_k(x) = d_k \sin(P_k(a_k))(x - a_k). \tag{3.7}$$

Summing all these relations yields

$$P_{k+1}(x) = P_1(x) + \sum_{i=1}^{k} d_i \sin(P_i(a_i))(x - a_i). \tag{3.8}$$

Polynomials (3.4) and (3.6) show by construction that the constants H, j, and ϕ_1 determine completely the solution of (2.6) if one exists. In the same way, we can show that the three other constants, j, $\phi'(a_1)$, and ϕ_1, fix ϕ. From (3.8), we give an expression of ϕ:

$$\phi(x) = P_1(x) + \sum_{i=1}^{n} \mathcal{H}_{\{x \ge a_i\}} d_i \sin(\phi_i)(x - a_i),$$

where

$$\mathcal{H}_{\{x \ge a_i\}} = \begin{cases} 1, & x \ge a_i, \\ 0, & x < a_i, \end{cases}$$

is the Heaviside function.

3.2. An *n*th order transcendental system. Another way to solve (2.6) for ϕ is to write it as a coupled system of n transcendental equations. For that, we first eliminate the constant term by introducing ψ such that

$$\phi = \psi - \nu\frac{\gamma}{l}\frac{x^2}{2} \equiv \psi - f(x)$$

and obtain

$$-\psi'' + \sum_{i=1}^{n} d_i \delta(x - a_i)\sin(\psi - f(a_i)) = 0, \tag{3.9}$$

with the boundary conditions

$$\psi'(0) = H - \frac{(1-\nu)\gamma}{2}, \quad \psi'(l) = H + \frac{(1+\nu)\gamma}{2}.$$

To simplify the notation we will write $f_i \equiv f(a_i)$ and $\psi_i \equiv \psi(a_i)$. Integrating (3.9) over the intervals $[0, a_2^-]$, $[a_1^+, a_3^-]$, ..., we obtain the relations

$$\begin{aligned}
-[\psi']_0^{a_2^-} + d_1\sin(\psi_1 - f_1) &= 0,\\
-[\psi']_{a_1^+}^{a_3^-} + d_2\sin(\psi_2 - f_2) &= 0,\\
-[\psi']_{a_2^+}^{a_4^-} + d_3\sin(\psi_3 - f_3) &= 0,\\
-[\psi']_{a_3^+}^{a_5^-} + d_4\sin(\psi_4 - f_4) &= 0,\\
-[\psi']_{a_4^+}^{l} + d_5\sin(\psi_5 - f_5) &= 0,
\end{aligned} \tag{3.10}$$

where we have assumed $n = 5$ as an example. Now we can use the fact that $\psi'' = 0$ in the intervals between the junctions and the boundary conditions to obtain the final system

$$\begin{aligned}
H - (1-\nu)\frac{\gamma}{2} - \frac{\psi_2 - \psi_1}{a_2 - a_1} + d_1\sin(\psi_1 - f_1) &= 0,\\
-\frac{\psi_3 - \psi_2}{a_3 - a_2} + \frac{\psi_2 - \psi_1}{a_2 - a_1} + d_2\sin(\psi_2 - f_2) &= 0,\\
-\frac{\psi_4 - \psi_3}{a_4 - a_3} + \frac{\psi_3 - \psi_2}{a_3 - a_2} + d_3\sin(\psi_3 - f_3) &= 0,\\
-\frac{\psi_5 - \psi_4}{a_5 - a_4} + \frac{\psi_4 - \psi_3}{a_4 - a_3} + d_4\sin(\psi_4 - f_4) &= 0,\\
-H - (1+\nu)\frac{\gamma}{2} + \frac{\psi_5 - \psi_4}{a_5 - a_4} + d_5\sin(\psi_5 - f_5) &= 0.
\end{aligned} \tag{3.11}$$

We will use this formulation as well as the one in the previous subsection to establish properties of the solutions and solve the problem numerically using Newton's method.

4. General properties of $\gamma_{max}(H)$ for an n junction array. The general problem is

$$-\phi''(x) + \sum_{i=1}^{n} d_i \delta(x - a_i)\sin(\phi) = \nu j, \tag{4.1}$$

with the boundary conditions

$$\phi'(0) = H - \frac{(1-\nu)\gamma}{2}, \quad \phi'(l) = H + \frac{(1-\nu)\gamma}{2}.$$

Experimentalists measure the maximum current γ for a given magnetic field H and plot this as a curve $\gamma_{max}(H)$. To compare with real data it is therefore important to compute and analyze this quantity. In this section, we give some properties of the $\gamma_{max}(H)$ curve. In the appendix some analytical estimates on the influence of the geometry on the maximal current will be presented.

4.1. Periodicity. We introduce

$$l_j \equiv a_{j+1} - a_j,$$

the distance between two consecutive junctions. Let l_{min} be the smallest distance l_j. We define the array as harmonic if l_i is a multiple of l_{min} $\forall\, i$.

PROPOSITION 4.1 (periodicity of the device). *For a harmonic array, the $\gamma_{max}(H)$ curve is periodic with a period $2\pi/l_{min}$.*

Proof. Let ϕ be a solution of (4.1) for a current γ and a magnetic field H. We introduce $f(x) = (2\pi/l_{min})(x - a_1)$ and $\psi(x) = \phi(x) + f(x)$. So ψ verifies

$$-\psi''(x) + \sum_{i=1}^{n} \delta(x - a_i) \sin(\psi - f) = \nu j, \tag{4.2}$$

with $\psi'(0) = H + 2\pi/l_i - (1-\nu)\gamma/2$ and $\psi'(l) = H + 2\pi/l_i + (1-\nu)\gamma/2$. Since $f(a_j) = 2k\pi$ $\forall i \in \{1, \ldots, n\}$, then ψ is a solution of (4.1) for $H + H_p \equiv H + 2\pi/l_{min}$ and the same γ, and so $\gamma_{max}(H + H_p) \geq \gamma_{max}(H)$.

Conversely, by subtracting f from a solution associated with $H + H_p$ and a current γ, we obtain a solution for H and the same current γ, and so $\gamma_{max}(H + H_p) \leq \gamma_{max}(H)$. From the two inequalities we get

$$\gamma_{max}(H + H_p) = \gamma_{max}(H) \tag{4.3}$$

with $H_p = 2\pi/l_{min}$. ∎

In the nonharmonic case, if the junctions are set such that $l_j = p_j/q_j$, where p_j and q_j are integers, prime with each other, then γ_{max} is periodic with period H_p such that

$$H_p = 2\pi \frac{LCM(q_1, \ldots, q_{n-1})}{HCF(p_1, \ldots, p_{n-1})} \tag{4.4}$$

(see Figure 4.1), where LCM is the lowest common multiple and HCF the highest common factor. To prove this write $f(x) = p(x - a_1)$ and use again the previous argument. In Figure 4.1 we show the $\gamma_{max}(H)$ curve for a three-junction unit such that $l_1 = 3/2$ and $l_2 = 5/3$, so that the period of $\gamma_{max}(H)$ is $H_p = 2\pi LCM(2,3)/HCF(3,5) = 12\pi$. In the following plots we will show only one period of $\gamma_{max}(H)$.

In the general case, we have only an approximate periodicity of $\gamma_{max}(H)$, which can be estimated using (4.4). Also, real junctions have a finite size, which causes $\gamma_{max}(H) \to 0$ when $H \to +\infty$. Our model is thus valid as long as the dimensionless magnetic field H is not larger than $1/w_j$.

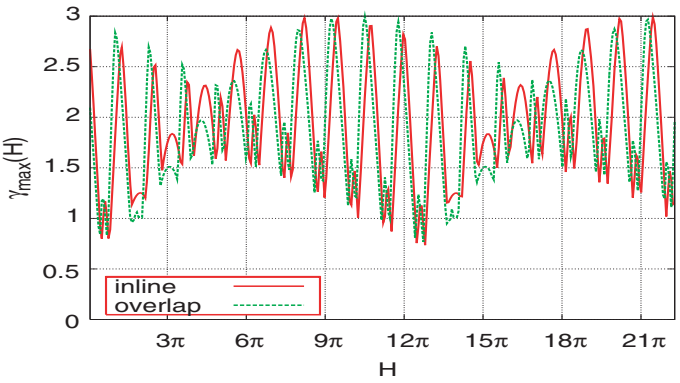


FIG. 4.1. *$\gamma_{max}(H)$ curve for an inline current feed, $\nu = 0$ (solid line), and overlap feed, $\nu = 1$ (dotted line), for a three-junction unit $\{1, 5/2, 5/2 + 5/3\}$, with $d_1 = d_2 = d_3 = 1$. Thus $l_1 = 3/2$ and $l_2 = 5/3$.*

4.2. Influence of the position of the junction unit. In this section, we examine how the position of the set of junctions in the microstrip (linear domain) will affect the $\gamma_{max}(H)$ curve. For an array of junctions placed at the distances $\{a_i, i = 1, n\}$, we define a junction unit as the set $\{l_i, i = 1, n-1\}$. Then the array where the junctions are at $\{a_1 + c, a_2 + c, \ldots, a_n + c\}$ is the same junction unit. We define a_1 as the position of the junction unit. The length of the junction unit is $l_b = a_n - a_1$. The array is centered if $(a_n + a_1)/2 = l/2$.

Inline current feed: $(\nu = 0)$. Then the boundary conditions at the edge of the junction unit are $\phi'(a_1^-) = \phi'(0) = H - \gamma/2$ and $\phi'(a_n^+) = \phi'(l) = H + \gamma/2$, independent of the position of the junction unit.

PROPOSITION 4.2 (inline junction unit). *For inline current feed, $\gamma_{max}(H)$ is independent of a_1 (the position of the junction unit) and of the length l of the cavity.*

Proof. Let $\phi_1(x)$ be a solution of (4.1) for given γ, H. Let us change the position of the junction unit to $a_1 + c$ so that the junctions are now placed at $\{a_1 + c, a_2 + c, \ldots, a_n + c\}$. It is easy to see that $\phi_2(x) = \phi_1(x - c)$ satisfies the boundary conditions and is a solution. This one-to-one map between ϕ_1 and ϕ_2 exists $\forall\, c$, H, and γ, and so the two junction units have the same $\gamma_{\max}(H)$. ∎

Then the $\gamma_{max}(H)$ curve is independent of the position of junction unit when $\nu = 0$. By the same argument, we can show that $\gamma_{max}(H)$ is independent of the length l of the circuit (see Figure 4.3). This curve depends only on the junction unit.

General current feed: $(0 < \nu \leq 1)$. In this case the boundary conditions at the edge of the junction unit are

$$\phi'(a_1^-) = -\nu j a_1 + H + (1 - \nu)\gamma/2, \quad \phi'(a_n^+) = H - (1 - \nu)\gamma/2 + \nu j(l - a_n).$$

Contrary to the inline feed, we cannot shift the phase to find a solution when the junction unit has been shifted, because now the boundary conditions depend on the position of the junction unit. Consider the derivative ϕ' at the boundaries of the junction unit. We will compare the curves $\gamma_{max}(H)$ for a centered unit and for a noncentered unit. For a centered unit, $a_n - a_1 = l/2$ so that

$$\phi'(a_1^-) - H = -\left(\phi'(a_n^+) - H\right),$$

but this equality is false for a noncentered unit. It is possible to choose a correction

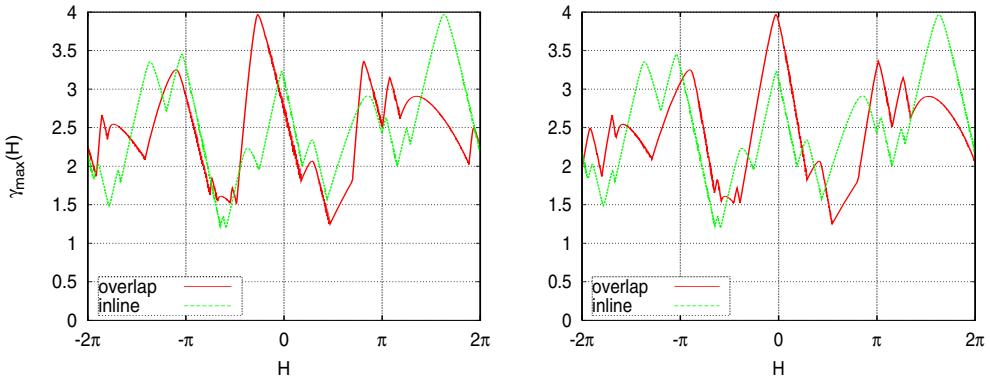


FIG. 4.2. *Plot of $\gamma_{max}(H)$ for a four-junction device ($l = 10$, $d_i = 1$) such that $l_1 = 1.5$, $l_2 = 2.5$, $l_3 = 2$. The left panel shows a noncentered junction unit with $a_1 = 0.1$, and the right panel a centered unit with $a_1 = 2$. Notice the current-dependent shift* (4.5) *for the overlap solution as one goes from a centered junction unit (right panel) to an off-centered junction unit (right panel). The junction unit was moved to the left.*

H_ν to the magnetic field H in order to obtain the equality

$$\begin{aligned} \phi'(a_1^-) - H + H_\nu &= -\left(\phi'(a_n^+) - H + H_\nu\right), \\ -\nu j a_1 + (1-\nu)\frac{\gamma}{2} + H_\nu &= -\left[\nu j(l - a_n) - (1-\nu)\frac{\gamma}{2} + H_\nu\right], \\ H_\nu &= \nu j\left(\frac{l_b - l}{2} + a_1\right). \end{aligned} \tag{4.5}$$

Let us consider two arrays, circuit 1 with a centered junction unit and circuit 2 with the same junction unit but noncentered.

PROPOSITION 4.3 (magnetic shift). *Let (H, γ_{max}) be the coordinates of a point of the $\gamma_{max}(H)$ curve for the circuit* 1. *Then $(H + H_\nu, \gamma_{max})$ is a point of the curve for the circuit* 2.

Thus, moving a junction unit translates the γ_{max} curve by $\nu j a_1$. Figure 4.2 shows a $\gamma_{max}(H)$ for a four-junction device with a noncentered junction unit in the left panel and a centered junction unit in the right panel. Both inline and overlap current feeds are presented. Notice the unchanged behavior for the inline current feed and the effect of H_ν ($= -4.1\gamma/10$) from (4.5) in the overlap case.

Proof. Let $\phi_{\{H,\gamma,a_1\}}$ be a solution for an array $A_1 \equiv \{a_1, \ldots, a_n\}$ with a centered junction unit with γ and H given. Consider another array A_2 with the same junction unit moved by s, $A_2 \equiv \{a_1 + s, \ldots, a_n + s\} \equiv A_1 + s$, the coefficients $d_1, \ldots, d_n$ being equal for the two circuits. From the solution $\phi_{\{H,\gamma,a_1\}}$ for A_1 we can deduce a solution $\psi_{\{H+H_\nu,\gamma,a_1+s\}}$ for A_2. From (4.5) we have

$$\psi'_{\{H+H_\nu,\gamma,a_1+s\}}(a_1^- + s) = \phi'_{\{H,j,a_1\}}(a_1^-).$$

Taking

$$\psi_{\{H+H_\nu,\gamma,a_1+s\}}(a_1^- + s) = \phi_{\{H,\gamma,a_1\}}(a_1^-),$$

and from the unicity of the solution, we obtain $\phi \equiv \psi$ in the two junction units. Thus, if ϕ is a solution for $\{H, \gamma\}$ given for A_1, then ψ is a solution for $\{H + H_\nu, \gamma\}$ for A_2, and vice versa.

Let $\gamma_{max,1}$ and $\gamma_{max,2}$ be the γ_{max} curves for the arrays A_1 and A_2. From the solutions obtained for A_1 we build solutions for A_2. As a consequence, $\gamma_{max,1}(H + H_\nu) \leq \gamma_{max,2}(H)$. On the other side, from solutions of A_2 we build solutions for A_1, then $\gamma_{max,1}(H + H_\nu) \geq \gamma_{max,2}(H)$. So, we obtain the equality

$$\gamma_{max,1}(H + H_\nu) = \gamma_{max,2}(H) \ . \qquad \square$$

Notice that this equality is independent of the number of junctions.

4.3. Comparison between inline and overlap current feeds. We now compare the γ_{max} curves for inline and overlap current feed. For one junction, the problem can be solved exactly using polynomials by parts (see remark (3.1)). We obtain $\gamma_{max}(H) = d_1 \ \forall \nu$. For two junctions there is the possibility of $d_1 \neq d_2$, and this will change $\gamma_{max}(H)$ qualitatively. Let us study the phase difference between two junctions. We use remark (3.1) and the boundary conditions to get

$$\begin{aligned} &\phi_2 - \phi_1 = -\frac{\nu j}{2}(a_2 - a_1)^2 + (P'(a_1) + d_1 \sin(\phi_1))\,(a_2 - a_1), \\ &\frac{\phi_2 - \phi_1}{a_2 - a_1} = -\nu j \frac{a_2 + a_1}{2} + H - \frac{1-\nu}{2}\gamma + d_1 \sin(\phi_1). \end{aligned} \tag{4.6}$$

If $(a_2 + a_1)/2 = l/2$ (the junction unit is also centered), as $\gamma = jl$, (4.6) becomes

$$\frac{\phi_2 - \phi_1}{a_2 - a_1} = H - \frac{\gamma}{2} + d_1 \sin(\phi_1). \tag{4.7}$$

Note that we can obtain (4.7) from (4.6) with $\nu = 0$. We have shown the following result.

PROPOSITION 4.4 (equivalence of all current feeds for a centered SQUID). *For a centered two-junction device, all current feeds give the same γ_{max} curve.*

For an inline current feed, $\nu = 0$, so that the phase difference $\phi_2 - \phi_1$ is independent of the position of the junction unit. This is not true for the overlap feed, where moving the junction unit causes a "magnetic shift" as seen above in H_ν equation (4.5). When the number of the junctions $n \geq 3$, the γ_{max} curve depends on ν. The effect of moving the junction unit on the γ_{max} curve was shown above. Thus, we can reduce the study to a centered junction unit. In this case, we have $a_1 = (l - l_b)/2$, and

$$\begin{aligned} \phi'(a_1^-) &= \phi'(0) + \int_0^{a_1^-} -\nu \frac{\gamma}{l} dx \\ &= H - (1-\nu)\frac{\gamma}{2} - \nu\frac{\gamma}{2} + \frac{\nu l_b}{l}\frac{\gamma}{2} = H - \left(1 - \frac{\nu l_b}{l}\right)\frac{\gamma}{2}, \\ \phi'(a_n^+) &= H + \left(1 - \frac{\nu l_b}{l}\right)\frac{\gamma}{2}, \end{aligned} \tag{4.8}$$

with $l_b = a_n - a_1$. We can write $\nu j = (\nu l_b/l)(\gamma/l_b)$ and $\nu l_b/l = \mu$. Thus, (4.1) is equivalent to the system

$$-\phi''(x) + \sum_{i=1}^{n} d_i \delta(x - a_i)\sin(\phi) = \mu \frac{\gamma}{l_b} \ , \tag{4.9}$$

with

$$\phi'(a_1^-) = H - \frac{(1-\mu)\gamma}{2}, \quad \phi'(a_n^+) = H + \frac{(1-\mu)\gamma}{2}.$$

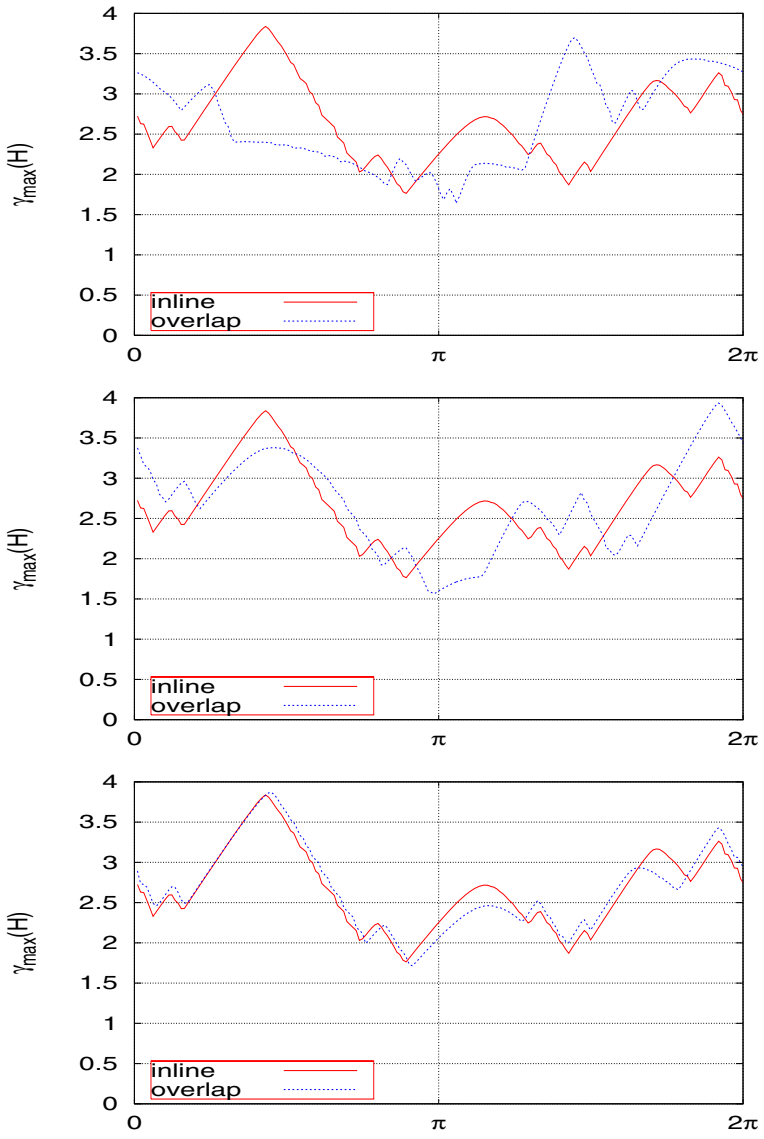


FIG. 4.3. *Plot of $\gamma_{max}(H)$ for the same centered junction unit $l_1 = 1$, $l_2 = 4$, $l_3 = 3$, $d_1 = d_2 = d_3 = d_4 = 1$ and different lengths l of the microstrip; from top to bottom, $l = 8$, 16, and 64. Notice how the overlap solution tends to the inline solution as one increases l.*

As $0 \le \nu \le 1$, $0 \le \mu \le l_b/l$. Note that l_b can be considered as the reference length of the device. Also note that if $l \to +\infty$, then $\mu \to 0$ and (4.9) and boundary conditions tend to the situation of inline current feed. Figure 4.3 illustrates this convergence when we increase the microstrip length l for a centered junction unit. Notice that the solution for the inline feed is not modified by the variation of length. With $l = 8$, we have a maximum difference between the solutions for the overlap and inline current feeds. As l increases, the solution for the overlap current feed tends to the solution for the inline current feed. We prove this in section 9.3.

Conclusion. In the appendix, we show that when $\nu\gamma/l$ tends to 0 the solution

tends to the one for inline current feed. We can get this by increasing the length l or shrinking the junction area. (See section 9.3.) We have three parameters: ν, l, and γ_{max}. l is determined by the circuit. ν comes from the 2D model, and depends on the width of the circuit. The third parameter can be bounded from above: $0 \leq \gamma \leq \sum_i d_i$. We will see in the next section what the limit of $\gamma_{max}(H)$ is for inline and overlap feeds when d_i are small.

4.4. The relation between inline and magnetic approximation. The size of the junctions $w_i < 1 < w$ so that $d_i \ll 1$; therefore the jump of the gradient of the phase across the junctions can be neglected. This is the magnetic approximation where only H fixes the phase gradient. In the previous section, we have shown that the solutions for inline and overlap current feeds converge to the same $\gamma_{max}(H)$ curve for small d_i. We will show that this limit is the magnetic approximation.

Since $[\phi']_{a_i^-}^{a_i^+} = d_i \sin(\phi_i)$ (remark (3.2)) and $j \leq \sum_i d_i/l$, then for small d_i, ϕ tends to the linear function $\phi(x) = Hx + c$. This magnetic approximation seems crude, but we show that it approaches the solution for inline feed; see section 9.4. There we bound the difference between the γ curves for the inline feed and the magnetic approximation. Figure 4.4 illustrates this convergence as d_i decreases.

This approximation gives very good results because we work on very small junctions and the corresponding $d_i \approx 10^{-2}$ (compared with the values taken in Figure 4.4).

4.5. Magnetic approximation. The magnetic approximation is very interesting because it gives an analytic expression of $\gamma_{max}(H)$ and is independent of the value of the current and of the scale of the circuit. Here we consider that $\phi(x) = Hx + c$, and from (3.3),

$$\gamma = \sum_{i=1}^{n} d_i \sin(Ha_i + c) \ .$$

Notice that c is the only parameter which can be adjusted to reach the maximum.

To find the $\gamma_{max}(H)$ curve of the magnetic approximation, we take the derivative

$$\frac{\partial \gamma}{\partial c} = -\sin(c) \left(\sum_{i=1}^{n} d_i \sin(Ha_i) \right) + \cos(c) \left(\sum_{i=1}^{n} d_i \cos(Ha_i) \right) . \tag{4.10}$$

The values of c canceling $\partial\gamma/\partial c$ are

$$c_{max}(H) = \arctan \left(\frac{\sum_{i=1}^{n} d_i \cos(Ha_i)}{\sum_{i=1}^{n} d_i \sin(Ha_i)} \right) , \tag{4.11}$$

and using (4.10), we have the solution:

$$\gamma_{max}(H) = \left| \sum_{i=1}^{n} d_i \sin(Ha_i + c_{max}(H)) \right| . \tag{4.12}$$

This γ_{max} curve is a function of H. A similar expression was given by Miller et al. [16] for homogeneous arrays. Here we generalize this approach to nonhomogeneous arrays and justify it rigorously.

Remark. If $d_i = d \ \forall i \in \{1, \ldots, n\}$, we can simplify:

$$c_{max} = \arctan \left(\frac{\sum_{i=1}^{n} \cos(Ha_i)}{\sum_{i=1}^{n} \sin(Ha_i)} \right) .$$

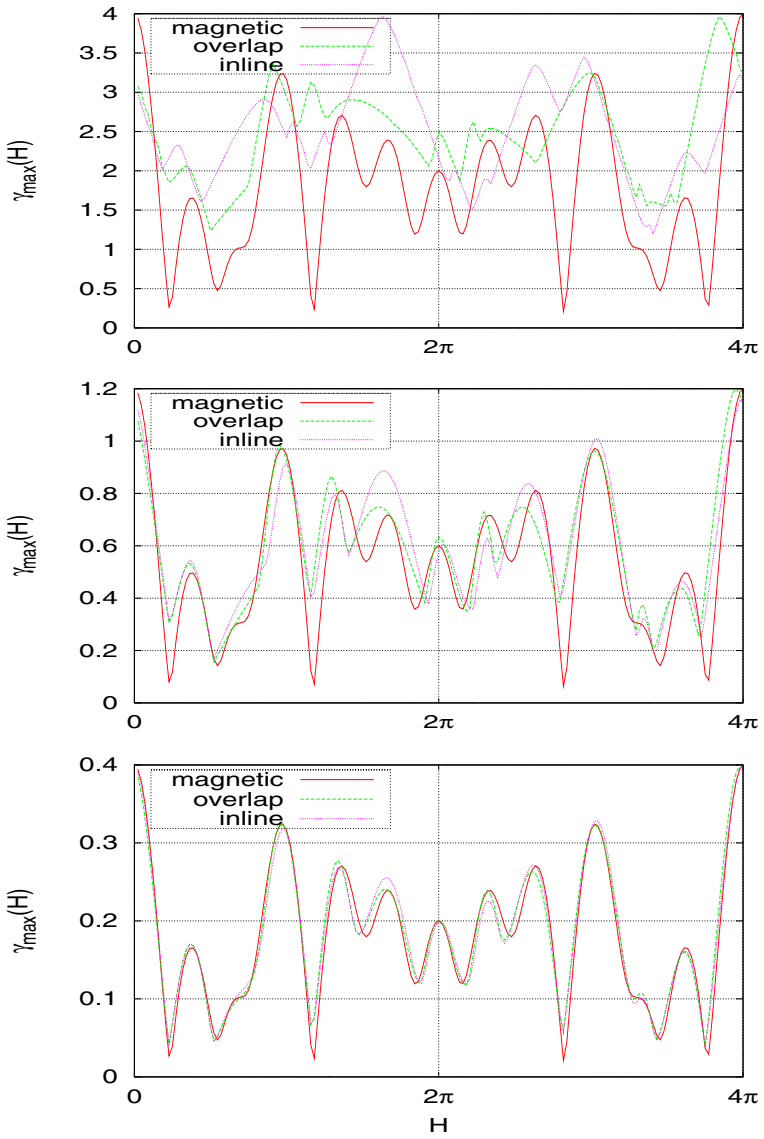


FIG. 4.4. *Plot of $\gamma_{max}(H)$ for the same junction unit and different coefficients d_i, which are all equal; from top to bottom, $d_i = 1$, 0.3, and 0.1. The distances between the junctions are $l_1 = 1.5$, $l_2 = 2.5$, $l_3 = 2$, $l = 10$.*

The maximum current (4.12) is then

$$\gamma_{max}(H) = d \left| \sum_{i=1}^{n} \sin(Ha_i + c_{max}(H)) \right|.$$

Here changing the value of d will change linearly the amplitude of the γ_{max} curve; this is not the case for the solutions of the boundary value problem (2.6). We can notice, too, that the $\gamma_{max}(H)$ obtained from this approximation is invariant by the

transformation

$$\forall t \in \Re, \ \begin{cases} a_i & \to & t a_i, \\ H & \to & H/t. \end{cases}$$

We will show in the next sections that when $d_i \ll 1$, (4.12) provides a good estimate of the $\gamma_{max}(H)$ curve of a circuit. In addition, from its invariant properties we can compare different models and estimate the parameters of the circuit. A cooperation has begun with Boussaha and Salez from the Observatoire de Paris to match theory and design for this type of circuit with specific properties [3].

5. Numerical solutions. We used two different methods, a stepping in the (H, γ) plane using a Newton iteration, and what we call the method of implicit curves to find the maximal current of (4.1) for H given.

5.1. Newton's method. We start from the system of nonlinear transcendental equations (3.11), which is written for $n = 5$. Introducing the vector $X = (\phi_1, \phi_2, \ldots, \phi_n)$, (3.11) can be written as $F(X) = 0$, where F is a nonlinear map from R^n to R^n. To solve this equation numerically, we use the Newton method:

$$X_{k+1} = X_k - (\nabla F(X_k))^{-1} F(X_k),$$

where $\nabla F(X_k)$ is the gradient of F evaluated at $X = X_k$. A first problem is to choose the initial vector X_0. For that consider $H = 0$; there we expect a solution such that $\gamma \approx \sum_i^n d_i$ and consequently $\phi_i \approx \pi/2\,[2\pi]$. We have our initial vector. After finding the solution, we step in H and take as an initial guess the previous solution found, which for a small step in magnetic field is assumed to be close to the one we are looking for. In this way, we obtain a solution with a magnetic field $H + dH$ and a current γ. We can then increase γ until the method does not converge, and this gives the maximum current $\gamma_{\max}(H + dH)$ for increasing H. Similarly we can compute $\gamma_{max}(H)$ by starting with a large magnetic field and decreasing H to 0. This curve will in general be different from the one obtained when increasing H due to hysteresis. The two curves need to be overlapped to find $\gamma_{max}(H)$. So, we introduce another method to be sure to obtain the γ_{max} curve directly.

5.2. Implicit curves method. The polynomials (3.4) and (3.6) establish the existence and value of ϕ at the junctions. This function should satisfy the boundary conditions. The first one,

$$\phi'(0) = P_1'(0) = H - (1 - \nu)\gamma/2,$$

is true by construction; the second (for n junction circuit),

$$P_{n+1}'(l) = H + (1 - \nu)\frac{\gamma}{2}, \tag{5.1}$$

is true only for the solutions of (4.1). As we have remarked in section 3.1, "The solution as a piecewise polynomial," ϕ is entirely determined by ϕ_1, γ, and H. For H given, the solutions of (5.1) define a relation between ϕ_1 and γ. Thus, the maximal current solution depends on ϕ_1 and γ, and (5.1) is the constraint it should satisfy. As the solutions ϕ are defined modulo 2π from (3.1), we can assume $\phi_1 \in [-\pi, \pi]$. On the other hand, because of (3.3), $\gamma \in [0, \sum_i d_i]$. To solve this problem using the software

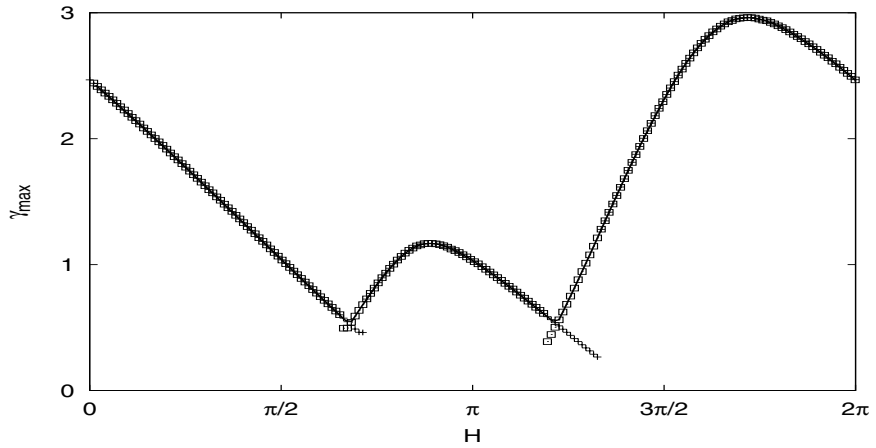


FIG. 5.1. *Comparison between the Newton method and the implicit curve method for the γ_{max} curve for a three-junction unit; $a_1 = 1$, $a_2 = 2$, $a_3 = 3$, $d_1 = d_2 = d_3 = 1$, $\nu = 1$, and $l = 10$. The square (resp., the +) symbols correspond to the Newton results for decreasing (resp., increasing) H, and the continuous line corresponds to the results of the implicit curve method.*

package Maple [15], we plot the implicit function (the constraint) of the two variables ϕ_1 and γ with H and ν fixed, defined by

$$P'_{n+1}\big|_{x=l}(\phi_1, \gamma, \nu, H) - H - \frac{1-\nu}{2}\gamma = 0, \tag{5.2}$$

with $(\phi_1, \gamma) \in [-\pi, \pi] \times [0, \sum_i d_i]$. The program searches, in an exhaustive way, the biggest value of γ of this implicit curve. Incrementing H, we obtain the relation $\gamma_{max}(H)$. We give an expression of P'_{n+1} for two and three junctions, in the appendix (section 9.1).

Compared to the Newton method detailed in the previous section, this method has the advantage of converging to a global maximum γ_{max}, as long as we give enough points to plot the implicit curve. Figure 5.1 compares $\gamma_{max}(H)$ using the two methods for a three-junction unit. The solution given by the implicit curve method is shown as a continuous line and superposes exactly with the other two plots. With the Newton method we can get trapped in local maxima, while the implicit curve method always gives the global maximum. On the other hand, the Newton method is much faster.

6. Two junctions. We have seen two methods to solve the problem numerically and established general properties. Now let us use these results for an array with a few junctions.

6.1. Same junction strength ($d = d_1 = d_2$). In Figure 6.1 we plot in the left panel $\gamma_{max}(H)$ of a two-junction unit. We find the expected periodicity $H_p = 2\pi/(a_2 - a_1)$, with a maximum for $H = 0$ in the inline case ($\nu = 0$). For the overlap feed, we have exactly the inline curve plus a magnetic shift. Notice that for the inline feed the amplitude of the γ_{max} curve is not proportional to d_i, contrary to the magnetic approximation. The larger the d_i, the further away the $\gamma_{max}(H)$ curves are from the ones given by the magnetic approximation. This is expected because the magnetic approximation neglects the effect of d_i on the phase. In this section, to simplify the discussion we will restrict ourselves to the inline current feed. However, the results will be valid for the general case. The maximum of γ_{max} corresponding

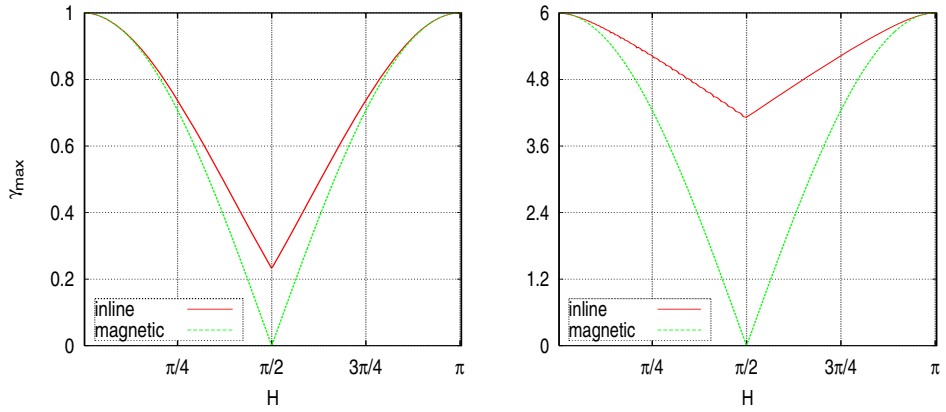


FIG. 6.1. *Plot of the γ_{max} curve for a two-junction unit such that $l_1 = 2$. In the left panel, $d_1 = d_2 = 0.5$, while in the right panel, $d_1 = d_2 = 3$.*

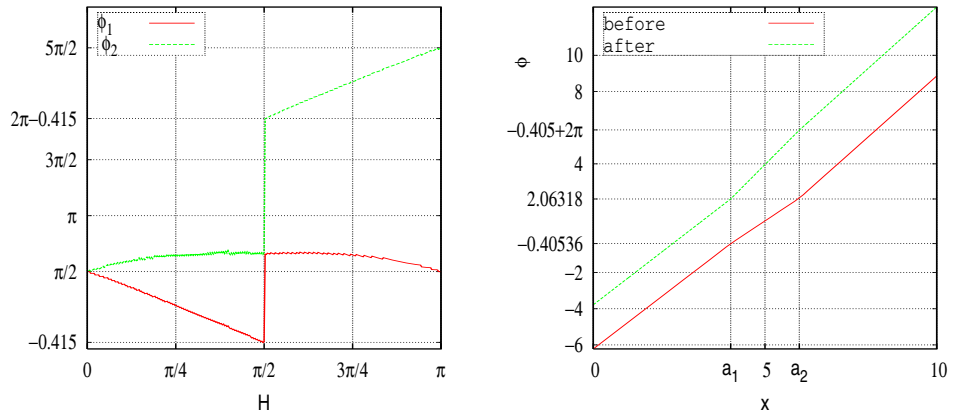


FIG. 6.2. *Left panel: plot of the phases ϕ_1 and ϕ_2 as a function of the magnetic field H for the same junction unit as the one shown in the left panel of Figure* 6.1 *Right panel: plot of $\phi(x)$ for the same device for $H < \pi/(a_2 - a_1)$ (solid line) and $H > \pi/(a_2 - a_1)$ (dashed line).*

to $H \equiv 0(H_p)$ is the only case where $(\phi_2 - \phi_1)/(a_2 - a_1) = H$. On the other hand, by construction, in the magnetic approximation $(\phi_2 - \phi_1)/(a_2 - a_1) = H \; \forall H$. In the general case, the closer H is to $\pi/(a_2 - a_1)$, the further $(\phi_2 - \phi_1)/(a_2 - a_1)$ is from H. This can be seen in the right panel of Figure 6.2. Thus, there will be more tunneling current in one junction than in the other. This phenomenon increases as H increases from 0 to $\pi/(a_2 - a_1)$. For that value, we have two possible solutions for γ_{max}, as shown in the left panel of Figure 6.2 for $H = \pi/2$. As the field crosses $\pi/(a_2 - a_1)$ the two junctions behave in the opposite fashion, as shown by the switch of the jumps in ϕ_x at the junctions; see the right-hand panel of Figure 6.2. These two solutions or reversing behavior of the junction cause a jump in the derivative $\gamma'_{max}(H)$ of $\gamma_{max}(H)$. As long as the evolution of ϕ_1 (or ϕ_2) is continuous there is no jump in γ'_{max}. To summarize, the smaller d is, the closer $(\phi_2 - \phi_1)/(a_2 - a_1)$ is to H. Another way of relaxing this constraint on $(\phi_2 - \phi_1)/(a_2 - a_1)$ for a constant d is to separate the junctions and, we can show that if $l_b = a_2 - a_1 \to +\infty$, then $\gamma_{max}(H) \to d_1 + d_2$.

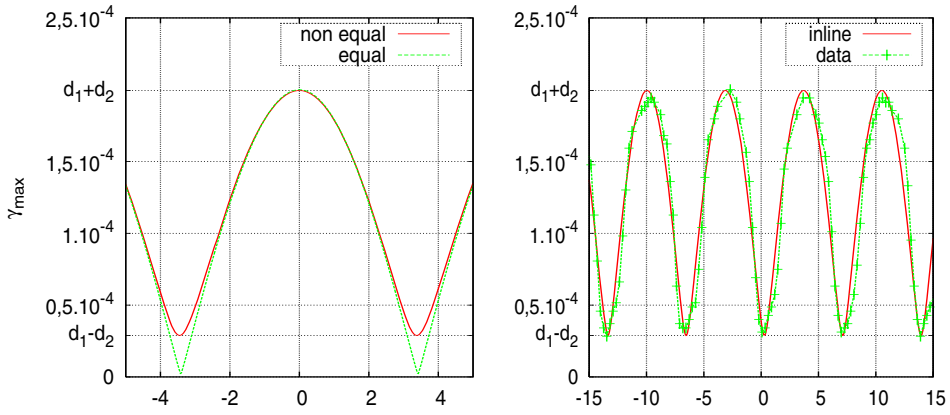


FIG. 6.3. *Plot of $\gamma_{max}(H)$ curves for two two-junction units with inline current feed. Left panel: comparison of $\gamma_{max}(H)$ given by the model for the cases $d_1 = d_2$ and $d_1 \neq d_2$. Right panel: the fit of the experimental data from a two-junction unit of the Observatory of Paris (reproduced from* [3] *with permission of Morvan Salez).*

6.2. Regularity of $\gamma_{max}(H)$. Junctions are never perfectly similar, and small differences in their areas or their critical currents will affect γ_{max}. In the left panel of Figure 6.3, showing $\gamma_{max}(H)$ for a two-junction device, there is no discontinuity of the slope of the curve $\gamma_{max}(H)$ labeled "nonequal"; $\partial\gamma_{max}/\partial H$ exists everywhere. In this case, the values of $\phi_1(H)$ and $\phi_2(H)$ associated with $\gamma_{max}(H)$ vary continuously.

To show this, consider a circuit such that $d_1 > d_2$. If $\phi_1 = \pi/2$, remark (3.3) implies $d_1 - d_2 \leq \gamma_{max} \leq d_1 + d_2$. Now let us find the values of H for which these bounds can be reached. From (3.3) and (3.6) we have

$$\begin{aligned} \gamma &= d_1 \sin(\phi_1) + d_2 \sin(\phi_2), \\ \phi_2 &= -\frac{\nu j}{2} l_1^2 + \left(H - \left(\nu a_1 + \frac{1-\nu}{2} l \right) j + d_1 \sin\phi_1 \right) l_1 + \phi_1, \end{aligned} \tag{6.1}$$

with $l_1 = a_2 - a_1$. By substituting the second equality into the first and taking the derivative with respect to ϕ_1, we obtain

$$\frac{\partial\gamma}{\partial\phi_1} = d_1 \cos\phi_1 + d_2 \left[\left(- \left(\nu \frac{a_2 + a_1}{2l} + \frac{1-\nu}{2} \right) \frac{\partial\gamma}{\partial\phi_1} + d_1 \cos\phi_1 \right) l_1 + 1 \right] \cos\phi_2.$$

Since we search for the maximum of γ, then $\partial\gamma/\partial\phi_1 = 0$, so that

$$d_1 \cos\phi_1 = -d_2 \left(d_1 l_1 \cos(\phi_1) + 1 \right) \cos\phi_2. \tag{6.2}$$

When $\phi_1 = \pi/2$, this condition gives $\phi_2 = \pi/2[\pi]$. Now, inserting these solutions into (6.1), we obtain the values of H for which these solutions are possible:

$\gamma_{max}(H)$	H
$d_1 + d_2$	$2k\pi/l_1 + [\nu(a_1 + l_1/2) + (1-\nu)l/2](d_1 + d_2)/l - d_1$
$d_1 - d_2$	$(2k+1)\pi/l_1 + [\nu(a_1 + l_1/2) + (1-\nu)l/2](d_1 - d_2)/l - d_1$

This enables us to estimate d_1 and d_2 from the curve $\gamma_{max}(H)$.

We now proceed to give the conditions that d_1 and d_2 should satisfy in order to have an angular or smooth $\gamma_{max}(H)$. Since $\phi_2(H)$ varies continuously, $\cos(\phi_2)$ takes

all the values between -1 and 1. We assume that $\forall\phi_1$, $d_1 l_1 \cos(\phi_1) + 1 \geq 0$, where $l_1 = a_2 - a_1$.[1] We consider the following two cases:

1. $\cos\phi_1 \leq 0$: as $\cos\phi_1 \geq -1$ from (6.2), we obtain

$$d_2 \cos(\phi_2) \leq \frac{d_1}{1 - d_1 l_1} \ .$$

Since $\cos(\phi_2)$ must take all values between -1 and 1 and $d_2 > 0$,

$$d_2 \leq \frac{d_1}{1 - d_1 l_1} \ . \tag{6.3}$$

This is the maximal value that d_2 can take compared to d_1.

2. $\cos\phi_1 \geq 0$: as $\cos\phi_1 \leq 1$, for the same reason we obtain

$$d_2 \geq \frac{d_1}{1 + d_1 l_1} \ . \tag{6.4}$$

To summarize, $d\gamma_{max}(H)/dH$ does not vary continuously if

$$\frac{d_1}{1 - d_1 l_1} \leq d_2 \leq \frac{d_1}{1 + d_1 l_1}. \tag{6.5}$$

To illustrate this effect we consider the configuration of a a microstrip with inline current feed with two Josephson junctions built by Salez and Boussaha at the Observatoire de Paris [17, 4]. The results are shown in Figure 6.3. The square junctions have an area of $w_j^2 \approx 1\mu m^2$, the Josephson length is $\lambda_J = 5.6\mu m$, and $l_1 = a_2 - a_1 = 13\mu m$ (using the junction centers). This gives $d_1 = d_2 \approx 0.0357$, $l_1 \approx 2.32$ if the areas are equal. However, the experimental data does not go to 0, and so the junctions are probably slightly different than expected from (6.3) and (6.4),

$$0.032969 \leq d_2 \leq 0.038923.$$

Only a 10% difference in area is enough to give a regular $\gamma_{max}(H)$. From the fit of the experimental data (right panel of Figure 6.3) we can estimate the areas of the junctions as $w_1^2 = 0.85255\mu m^2$ and $w_2^2 = 1.1417\mu m^2$.

As we have seen in the previous section, when the γ_{max} curve does not show any spike, it is bounded by $d_1 + d_2$ and $|d_1 - d_2|$. From this we can obtain the characteristics of the two junctions, their critical current density, and area, except that we do not know which junction corresponds to d_1 and which to d_2. However, if the γ_{max} does not have any spikes, then we can give the exact area of the junctions, assuming that the critical density current is known.

7. Many junctions. A two-junction circuit behaves as a simple SQUID and shows a regular $\gamma_{max}(H)$. To obtain specific properties for advanced detectors, experimentalists make devices with more junctions.

7.1. 3-device junction. When we add a new junction to a circuit with two junctions, new oscillations appear on $\gamma_{max}(H)$. We cannot predict the amplitude of the oscillations, but we can have an idea of their number in one period, i.e., the interval $[0, H_p]$. We introduce the phase difference for $H = 0$, $\delta\phi_i = \phi_i - \phi_1$. Using

[1] For small junctions this is not a strong constraint, because since $d_i = w_i^2/w \ll 1$, w_i, $w \ll 1$, and $l_i \ll 1$ are about 10^{-2}.

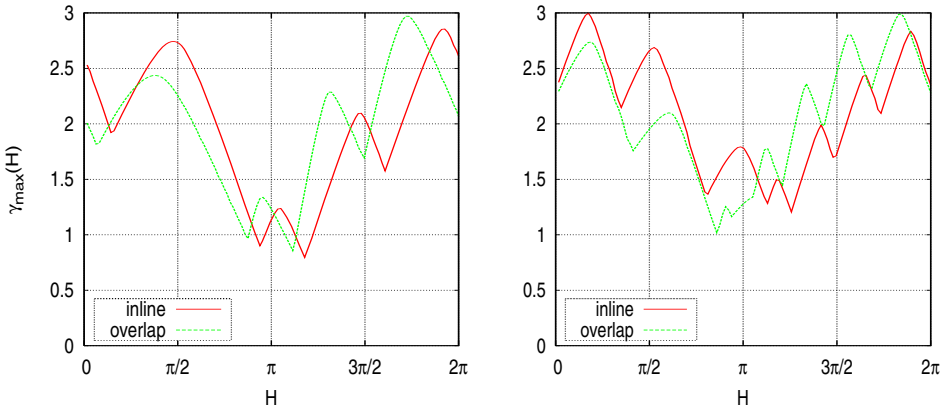


FIG. 7.1. *Plot of $\gamma_{max}(H)$ curves for a two-junction unit $a_1 = 1$, $a_2 = 2$ ($l_1 = 1$) and a third junction placed at $a_3 = 5$, $l_2 = 3$ (left panel) and $a_3 = 8$, $l_2 = 6$ (right panel). All the junctions have the coefficient $d_i = 1$.*

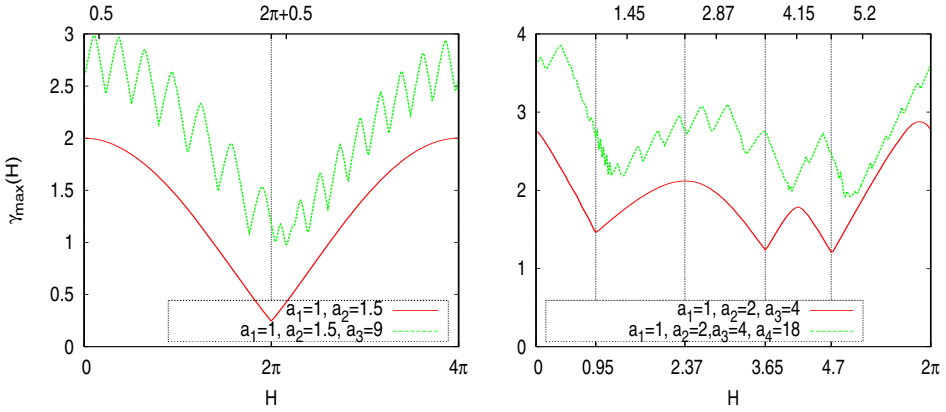


FIG. 7.2. *$\gamma_{max}(H)$ curves for all devices $d_i = 1$. Left panel: we compare a circuit of two junctions with one of three junctions: same junction unit plus one. Right panel: a three- and a four-junction circuit. We report on top of the box the expected shift to the left given by $d_{n+1}/2 = 0.5$.*

Proposition 4.1, we can state that as H goes from 0 to H_p, $\phi_2 - \phi_1$ goes from $\delta\phi_2$ to $\delta\phi_2 + 2\pi l_1/l_1 = \delta\phi_2 + 2\pi$. Similarly $\phi_3 - \phi_1$ goes from $\delta\phi_3$ to $\delta\phi_3 + 2\pi(l_2 + l_1)/l_1 = \delta\phi_3 + 2\pi(l_2/l_1 + 1)$, which becomes $\delta\phi_3 + 2\pi(k + 1)$ if the junctions are placed harmonically so that $l_2 = kl_1$. In that case we expect the $\gamma_{max}(H)$ curve to present $k + 1$ bumps within one period. In Figure 7.1, the junctions are placed in a harmonic way $a_3 - a_2 = k(a_2 - a_1)$, where $k = 3$ (left panel) and $k = 6$ (right panel). As expected, we see the four intermediate "bumps" in the $\gamma_{max}(H)$ curve in the left panel and seven bumps in the curve of the right panel. We can see the periodicity given by $H_p = 2\pi/(a_2 - a_1) \equiv 2\pi/l_1$, which adds new oscillations. This picture shows that the closer the third junction is to the junction unit, the fewer oscillations there are. These estimations hold approximately in the case of an array with more junctions.

In other words, when $a_3 - a_1$ is large, as in the right panel of Figure 7.1 and the left panel of Figure 7.2, the shape of the γ_{max} curve tends to that for a two-junction circuit. We explain this below.

7.2. Influence of a faraway single junction for the inline current feed. In Figure 7.2, for each panel we plot a γ_{max} curve, for an n-junction unit and another with the same junction unit plus a distanced junction. $(n+1)$-junction γ_{max} curves look like n-junction curves to which a shift has been added. We want to evaluate this shift.

Note that for the junction n, using the notation of (3.4) and (3.6), we know that ϕ_{n+1} is determined by P_{n+1}. If we increase P'_{n+1} of ϵ, then ϕ_{n+1} increases by $\epsilon(a_{n+1} - a_n)$. Thus, a variation at ϕ_n of $\epsilon = \pm\pi/(a_{n+1} - a_n)$ is enough to obtain $\sin\phi_{n+1} = 1$. The farther the last junction, the smaller ϵ, and consequently this junction has the smallest action on the junction unit. So, in the search of γ_{max}, the value of $\sin\phi_{n+1}$ is near 1. The γ^{n+1}_{max} curve of a circuit with $n+1$ junctions is close to $\gamma^n_{max} + d_{n+1}$, i.e., the curve for the n-junction circuit with n junctions plus the maximal contribution of the last junction.

Now let us assume that $\sin\phi_{n+1} = 1$. We must not forget the boundary conditions of our problem: $\phi'|_{\{0,l\}} = H \mp \gamma/2$. But boundary conditions at the junction unit are $\phi'(0) = H - \gamma/2$ and $a_n < x_0 < a_{n+1}$, $\phi'(x_0) = H + \gamma/2 - d_{n+1}$. We cannot compare this junction unit with the n-junction problem because of different boundary conditions. As we have done in section 5.2, let $H' = H - d_{n+1}/2$. The previous boundary conditions become

$$\phi'|_{\{0,a_n^+\}} = H' \mp \frac{\gamma}{2} \ .$$

We find wanted boundary values. Finally we show that

$$\lim_{a_{n+1}-a_n\to+\infty} \gamma^{n+1}_{max}\left(H + \frac{d_{n+1}}{2}\right) = \gamma^n_{max}(H) + d_{n+1}. \tag{7.1}$$

Figure 7.2 illustrates this convergence.

This argument cannot be extended simply to the overlap or general current feed, for two reasons. First, introducing or taking out the last junction a_{n+1} induces a variation of the magnetic shift H_ν given by (4.5). We could estimate it, but we have the problem that the curvature of ϕ for an n-junction device is $\nu j/2$, where j depends on the number of the junctions. This will affect the shift between the junctions and consequently the curve $\gamma_{\max}$.

However, numerical simulation shows that (7.1) remains a good approximation for the general case, even with a small number of junctions. The general feed and inline feed problems coincide when $d_{n+1}/\sum_1^n d_i$ tends to zero. Going back to the physical device, this means that the forces of the junctions are very small, $d_i \approx 10^{-2}$, and for these values the inline and overlap results are practically indiscernible from the magnetic approximation. Then (7.1) can be used.

7.3. A real device with five Josephson junctions. We have compared our theory to the experimental results for a device with two Josephson junctions. The same team at the Observatoire de Paris has made a device with five junctions. Here the γ_{max} curve obtained is totally different from the one for a simple SQUID. The parameters are $l_1 = 20$, $l_2 = 42$, $l_3 = 12$, and $l_4 = 6$. Figure 7.3 shows the γ_{max} curve where the current and magnetic field have been scaled using approximately the same factors as for the SQUID of Figure 6.2. Our modeling approach also gives excellent agreement for experimental uniform arrays of 5, 10, and 20 junctions.

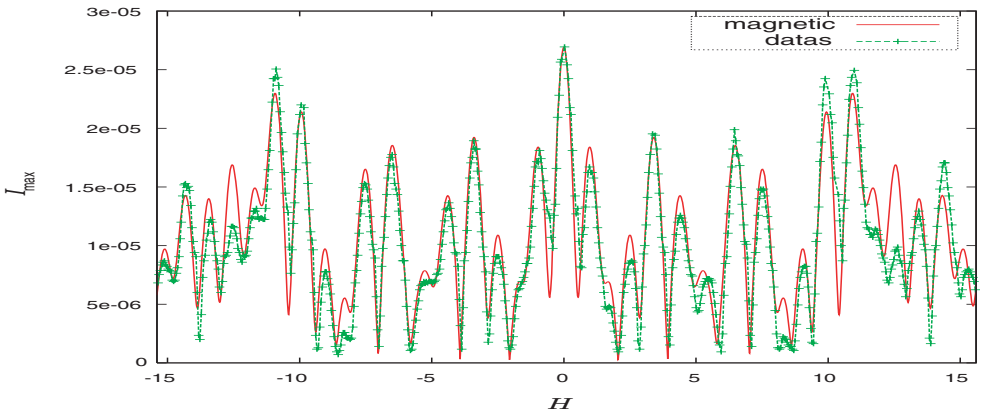


FIG. 7.3. *Experimental $I_{\max}(H)$ for an array of five junctions in a 2D microstrip line built by Salez and Boussaha of the Observatory of Paris (reproduced with their permission). The measured data is presented by the + symbols, and the magnetic approximation result is shown with a solid line.*

8. Conclusion. We have mathematically analyzed a new continuous/discrete model for describing arrays of small Josephson junctions. Compared to standard "lumped" approaches we do not approximate the equations, except for neglecting the phase variation in the junction. In particular, our approach preserves the matching at the interface.

We establish the periodicity of the $\gamma_{max}(H)$ curve and show how it depends on the position of the array with respect to the microstrip. This is particularly interesting for estimating the proportion of inline current feed versus overlap feed. We show how separating a junction from an array will influence $\gamma_{max}(H)$.

We introduce a numerical method for estimating $\gamma_{max}(H)$ which is more reliable than the standard Newton method used up to now.

The relative simplicity of the model allows in-depth analysis that is out of reach for the 2D model. In particular, we show that solutions for general current feed tend to the solutions of inline feed when $\nu j/l \to 0$. All models reduce to what we call the magnetic approximation for small d_i.

Our global model gives a very good agreement with experimental curves obtained for arrays of up to five junctions. The simplicity of the magnetic approximation allows us to address the inverse problem of determining features of the array from $\gamma_{max}(H)$.

9. Appendix.

9.1. Implicit curves. In this part, we give an example of $P'_{n+1}(x)$ for systems with three junctions. We define the following:

$$\begin{cases} \sin_1 &= \sin(\phi_1), \\ C_1 &= \left(d_1 \sin(\phi_1) - \dfrac{\nu\gamma a_1}{l} + H - (1-\nu)\,\dfrac{\gamma}{2}\right)(a_2 - a_1) + \phi_1, \\ D_j &= \dfrac{\nu\gamma(a_{j+1} - a_j)^2}{2l}. \end{cases}$$

Then (3.4) and (3.6) give

$$P_3'(x) = -\frac{\nu\gamma x}{l} + d_2\sin(-D_1 + C_1) + d_1\sin_1 + H - (1-\nu)\frac{\gamma}{2}, \tag{9.1}$$

$$\begin{aligned} P_4'(x) = -\frac{\nu\gamma x}{l} + d_3\sin\Big[-D_2 + \Big\{-d_2\sin(D_1 - C_1) - \frac{\nu\gamma a_2}{l} \\ + d_1\sin_1 + H - (1-\nu)\frac{\gamma}{2}\Big\}(a_3 - a_2) - D_1 + C_1\Big] \\ + d_2\sin(-D_1 + C_1) + d_1\sin_1 + H - (1-\nu)\frac{\gamma}{2}. \end{aligned} \tag{9.2}$$

This example shows that $P_k'(x)$ is C^∞ in the variables $(\gamma, \phi_1, \nu, H, x)$. In particular, $P_n'(l)$ is C^∞ in the variables (γ, ϕ_1, ν, H).

9.2. The current feed factor ν: Analytical estimates. Equation (4.9) shows that we tend to an inline current feed when l is large. However, we should show that the γ_{max} curve tends to that for the inline feed.

LEMMA 9.1 (solution). *$\forall\,\phi_1$ and H $\exists$ a γ such that* (4.1) *has a solution.*

Proof. As we have seen in section 5.2, it is sufficient to solve (5.1), $P_{n+1}'(l) = H + (1-\nu)\gamma/2$, to find a solution. Let us fix a value for ϕ_1 with ν, H, l given. If $\gamma < -\sum_{i=1}^n d_i$, then $P_{n+1}'(l) < H + (1-\nu)\gamma/2$. Conversely when $\gamma > \sum_{i=1}^n d_i$, we obtain $P_{n+1}'(l) > H + (1-\nu)\gamma/2$. But by construction, $P_{n+1}'(l)$ is a function continuous in all its variables, in particular γ. Thus we have at least one value of γ in $[-\sum_{i=1}^n d_i, \sum_{i=1}^n d_i]$ such that $P_{n+1}'(l) = H + (1-\nu)\gamma/2$, and thus it is a solution for that value of ϕ_1. ∎

We want to study the variation of $\gamma(H)$ versus the current feed ν. At this point, we do not consider the γ_{max} curve. Let us fix ϕ_1. Using the previous property, we know that there exists at least one solution of (4.1), and particularly almost one γ. Without changing ϕ_1 or H, we plot all the possible γ versus ν in Figure 9.1. We call this curve the $\gamma(\nu)$ curve. To plot this $\gamma(\nu)$ curve, we use the same parameter as in Figure 4.3, with $H = 1.3617$ (see the top panel; we choose this H because there is a big difference between the solution for inline and overlap current feeds). We choose for ϕ_1 the value found with Maple, giving the maximum γ_{max} for the inline feed. Figure 9.1 (top panel), for $\nu = 0$, confirms the γ_{max} value found in Figure 4.3. But for overlap the maximum current we can obtain is near 0. Thus, there is another value of ϕ_1 for γ_{max} of overlap current feed ($\phi_1 \approx 0.252$).

Let us study the curve $\gamma(\nu)$. By definition, $\gamma = \sum_{i=1}^n d_i\sin(\phi_i)$. Let ϕ_1 be a value such that $\partial\gamma/\partial\nu$ exists; then

$$\begin{aligned} \frac{\partial\gamma}{\partial\nu} &= \sum_{i=1}^n d_i\frac{\partial\phi_i}{\partial\nu}\cos(\phi_i), \\ \left|\frac{\partial\gamma}{\partial\nu}\right| &\le \sum_{i=1}^n d_i\left|\frac{\partial\phi_i}{\partial\nu}\right|. \end{aligned} \tag{9.3}$$

With $\phi_i = \phi(a_i)$, we note in the following: $\phi_i' = \lim_{\epsilon\to 0}\phi'(a_i - \epsilon)$ (the left derivative of ϕ). Let us make some remarks: as ϕ_1 is fixed,

$$\left.\frac{\partial\phi_1}{\partial\nu}\right|_{\nu=0} = 0,$$

$$\frac{\partial\phi_1'}{\partial\nu} = \frac{\partial}{\partial\nu}\left\{H - \left(\frac{1}{2} - \frac{\nu l_b}{2l}\right)\frac{\gamma}{2}\right\} = -\left(\frac{1}{2} - \frac{\nu l_b}{2l}\right)\frac{\partial\gamma}{\partial\nu} + \frac{l_b\gamma}{4l};$$

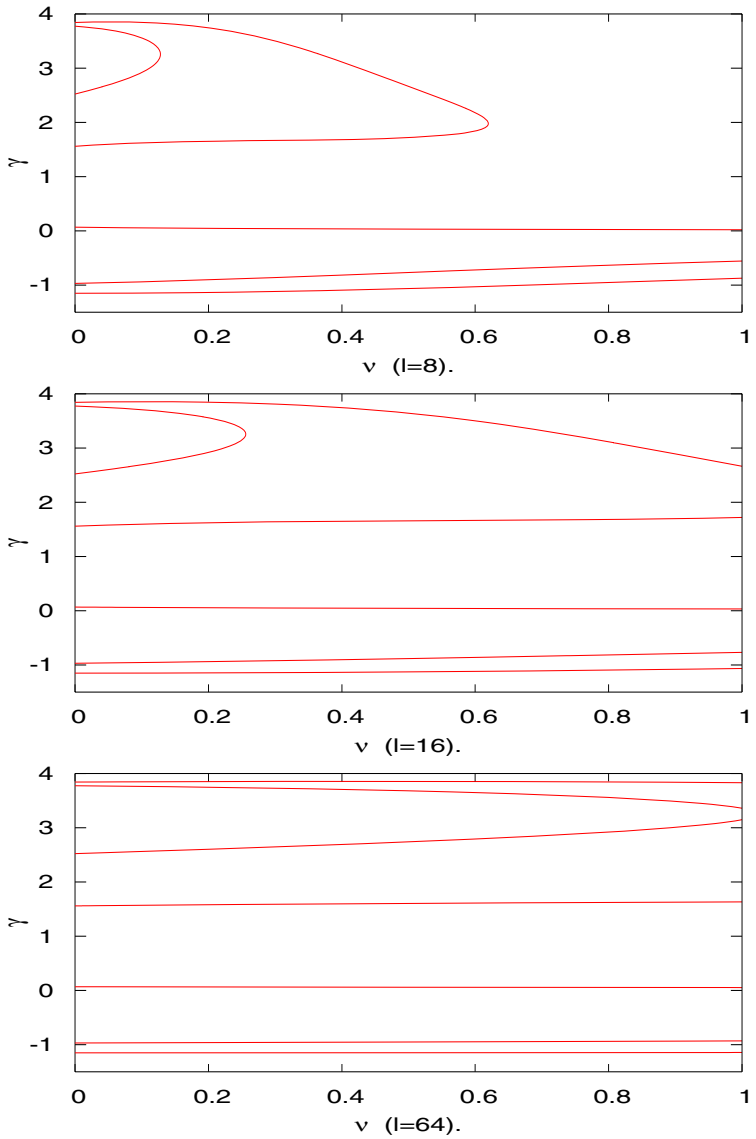


FIG. 9.1. *Each panel corresponds to the devices studied in the panels of Figure* 4.3. *We plot the implicit curve* $\gamma(\nu)$ *curve for* $H = 1.3617$, $\phi_1 = 1.3897$, *corresponding to field* H *giving the maximum current* $\gamma_{max}(H)$ *for the inline feed. From top to bottom we increase the length of the device and notice the stretching of the* $\gamma(\nu)$ *curve (with the coefficient found in* (4.9): l_b/l*).*

using (3.4) and (3.6) we can begin iteration,

$$\frac{\partial \phi_i}{\partial \nu} = -\left(\nu \frac{\partial \gamma}{\partial \nu} + \gamma\right) \frac{l_{i-1}^2}{2l} + l_{i-1} \frac{\partial \phi'_{i-1}}{\partial \nu} + \frac{\partial \phi_{i-1}}{\partial \nu}(d_{i-1} l_{i-1} \cos(\phi_{i-1}) + 1),$$

$$\frac{\partial \phi'_i}{\partial \nu} = -\left(\nu \frac{\partial \gamma}{\partial \nu} + \gamma\right) \frac{l_{i-1}}{l} + d_{i-1} \frac{\partial \phi_{i-1}}{\partial \nu} \cos(\phi_{i-1}) + \frac{\partial \phi'_{i-1}}{\partial \nu},$$

with $l_i = a_{i+1} - a_i$. This last equation can be written, for $i \geq 3$,

$$\left.\frac{\partial \phi_i'}{\partial \nu}\right|_{\nu=0} = -\left(\nu \frac{\partial \gamma}{\partial \nu} + \gamma\right) \frac{a_i - a_1}{l} + \left.\frac{\partial \phi_1'}{\partial \nu}\right|_{\nu=0} + \sum_{k=2}^{k-1} d_k \frac{\partial \phi_k}{\partial \nu} \cos(\phi_k).$$

We obtain

$$\begin{aligned}
\frac{\partial \phi_{i+1}}{\partial \nu} &= -K_1^i \frac{\partial \gamma}{\partial \nu} - K_2^i \gamma + \frac{\partial \phi_i}{\partial \nu} + l_i \sum_{k=2}^{i} d_k \frac{\partial \phi_k}{\partial \nu} \cos(\phi_k), \\
K_1^i &= l_i \left[\frac{\nu l_i}{2l} + \nu \frac{a_{i+1} - a_1}{l} + \frac{1}{2} - \frac{\nu l_b}{2l}\right], \\
K_2^i &= l_i \left[\frac{l_i}{2l} + \frac{a_{i+1} - a_1}{l} - \frac{l_b}{2l}\right].
\end{aligned}$$

Applying absolute values, we obtain

$$\left|\frac{\partial \phi_{i+1}}{\partial \nu}\right| \leq K_1^i \left|\frac{\partial \gamma}{\partial \nu}\right| + K_2^i |\gamma| + \left|\frac{\partial \phi_i}{\partial \nu}\right| + l_i \sum_{k=2}^{i} d_k \left|\frac{\partial \phi_k}{\partial \nu}\right|. \tag{9.4}$$

We do not need to find the exact expression of $|\partial \phi_{i+1}/\partial \nu|$; we know that it is a linear combination of $|\partial \gamma/\partial \nu|$ and $|\gamma|$ and so is $|\partial \phi_2/\partial \nu|$. Using (9.4), we can show by iteration that $|\partial \phi_i/\partial \nu|$ is a linear combination of $|\partial \gamma/\partial \nu|$ and $|\gamma|$. Applying this last remark to inequality (9.3), we obtain that there exist two constants, C_1 and C_2, such that

$$\left|\frac{\partial \gamma}{\partial \nu}\right| \leq C_1 \left|\frac{\partial \gamma}{\partial \nu}\right| + C_2 |\gamma|;$$

C_1 and C_2 are a combination of d_i, l_i, K_1^i, and K_2^i. For ν and d_i sufficiently small, $|C_1| < 1$; then

$$\left|\frac{\partial \gamma}{\partial \nu}\right| \leq \frac{C_2}{1 - C_1} |\gamma|. \tag{9.5}$$

This last equation implies local continuity of the γ curve as a function of ν. As we have seen in section 4.3, increasing l is equivalent to decreasing the range of ν (given by μ). $\forall \epsilon$, $\exists L /\ l \geq L \Rightarrow \mu \leq \epsilon$. This shows the convergence of the γ_{max} $(0 \leq \nu \leq 1)$ curve to the one for inline current feed when $l \to +\infty$.

9.3. Convergence by the junction coefficient d_i. We want to show that the solution for the general case converges to the solution of the inline case, for small d_i. We have shown in Lemma 9.1 that for H and ϕ_1 given, we can find at least one solution, whatever ν. This shows that for the same ϕ_1, we can find a general and an inline solution. Let us define the following:

1. $P_n^i(x)$, $\forall x \in]a_n, a_{n+1}[$, a solution of inline problem ($\nu = 0$) of this circuit, γ^i the maximal current associated to the value ϕ_1.

2. $P_n^g(x)$, $\forall x \in]a_n, a_{n+1}[$, a general solution ($\nu \neq 0$, same l, and same junction unit), γ^g the maximal current associated with the value ϕ_1.

3. α_j and β_j by

$$\begin{cases} \alpha_j = P_j^{g'}(a_j) - P_j^{i'}(a_j), \\ \beta_j = P_j^g(a_j) - P_j^i(a_j). \end{cases}$$

As $P_1^g(a_1) = \phi_1 = P_1^i(a_,)$, we have $\beta_1 = 0$. We can calculate α_1 using (4.8),

$$\alpha_1 = P_1^{g'}(a_1) - P_1^{i'}(a_1) = -\frac{\gamma^i - \gamma^g}{2} + \frac{\nu l_b}{2l}\gamma^g.$$

However, γ^i and γ^g are positive, and so

$$|\alpha_1| \leq \left(\frac{1}{2} + \frac{\nu l_b}{2l}\right) \sum_{i=1}^{n} d_i. \tag{9.6}$$

The aim of the following is to give an upper bound for β_i. We proceed by iteration. We recall that $l_k = a_{k+1} - a_k$. Using (3.6), we estimate β_{k+1}:

$$\beta_{k+1} = \frac{-\nu\gamma^g}{2l} l_k^2 + \left[d_k(\sin(P_k^g(a_k)) - \sin(P_k^i(a_k))) + \alpha_k\right] l_k + \beta_k. \tag{9.7}$$

Let us focus on the sine terms,

$$\begin{aligned}
\sin(P_k^g(a_k)) - \sin(P_k^i(a_k)) &= \sin(P_k^i(a_k) + \beta_k) - \sin(P_k^i(a_k)) \\
&= \sin(P_k^i(a_k))\left[\cos(\beta_k) - 1\right] + \sin(\beta_k)\cos(P_k^i(a_k)) \\
&\leq |\beta_k|^2 + |\beta_k|.
\end{aligned}$$

We assume $\beta_k \gg 1$, and thus we obtain the equivalences $\beta_k \approx \sin(\beta_k)$ and $-\beta_k^2 \approx \cos(\beta_k) - 1$, but we cannot predict the sign of $\sin(P_k^i(a_k))$ or $\cos(P_k^i(a_k))$. We neglect $|\beta_k|^2$. From (9.7),

$$\begin{aligned}
|\beta_{n+1}| &\leq \left|\frac{\nu\gamma^g}{2l} l_n^2\right| + (d_n|\beta_n| + |\alpha_n|)\, l_n + |\beta_n|, \\
|\alpha_{n+1}| &\leq \left|\frac{\nu\gamma^g}{l} l_n\right| + d_n|\beta_n| + |\alpha_n|.
\end{aligned}$$

Let us note $G = \nu \sum_{i=1}^{n} d_i/l$; we then obtain a simple system

$$\zeta_{n+1} \leq M_n \zeta_n + G V_n, \tag{9.8}$$

with, $\zeta_n = \begin{pmatrix} c|\beta_n| \\ |\alpha_n| \end{pmatrix}$, $M_n = \begin{pmatrix} cc d_n l_n + 1 & l_n \\ d_n & 1 \end{pmatrix}$, and $V_n = \begin{pmatrix} c l_n^2/2 \\ l_n \end{pmatrix}$. So, we bound $|\beta_n|$ and $|\alpha_n|$ from above, with $|\beta_1|$ and $|\alpha_1|$:

$$\zeta_n \leq M_{n-1}(\dots(M_2(M_1\zeta_1 + GV_1) + GV_2)\dots) + GV_{n-1}\ . \tag{9.9}$$

When $d_i \to 0$,

1. $G \to 0$; then equation (9.9) tends to $\zeta_n \leq M_{n-1} \dots M_2 M_1 \zeta_1$.
2. $M_k \to \begin{pmatrix} cc1 & l_k \\ 0 & 1 \end{pmatrix}$; then, $M_k \dots M_2 M_1 \to \begin{pmatrix} cc1 & \sum_{i=1}^{k-1} l_i \\ 0 & 1 \end{pmatrix}$.

From the two previous points, we obtain that

$$|\beta_i| \leq |\beta_1| + |\alpha_1|(a_i - a_1) + O\left(\sum_{i=1}^{n} d_i\right).$$

Using (9.6), we have $|\alpha_1|(a_i - a_1) \leq \left(l_b/2 + \nu l_b^2/(2l)\right) \sum_{i=1}^{n} d_i$, and the previous inequality becomes, $\forall i \in \{1, \dots, n\}$,

$$|\beta_i| \leq |\beta_1| + O\left(\sum_{i=1}^{n} d_i\right). \tag{9.10}$$

Remember that $\beta_1 = 0$. Then (9.10) shows that γ^g tends to γ^i. Since this convergence occurs independently of ϕ_1, we obtain the convergence of the γ_{max} curve.

9.4. Inline—Magnetic convergence. We show in this section the convergence of an inline solution to the magnetic approximation when $d_i \ll 1$. We already know that in this case the γ_{max} curve in overlap (or in general case) and inline tend to the same behavior. So, in this way, we show that $\forall\, \nu$ the γ_{max} curve of (4.1) tends to the magnetic approximation when $d_i \ll 1$.

We know that the magnetic approximation is given by $f(x) = Hx + c_{max}(H)$. Notice that c_{max} does not depend on the γ value (see (4.11)). We are going to compare the magnetic approximation and the inline current feed solution for the same geometry. We proceed as in the previous part, choosing $\phi_1 = Ha_1 + c_{max}$. Remember that in the inline case, ϕ is piecewise linear. Then $\forall x \in]a_i, a_{i+1}[$,

$$P_{i+1}(x) = (d_i \sin(P_i(a_i)) + P_i'(a_i))(x - a_i) + P_i(a_i).$$

Let us define

$$\begin{aligned} \alpha_i &= P_i'(a_i) - f'(a_i) = P_i'(a_i) - H, \\ \beta_i &= P_i(a_i) - f(a_i). \end{aligned}$$

We obtain that $\alpha_1 = -\gamma/2$, $\beta_1 = 0$ and, for an n-junction circuit, $\alpha_{n+1} = \gamma/2$. We estimate α_{i+1}:

$$\alpha_{i+1} = d_i \sin(P_i(a_i)) + P_i'(a_i) - H = d_i \sin(P_i(a_i)) + \alpha_i.$$

We obtain $\alpha_{i+1} = \sum_{j=1}^{i} d_j \sin(P_j(a_j)) + \alpha_1$, and thus

$$|\alpha_{i+1}| \le \sum_{k=1}^{n} d_k \ . \tag{9.11}$$

We write β_{i+1}:

$$\begin{aligned} \beta_{i+1} &= (d_i \sin(P_i(a_i)) + P_i'(a_i))(a_{i+1} - a_i) + P_i(a_i) - Ha_{i+1} + b \\ &= (d_i \sin(P_i(a_i)) + P_i'(a_i) - H)(a_{i+1} - a_i) + P_i(a_i) - Ha_i + b \\ &= \alpha_{i+1}(a_{i+1} - a_i) + \beta_i. \end{aligned}$$

Thus if $\beta_1 = 0$, then

$$\beta_i = \sum_{k=1}^{i-1} \alpha_{k+1}(a_{k+1} - a_k) \ . \tag{9.12}$$

Now using the bounds on the α's and bounding the l_i's, we get

$$|\beta_i| \le n l_b \sum_{k=1}^{n} d_k \ . \tag{9.13}$$

This shows that γ_{max} of (4.1) tends to the magnetic approximation when $\sum_{k=1}^{n} d_k$ tends to 0.

Acknowledgments. Both authors wish to thank Faouzi Boussaha and Morvan Salez for helpful discussions and for their experimental results. The authors also thank Yuri Gaididei for useful suggestions. The computations were done at the Centre de Ressources Informatiques de Haute-Normandie (CRIHAN).

REFERENCES

[1] A. Barone and G. Paterno, *Physics and Applications of the Josephson Effect*, Wiley, New York, 1982.

[2] A. Benabdallah, J. G. Caputo, and N. Flytzanis, *The window Josephson junction: A coupled linear nonlinear system*, Phys. D, 161 (2002), pp. 79–101.

[3] F. Boussaha, J. G. Caputo, L. Loukitch, and M. Salez, *Statics of nonuniform Josephson junction parallel arrays: Model vs. experiment*, http://arXiv.org/abs/cond-mat/0609757.

[4] M.-H. Chung and M. Salez, *Numerical simulation based on a five-port model of the parallel SIS junction array mixer*, in Proceedings of the 4th European Conference on Applied Superconductivity (EUCAS99), Sitges, Spain, 1999, p. 651–653.

[5] J. G. Caputo and Y. Gaididei, *Two point Josephson junctions in a superconducting stripline: Static case*, Phys. C, 402 (2004), pp. 160–173.

[6] J. G. Caputo and L. Loukitch, *Dynamics of point Josephson junctions in microstrip line*, Physica, 425 (2005), pp. 69–89.

[7] J. G. Caputo, N. Flytzanis, Y. Gaididei, and M. Vavalis, *Two-dimensional effects in Josephson junctions:* I. *Static properties*, Phys. Rev. E, 54 (1996), pp. 2092–2101.

[8] J.-G. Caputo, N. Flytzanis, A. Tersenov, and E. Vavalis, *Analysis of a semilinear PDE for modeling static solutions of Josephson junctions*, SIAM J. Math. Anal., 34 (2003), pp. 1356–1379.

[9] J. G. Caputo, N. Flytzanis, and M. Vavalis, *A semi-linear elliptic pde model for the static solution of Josephson junctions*, Internat. J. Modern Phys. C, 6 (1995), pp. 241–262.

[10] R. Fehrenbacher, V. B. Geshkenbein, and G. Blatter, *Pinning phenomena and critical currents in disordered long Josephson junctions*, Phys. Rev. B, 45 (1992), p. 5450.

[11] M. A. Itzler and M. Tinkham, *Flux pinning in large Josephson junctions with columnar defects*, Phys. Rev. B, 51 (1995), p. 435.

[12] M. A. Itzler and M. Tinkham, *Vortex pinning by disordered columnar defects in large Josephson junctions*, Phys. Rev. B, 53 (1996), p. 11949.

[13] B. D. Josephson, *Possible new effects of superconductive tunneling*, Phys. Lett., 1 (1962), pp. 251–253.

[14] K. Likharev, *Dynamics of Josephson Junctions and Circuits*, Gordon and Breach, New York, 1986.

[15] *Maplesoft website*, http://www.maplesoft.com/.

[16] J. H. Miller, Jr., G. H. Gunaratne, J. Huang, and T. D. Golding, *Enhanced quantum interference effects in parallel Josephson junction arrays*, Appl. Phys. Lett., 59 (1991), pp. 3330–3332.

[17] M. Salez, Y. Delorme, I. Peron, B. Lecomte, F. Dauplay, F. Boussaha, J. Spatazza, A. Feret, J. M. Krieg, and K. F. Schuster, *A* 30% *bandwith tunerless SIS mixer of quantum-limited sensitivity for Herschell/HIFI band* 1, in Proceedings of the SPIE Conference on Telescopes and Astronomical Instrumentation, HI, 2002, T. G. Phillips and J. Zmuidzinas, eds., 2003, vol. 4855, pp. 402–414.

[18] A. V. Ustinov, M. Cirillo, B. H. Larsen, V. A. Oboznov, P. Carelli, and G. Rotoli, Phys. Rev. B, 51 (1995), p. 3081.

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 837–853

FIRST-ORDER CONTINUOUS MODELS OF OPINION FORMATION*

GIACOMO ALETTI†, GIOVANNI NALDI†, AND GIUSEPPE TOSCANI‡

Abstract. We study certain nonlinear continuous models of opinion formation derived from a kinetic description involving exchanges of opinion between individual agents. These models imply that the only possible final opinions are the extremal ones, and they are similar to models of pure drift in magnetization. Both analytical and numerical methods allow us to recover the final distribution of opinion between the two extremal ones.

Key words. nonlinear nonlocal hyperbolic equation, sociophysics, opinion formation, magnetization

AMS subject classifications. 91C20, 82B21, 60K35

DOI. 10.1137/060658679

1. Introduction.

This paper is devoted to the analysis and large-time behavior of solutions of the equation

$$\frac{\partial f}{\partial t} = \gamma \frac{\partial}{\partial x}\left((1-x^2)(x-m(t))f\right), \tag{1.1}$$

where the unknown $f(x,t)$ is a time-dependent probability density which may represent the density of opinion in a community of agents. This opinion varies between the two extremal opinions represented by ± 1, so that $x \in \mathcal{I} = [-1,1]$. The constant γ is linked to the spreading ($\gamma = -1$) or to the concentration ($\gamma = +1$) of opinions. In (1.1) $m(t)$ represents the mean value of $f(\cdot,t)$,

$$m(t) = \int_{\mathcal{I}} x\, f(x,t)\, dx, \tag{1.2}$$

and its presence introduces a nonlinear effect into its evolution.

When $\gamma = -1$, the related linear equation

$$\frac{\partial f}{\partial t} = -\frac{\partial}{\partial x}\left(x(1-x^2)f\right) \tag{1.3}$$

has been introduced recently by Slanina in [15] to analyze the evolution of density opinions in the voter model on a complete graph. There, the equation was derived as the mean field limit of the Sznajd model [20] in the case of two opinions. Because of linearity, (1.3) allows for an analytical treatment, and it is possible (see, e.g., [1, 15]) both to obtain the exact solution and to control the rate of decay towards the equilibrium for all values of γ (see also, e.g., [3]). Equation (1.1) was introduced in [22], in

*Received by the editors May 2, 2006; accepted for publication (in revised form) December 18, 2006; published electronically April 3, 2007. This work was supported by 2005 INDAM project "Kinetic Innovative Models for the Study of the Behavior of Fluids in Micro/Nano Electromechanical Systems," Italian M.I.U.R. project "Mathematical Problems of Kinetic Theories," and Italian M.I.U.R. project "Numerical Modelling for Scientific Computing and Advanced Applications."
http://www.siam.org/journals/siap/67-3/65867.html

†Department of Mathematics, University of Milano, via Saldini 50, 20133 Milano, Italy (giacomo.aletti@unimi.it, giovanni.naldi@mat.unimi.it).

‡Department of Mathematics, University of Pavia, via Ferrata 1, 27100 Pavia, Italy (giuseppe.toscani@unipv.it).

connection with the asymptotic limit of a Boltzmann equation for the kinetic description of opinion formation involving binary exchange of opinion between individual agents. This kinetic description is based on two-body interactions involving both compromise and diffusion properties in exchanges between individuals. Compromise and diffusion were quantified in [22] by two parameters, which are mainly responsible of the behavior of the model, and allow for a rigorous asymptotic analysis in which the limiting model is a Fokker–Planck-type equation, where the second-order term is related to diffusion, while the drift term is due to compromise. In a compromise-dominated regime, the resulting equation is exactly (1.1). We point out that, contrary to (1.3), the presence of the mean value $m(t)$ in (1.1) takes into account the influence of the mean opinion on the compromise-dominated dynamics.

Microscopic models of both social and political phenomena describing collective behaviors and self-organization in a society have been recently introduced and analyzed by several authors (see, e.g., [2, 6, 7, 8, 11, 12, 14, 16, 17, 18, 20, 25, 26]). The leading idea is that collective behaviors of a society composed of a sufficiently large number of individuals (agents) can be hopefully described using the laws of statistical mechanics as it happens in a physical system composed of many interacting particles. The details of the social interactions between agents then characterize the emerging statistical phenomena.

Equation (1.3), or in general (1.1) with $\gamma = -1$, can also describe a pure drift in magnetization (see, e.g., [19] and the references therein), where the two extremal points of $\mathcal{I}$ represent the opposite attraction poles. In the kinetic picture of [22], it simply means that the compromise in the binary interaction is substituted by magnetic repulsion between agents. Moreover, the case $\gamma = +1$ is related to models of one-dimensional nonlinear friction equations considered in the study of granular flows (see, e.g., [13, 21]), in connection with the quasi-elastic limit of a Boltzmann equation for rigid spheres with dissipative collisions and variable coefficient of restitution.

The paper is organized as follows. In the next section we introduce the main properties of the model, which justify the treatment in terms of a suitable weak formulation. The qualitative analysis is given in section 3. The large-time behavior is considered in sections 4–6. It is shown that the problem can be solved in sufficiently high generality only in the case of concentration ($\gamma = +1$). This lack of generality in the analytical treatment of the large-time behavior of the solution in the spreading of opinion justifies the numerical treatment of the equation. The numerical approximation is included in section 7, where both the explicit solution of (1.3) and the knowledge of the steady state in the concentration case ($\gamma = +1$) are used as benchmark tests for the numerical scheme. Finally, we note that the mathematical methods used here are close to the recent framework considered in the context of kinetic theory of nonlinear friction equations [10] and made popular by the mass transportation community [24].

2. Main properties and weak description. As briefly described in the introduction, (1.1) describes the evolution of a probability density which represents the density of opinions in a community. For all values of the constant γ, we will show that the time-evolution driven by this equation leads the density towards a equilibrium state that is described in terms of two Dirac masses ($\gamma = -1$) or to a unique Dirac mass ($\gamma = 1$). Having in mind that the equilibrium solution to equation (1.1) is given by Dirac masses, any convergence result towards equilibrium holds in a weak*-measure sense. The recent analysis of [10] of the nonlinear friction equation introduced by McNamara and Young [13] suggests that a suitable way of treating

(1.1) is based on a rewriting of this equation in terms of pseudoinverse functions. It is immediate to show that the drift operator on the right-hand side of (1.1) preserves positivity and mass,

$$\int_{\mathcal{I}} f(x,t)\,dx = \int_{\mathcal{I}} f_0(x)\,dx. \tag{2.1}$$

Then, given a initial datum which is a probability density (nonnegative and with unit mass), the solution remains a probability density at any subsequent time. Let $F(x)$ denote the probability distribution induced by the density $f(x)$,

$$F(x) = \int_{(-\infty,x]} f(y)\,dy \tag{2.2}$$

and let μ denote the distribution on $\mathbb{R}$ associated to F. Since $F(\cdot)$ is not decreasing, we can define its pseudoinverse function (also called quantile function) by setting, for $\rho \in (0,1)$,

$$X^{\mu}(\rho) = X^{F}(\rho) = \inf\{x : F(x) \geq \rho\}.$$

Equation (1.1) for $f(x,t)$ takes a simple form if written in terms of its pseudoinverse $X(\rho,t)$. Theorem 3.1 shows in fact that the evolution equation for $X(\rho,t)$ reads

$$\frac{\partial X(\rho,t)}{\partial t} = -\gamma\,(X(\rho,t) - m(t))\,(1 - X^2(\rho,t)), \tag{2.3}$$

where now $\rho \in (0,1)$. Note that if we assume F to be absolutely continuous with respect to x and strictly increasing, then Theorem 3.1 reduces to elementary computations. In (2.3)

$$m(t) = \int_0^1 X(\rho,t)\,d\rho. \tag{2.4}$$

Let us set $\gamma = -1$ (spreading). Then the weak form (2.3) clarifies the evolution of $X(\rho,t)$ and the role of $m(t)$. In fact, if $X(\rho,t) > m(t)$, $X(\rho,t)$ increases towards 1, while $X(\rho,t) < m(t)$ implies that $X(\rho,t)$ decreases towards -1. Hence, the mean opinion $m(t)$ represent a barrier for the density of opinions to move towards one of the two extremal opinions. The fact that the mean opinion varies with time makes the nonlinear problem harder to handle with respect to the linear problem considered in [15] where the barrier is fixed equal to zero.

Among the metrics which can be defined on the space of probability measures, which metricize the weak convergence of measures [27], one can consider the L^p-distance ($1 \leq p < \infty$) of the pseudoinverse functions

$$d_p(X,Y) = \Big(\int_0^1 |X(\rho) - Y(\rho)|^p\,d\rho\Big)^{1/p}. \tag{2.5}$$

In what follows, we'll use the usual identifications

$$d_p(X,Y) = d_p(f_X,f_Y) = d_p(F_X,F_Y) = d_p(\mu_X,\mu_Y)\,,$$

where μ_X (μ_Y), F_X (F_Y), and f_X (f_Y) denote the distribution, the cumulative function, and the density associated to X (Y), respectively. By this identification, as one can see [10, 23, 24], $d_2(F,G)$ is nothing but the Wasserstein metric [23].

In addition to nonlinear friction equations arising in the modeling of granular gases [10], the strategy of passing to pseudoinverse functions has been recently applied to nonlinear diffusion equations of porous medium type [5] and to degenerate convection–diffusion equations [4]. This rewriting of nonlinear diffusion equations has been shown to be useful in order to obtain simple explicit numerical schemes that satisfy a contraction property with respect to the Wasserstein metric [9].

3. Existence, uniqueness, and well-posedness of the problem. In this section we will study the initial value problem for (1.1), with initial density

$$f(x, t = 0) = f_0(x), \quad x \in \mathcal{I}. \tag{3.1}$$

As before, we will denote by $X_0(\rho)$ the quantile function corresponding to f_0, so that

$$X(\rho, t = 0) = X_0(\rho) = \inf\{x : F_0(x) \geq \rho\}, \quad \rho \in [0, 1]. \tag{3.2}$$

The equivalence between (1.1) and (2.3) is contained in the following.

THEOREM 3.1. *There exists a weak solution of* (1.1)–(3.1) *if and only if there exists a solution of* (2.3)–(3.2).

Proof. Suppose first that there exists $f(x, t)$ which solves (1.1)–(3.1). Then $m(t)$ is a differentiable function of time. Let $y(t)$ be the maximal C^1 solution of the Abel differential equation:

$$\begin{cases} y' = -\gamma(1 - y^2)(y - m(t)), \\ y(0) = \bar{y}_0, \end{cases} \tag{3.3}$$

where, for any $y_0 \in [-1, 1]$, we denoted by $\bar{y}_0$ a C^1-extension of y_0 to $\mathbb{R}$. We have, in weak sense,

$$\begin{aligned} \frac{d}{dt} \int_{(-\infty, y(t)]} f(x, t)\, dx &= \int_{\mathbb{R}} \Big[\frac{\partial}{\partial t} \Big(\mathbb{1}_{(-\infty, y(t)]}(x) \Big) f(x, t) + \mathbb{1}_{(-\infty, y(t)]}(x) \frac{\partial}{\partial t} f(x, t) \Big]\, dx \\ &= \int_{\mathbb{R}} \big(y'(t) \delta_{y(t)}(x) + \gamma[\delta_{y(t)}(x)(1 - x^2)(x - m(t))] \big) f(x, t)\, dx \\ &= 0\,. \end{aligned} \tag{3.4}$$

Since $X(\rho, t) \leq x \iff \rho \leq \int_{-\infty}^{x} f(y, t)\, dy$, the first part of the proof has been shown.

Now, let $X(\rho, t)$ be a solution of (2.3)–(3.2). As a consequence of the properties of the solution to Abel's equation (3.3), given any initial datum $X(\rho, 0)$ satisfying

- $X(\rho, 0) \in [-1, 1]$;
- $X(\rho, 0)$ is nondecreasing;
- $X(\rho, 0)$ is left-continuous,

the same properties are preserved at any subsequent time $t > 0$. Hence, for any t, $\{X(\rho, t),\ \rho \in (0, 1)\}$ is the quantile function of a unique probability measure on $[-1, 1]$. We have only to prove that (1.1) holds. This is a consequence of the change

of variables formula. In fact, if h is a test function,

$$\begin{aligned}\int_{\mathbb{R}} h(x)\frac{\partial}{\partial t}f(x,t)\,dx &= \frac{\partial}{\partial t}\int_{\mathbb{R}} h(x)f(x,t)\,dx\\ &= \frac{\partial}{\partial t}\int_0^1 h(X_\rho(t))\,d\rho\\ &= \int_0^1 \frac{\partial}{\partial t}h(X_\rho(t))\,d\rho\\ &= \int_0^1 h'(X_\rho(t))\Big(-\gamma\left(X(\rho,t)-m(t)\right)\left(1-X^2(\rho,t)\right)\Big)\,d\rho\\ &= \int_{\mathbb{R}} h'(x)\big(-\gamma\left(x-m(t)\right)\left(1-x^2\right)\big)f(x,t)\,dx\\ &= \int_{\mathbb{R}} h(x)\frac{\partial}{\partial x}\Big(\gamma\left(x-m(t)\right)\left(1-x^2\right)f(x,t)\Big)\,dx\,. \qquad \square\end{aligned}$$

We call $(\mathcal{K},p)$ the (compact) set of probability distributions on $[-1,1]$ equipped with the p-Wasserstein distance. Note that all the p-Wasserstein distances on $(\mathcal{K},d)$ are equivalent. In fact, if $q\geq p\geq 1$,

$$\|X^{\mu_1}-X^{\mu_2}\|_p \leq \|X^{\mu_1}-X^{\mu_2}\|_q \leq 2^{1-p/q}\|X^{\mu_1}-X^{\mu_2}\|_p^{p/q}\,. \tag{3.5}$$

We will refer to $\mathcal{K}$ as the topological space of probability distributions on $[-1,1]$ induced by any of this metric: the weak*-topology. Before searching for a continuous solution of (1.1) in $\mathcal{K}$, we state the following trivial lemma.

LEMMA 3.2 (solution Abel). *Let $\phi(x,y)=-\gamma(1-x^2)(x-y)$. Then*

$$|\phi(x_1,y_1)-\phi(x_2,y_2)|\leq 4|x_1-x_2|+|y_1-y_2|\,.$$

Moreover, if f is a solution of $f'(t)=\phi(f(t),g(t))$ with $\sup|g(t)|\leq 1$ and $f(0)\in[-1,1]$,

$$|f(s)-f(t)|\leq 2|s-t|\,.$$

We call any function $\mu\in C^0(\mathbb{R},\mathcal{K})$ s.t. (1.1)–(3.1) holds a *solution of* (1.1)–(3.1). We have the following theorem.

THEOREM 3.3. *For any probability density $f_0(x)$ in* (3.1)*, there exists a unique function $\mu\in C^0(\mathbb{R},\mathcal{K})$ such that if $f(x,t)$ denotes the weak derivative of the probability distribution $\mu(t)$, $f(x,t)$ satisfies* (1.1) *with initial value* (3.1)*. Moreover, for any $t\in\mathbb{R}$, the solution depends in a continuous way on the initial datum: the problem* (1.1)–(3.1) *is well-posed in $C^0(\mathbb{R},\mathcal{K})$.*

Proof. [Existence] We prove the existence of a solution of the equivalent problem (2.3)–(3.2) (see Theorem 3.1) in a constructive way. More precisely,

(A) we construct a sequence $\{X_n,\, n\in\mathbb{N}\}$ which approximates a target solution;
(B) by compactness arguments, we find a convergent subsequence $X_{n_l}\to X$;
(C) the limit X satisfies (2.3)–(3.2).

Let $[-T,T]$ be fixed. For any $n\in\mathbb{N}$, we subdivide $[-T,T]$ into disjoint intervals of length $R/2^N$. Then we proceed as follows:

(A1) we compute $m_0^{(n)}=\int_0^1 X_0(\rho)\,d\rho$;
(A2) we solve (2.3) on $[-T/2^n,T/2^n]$ with $m^{(n)}(t)=m_0$, finding $X^{(n)}(\rho,t)$, $t\in[-T/2^n,T/2^n]$;

(A3) for any $k = 1, \dots, 2^n - 1$,

- we compute
$$m^{(n)}_{\pm k} = \int_0^1 X^{(n)}(\rho, \pm kT/2^n)\, d\rho;$$
- we solve (2.3) on $(kT/2^n, (k+1)T/2^n]$ with $m^{(n)}(t) = m^{(n)}_k$ and initial data $X^{(n)}(\rho, kT/2^n)$, finding $X^{(n)}(\rho, t)$, $t \in (kT/2^n, (k+1)T/2^n]$;
- we solve (2.3) on $[-(k+1)T/2^n, -kT/2^n)$ with $m^{(n)}(t) = m^{(n)}_{-k}$ and initial data $X^{(n)}(\rho, -kT/2^n)$, finding $X^{(n)}(\rho, t)$, $t \in [-(k+1)T/2^n, -kT/2^n)$.

We call $\mu^{(n)} : [-T, T] \to \mathcal{K}$ the sequence of function with value in $\mathcal{K}$ associated to $X^{(n)}$.

(B) For any $n \in \mathbb{N}$, it is possible to prove (by induction on k) that for any $t \in [-T, T]$, $X^{(n)}(\rho, t) \in [-1, 1]$ and $m^{(n)}(t) \in [-1, 1]$. As a consequence of Lemma 3.2, we have

$$\left|X^{(n)}(\rho, s) - X^{(n)}(\rho, t)\right| \le 2|t - s|, \tag{3.6}$$

i.e., for any $\rho \in (0, 1)$, $\{X^{(n)}(\rho, \cdot) : [-T, T] \to [-1, 1]\}_{n\in\mathbb{N}}$ is a uniformly equicontinuous sequence. A diagonal argument together with the Ascoli–Arzelà theorem ensure the existence of a subsequence n_l s.t. $\{X^{(n_l)}(\rho, \cdot) : [-T, T] \to [-1, 1]\}_{l\in\mathbb{N}}$ converges uniformly on $[-T, T]$ for each $\rho \in (0, 1) \cap \mathbb{Q}$.

Now, (3.6) implies

$$\int_0^1 |X^{(n)}(\rho, s) - X^{(n)}(\rho, t)|\, d\rho \le 2|t - s|\,, \tag{3.7}$$

i.e., $\mu^{(n)}$ is a equicontinuous sequence with respect to the distance d_1 on $\mathcal{K}$. Then the Ascoli–Arzelà theorem again ensures the existence of a subsection of n_l (we call it n_l again) such that

$$\sup_{t\in[-T,T]} d_1(\mu^{(n_k)}(t), \mu^{(n_l)}(t)) \le M(k \wedge l) \underset{k\wedge l\to\infty}{\longrightarrow} 0, \tag{3.8}$$

$$\sup_{t\in[-T,T]} |X^{(n_k)}(\rho, t) - X^{(n_l)}(\rho, t)| \le N_\rho(k \wedge l) \underset{k\wedge l\to\infty}{\longrightarrow} 0 \qquad \forall \rho \in \mathbb{Q} \cap (0, 1). \tag{3.9}$$

Now, let $\rho \in (0, 1) \cap \mathbb{Q}$ be fixed. $\{X^{(n_l)}(\rho, \cdot)\}_{l\in\mathbb{N}}$ is a uniform convergent sequence of derivable functions converging to $X(\rho, t) = \lim_l X^{(n_l)}(\rho, t))$. Left-continuity and monotonicity of $\{X(\rho, t),\ \rho \in (0, 1) \cap \mathbb{Q}\}$ extend the definition of $X(\rho, t)$ to all $\rho \in (0, 1)$.

(C) What remains to prove is

- $\lim_l m^{(n_l)}(t) = \int_0^1 X(\rho, t)\, d\rho =: m(t)$;
- $X(\rho, t)$ is differentiable, and (2.3) holds.

Note that, from (3.8), it follows immediately that $|\int_0^1 X^{(n_k)}(\rho, t)\, d\rho - \int_0^1 X^{(n_l)}(\rho, t)\, d\rho| \le M(k \wedge h)$ and hence

$$\left|\int_0^1 X(\rho, t)\, d\rho - \int_0^1 X^{(n_l)}(\rho, t)\, d\rho\right| \le M(h)\,. \tag{3.10}$$

By definition of $m^{(n)}$,

$$m^{(n)}(t) = m^{(n)}(\llbracket 2^n t \rrbracket/2^n) = \int_0^1 X^{(n)}(\rho, \llbracket 2^n t \rrbracket/2^n)\, d\rho,$$

where $[\![\cdot]\!]$ is the integer part of $\cdot$ closer to 0. Therefore,

$$\left|m^{(n)}(t) - \int_0^1 X^{(n)}(\rho,t)\,d\rho\right| \le 2^{-n+1} \tag{3.11}$$

and hence $m^{(n_l)}(t)$ is a Cauchy sequence on $[-1,1]$. Thus, there exists $\widehat{m}(t) = \lim_h m^{(n_l)}(t)$. By (3.10) and (3.11) it follows that $\widehat{m}(t) = m(t)$ since

$$\left|\widehat{m}(t) - \int_0^1 X(\rho,t)\,d\rho\right| \le \left|\widehat{m}(t) - m^{(n_l)}(t)\right| + 2^{-n_l+1} + M(h)\,.$$

Now, let $\rho \in (0,1)\cap\mathbb{Q}$ be fixed. For simplicity of notation, define $H(n,\rho,t) := \frac{\partial}{\partial s}X^{(n)}(\rho,s)\big|_{s=t}$. Moreover, we define

$$H(\rho,t) := \lim_{l\to\infty} H(n_l,\rho,t) = -\gamma(1-X(\rho,t)^2)(X(\rho,t)-m(t)).$$

The uniform convergence theorem states that $\frac{\partial}{\partial s}X(\rho,s)\big|_{s=t} = H(\rho,t)$ if $\{H(n_l,\rho,t)\}_{l\in\mathbb{N}}$ is a uniform converging sequence on $(-T,T)$. To prove this, let $k \ge l$. The triangular inequality

$$\begin{aligned}\left|H(n_k,\rho,t) - H(n_l,\rho,t)\right| &\le \left|H(n_k,\rho,t) - H(n_k,\rho,[\![2^{n_l}t]\!]/2^{n_l})\right| \\ &\quad + \left|H(n_k,\rho,[\![2^{n_l}t]\!]/2^{n_l}) - H(n_l,\rho,[\![2^{n_l}t]\!]/2^{n_l})\right| \\ &\quad + \left|H(n_l,\rho,[\![2^{n_l}t]\!]/2^{n_l}) - H(n_l,\rho,t)\right| \\ &= A_\rho(k,l,t) + B_\rho(l,k,t) + A_\rho(l,l,t)\end{aligned}$$

shows that we may prove that $\sup_{t\in[-T,T]} A_\rho(l,k,t) + B_\rho(l,k,t) + A_\rho(l,l,t) \underset{l\wedge k\to\infty}{\longrightarrow} 0$. As a consequence of (3.7) and (3.11), we have $|m^{(n)}(s) - m^{(n)}(t)| \le 2^{-n+2} + 2|t-s|$, which implies (see Lemma 3.2 and (3.6))

$$\begin{aligned}|H(n,\rho,s) - H(n,\rho,t)| &= \Big|(1-X^{(n)}(\rho,s)^2)(X^{(n)}(\rho,s) - m^{(n)}(s)) \\ &\qquad - (1-X^{(n)}(\rho,t)^2)(X^{(n)}(\rho,t) - m^{(n)}(t))\Big| \\ &\le 4|X^{(n)}(\rho,s) - X^{(n)}(\rho,t)| + |m^{(n)}(s) - m^{(n)}(t)| \\ &\le 10|t-s| + 2^{-n+2},\end{aligned}$$

and hence $A_\rho(k,l,t) \le 10\cdot 2^{-n_l} + 2^{-n_k+2}$. Now, let $k \ge l$ and $l \in \{-2^{n_l}+1,\dots,2^{n_l}-1\}$. Again, Lemma 3.2, (3.8), and (3.9) imply

$$\left|H(n_k,\rho,l/2^{n_l}) - H(n_l,\rho,l/2^{n_l})\right| \le 4N_\rho(l) + M(l),$$

and hence $B_\rho(k,l,t) \le 4N_\rho(l) + M(l)$. This completes the proof for $\rho \in (0,1)\cap\mathbb{Q}$. Now, fixing $y_0 \in [-1,1]$, let $y(t)$ be the maximal C^1 solution of the Abel differential equation (3.3). Since $X(\rho,t) \ge y(t) \iff X(\rho,0) \ge y(0)$ for all $\rho \in (0,1)\cap\mathbb{Q}$, left-continuity and monotonicity of $\{X_\rho(t),\ \rho\in(0,1)\}$ extend the proof to $\rho \in (0,1)$.

[Uniqueness and well-posedness] Let $Y(\rho,t), X(\rho,t)$ be two solutions of (1.1)–(1.2). Denote by $m_X(t)$ and $m_Y(t)$ the mean values of X and Y, respectively, at time

t. Since $|m_Y(t) - m_X(t)| \le d_1(X,Y)$ we have (by Lemma 3.2 and (3.5))

$$\begin{aligned}(3.12)\quad \frac{d}{dt} d_2(\mu_X(t),\mu_Y(t)) &= \frac{d}{dt}\int_0^1 \big(Y(\rho,t)-X(\rho,t)\big)^2\,d\rho \\ &\le 2\int_0^1 \big|Y(\rho,t)-X(\rho,t)\big|\Big(4\big|Y(\rho,t)-X(\rho,t)\big| + \big|m_Y(t)-m_X(t)\big|\Big)\,d\rho \\ &\le 10\cdot d_2(\mu_X(t),\mu_Y(t))\,.\end{aligned}$$

Gronwall's lemma completes the proof. ∎

4. Large-time behavior of solutions. Thanks to the uniqueness of solutions of Abel's equation, we obtain the following lemma.

LEMMA 4.1. *For any $t \ne s \in \mathbb{R}$ and $\rho \ne \rho' \in (0,1)$, we have*

$$X(\rho,t) = X(\rho',t) \iff X(\rho,s) = X(\rho',s)\,,$$

i.e., (1.1) *does not create or destroy delta masses in finite time.*

A direct consequence of the previous lemma is that the initial masses in $+1$, -1, and $(-1,1)$ remain unchanged in time. Let us call them p_{+1}, p_{-1}, and $1-(p_{+1}+p_{-1})$, respectively.

An important argument linked to the large-time behavior of solutions to nonlinear equations is the study both of conservation laws and of Lyapunov functionals. In addition to mass conservation, a second conserved quantity (when defined) is furnished by

$$(4.1)\qquad T(t) := \int_0^1 \log\Big(\frac{1+X(\rho,t)}{1-X(\rho,t)}\Big)\,d\rho\,.$$

In addition to the conservation of both mass and $T(t)$, equation (1.1) possesses a Lyapunov functional, simply given by the variance of the solution

$$(4.2)\qquad V(t) := \int_0^1 (X(\rho,t))^2\,d\rho - \Big(\int_0^1 X(\rho,t)\,d\rho\Big)^2.$$

We give below an easy-to-check condition which ensure both the the boundedness and the conservation in time of the functional (4.1).

LEMMA 4.2. *Let $\log\big((1+X(\rho,0))/(1-X(\rho,0))\big) \in L^1(0,1)$. Then for all $t \in \mathbb{R}$*

$$T(t) := \int_0^1 \log\Big(\frac{1+X(\rho,t)}{1-X(\rho,t)}\Big)\,d\rho$$

is well-defined. Moreover, $T(t)$ is differentiable and $T'(t) = 0$.

Proof. Since $\log\big((1+X(\rho,0))/(1-X(\rho,0))\big) \in L^1(0,1)$, then $X(\rho,0) \in (-1,1)$ for all $\rho \in (0,1)$. Now $(1-x^2)(x-1) \le (1-x^2)(x-m(t)) \le (1-x^2)(x-1)$, and hence the uniqueness of solutions of the Abel equations imply

$$X(\rho,t) \in (-1,1) \qquad \forall \rho \in (0,1)\,, \quad \forall t \in \mathbb{R}\,.$$

Let $G(\rho,t) = \log\big((1+X(\rho,t))/(1-X(\rho,t))\big)$. Since $|G_t(\rho,t)| \le 2$, we have

- $\exists G_t$ for all $\rho \in (-1,1)$, for all $t \in [-T,T]$;
- $|G(\rho,t)| \le |G(\rho,0)| + 2T$ and hence $T(t)$ exists;
- $|G_t(\rho,t)| \le 2$ and $2 \in L^1(0,1)$;

and hence it is possible to differentiate under the integral sign, obtaining $T'(t) = 0$. ∎

The following lemma shows that the variance is a Lyapunov functional for (1.1).

LEMMA 4.3. *The variance $V(t)$ of $\mu(t)$ is a monotone differentiable function with values in $[0,1]$.*

Proof. $V(t)$ is clearly differentiable, and

$$V'(t) = \frac{d}{dt}\int_0^1 (X_t(\rho) - m(t))^2\, d\rho = -2\gamma \int_0^1 (1 - X_t(\rho)^2)(X_t(\rho) - m(t))^2\, d\rho\,.$$

In the case $\gamma = -1$, $V(t)$ is monotonically increasing while bounded from above by 1. In fact, the maximum value of V is attained for $\mu = (\delta_1 + \delta_{-1})/2$. If, on the contrary, $\gamma = 1$, $V(t)$ is monotonically decreasing while bounded from below by 0. In this second case, the minimum value of V is attained for $\mu = \delta_a$, $a \in [-1,1]$. ∎

The most difficult problem linked to (1.1) is the study of the evolution of the mean $m(t)$ and to the exact evaluation of its limit value $\bar{m} = \lim_{t\to\infty} m(t)$. The knowledge of $\bar{m}$ is of primary importance, since in consequence of the structure of the limit state of the solution to (2.3), this value is enough to characterize completely the steady state.

Remark. The difficulty comes out from the evolution of the mean $m(t)$, which is given by the "nonclosed" equation

$$m'(t) = -\gamma m(t)\int_0^1 X^2(\rho)\, dr + \gamma \int_0^1 X^3(\rho)\, dr. \tag{4.3}$$

In what follows, we make use of the previous results to extract information on the behavior of the mean $m(t)$.

LEMMA 4.4. *For any t, $m(t) \in [-\sqrt{1 - V(t)}, \sqrt{1 - V(t)}]$.*

Proof. Since $V(t) = \int x^2 f(x,t)dx - (m(t))^2$, then $V(t) \le 1 - (m(t))^2$. ∎

LEMMA 4.5. *The function $m' : \mathbb{R} \to [-1,1]$ belongs to $L^2(\mathbb{R})$. Moreover, $\exists \lim_{t\to\infty} m'(t) = 0$.*

Proof. It is sufficient to note that

$$(m'(t))^2 \le \int \left(1 - X_\rho(t)^2\right)^2 (X_\rho(t) - m(t))^2\, d\rho \le V'(t),$$

and hence $m' \in L^2(\mathbb{R})$ by Lemma 4.3. Moreover, since m' is a Lipschitz function (in fact it is differentiable and $m'' \le 10$; cf. (4.3)), it follows that $\lim_{t\to\infty} m'(t) = 0$. ∎

5. Spreading of opinions. In this section we study the large behavior of (1.1) when $\gamma = -1$, leaving the opposite case $\gamma = 1$ to the following section.

Remark. If $X(\rho,0) = -X(1-\rho,0)$ (i.e., the initial distribution is symmetric), then from (4.3) it follows that $m(t) = 0$ for any subsequent time $t > 0$. If, in addition, $X(\rho_1,0) = X(\rho_1 + \delta, 0) = 0$, then $X(\rho_1,t) = X(\rho_1+\delta,t) = 0$ for all $t > 0$ (i.e., any initial mass in 0 is not moved away in time if the initial distribution is symmetric). In order to avoid these situations we will allow delta masses only in ± 1: $X(\rho_1,0) = X(\rho_2,0) \iff \rho_1 = \rho_2$ or $(X(\rho_1,0))^2 = 1$.

THEOREM 5.1. *Assume $X(\rho_1,0) = X(\rho_2,0) \iff \rho_1 = \rho_2$ or $(X(\rho_1,0))^2 = 1$. Then the limit distribution exists and is given by two masses located in -1 and $+1$.*

Proof. By Lemmas 4.3 and 4.4, $m(t) \in [-\sqrt{1 - V(0)}, \sqrt{1 - V(0)}]$ for all $t \ge 0$. Thus, if $(X(\rho,t_0)^2 > 1 - V(0)$, then $(X(\rho,t)^2 > 1 - V(0)$ for all $t \ge t_0$. Since

$X(\rho,t) < -\sqrt{1-V(0)} \Rightarrow \frac{\partial}{\partial t}X(\rho,t) \le 0$ and $X(\rho,t) > \sqrt{1-V(0)} \Rightarrow \frac{\partial}{\partial t}X(\rho,t) \ge 0$ by Lemmas 4.3 and 4.4, the two functions

$$p^h_{-1}(t) = \sup\{\rho \in (0,1)\colon X(\rho,t) < -\sqrt{1-V(0)}\} = \int_{[-1,-\sqrt{1-V(0)})} f(x,t)\,dx,$$
$$p^h_{1}(t) = \inf\{\rho \in (0,1)\colon X(\rho,t) > \sqrt{1-V(0)}\} = \int_{(\sqrt{1-V(0)},1]} f(x,t)\,dx$$

are monotone. We call $p^h_{\pm 1} = \lim_{t\to\infty} p^h_{\pm 1}(t)$. Equation (2.3) and monotonicity of Abel's solutions allow us to state that

$$\forall \rho \in [0, p^h_{-1}), \quad \lim_{t\to\infty} X(\rho,t) = -1, \tag{5.1}$$

$$\forall \rho \in (p^h_{+1}, 1], \quad \lim_{t\to\infty} X(\rho,t) = 1. \tag{5.2}$$

Hence the limit distribution has two masses in -1 and $+1$. It remains to characterize what happens for the remaining $p^h_{+1} - p^h_{-1}$ mass. Let us recall that

$$X(\rho,t) \in \big[-\sqrt{1-V(0)}, \sqrt{1-V(0)}\big] \qquad \forall \rho \in (p^h_{-1}, p^h_{+1}). \tag{5.3}$$

By Lemma 4.5, there exists $T > 0$ such that $|m'(t)| \le (V(0)/2)^2$ for all $t > T$. By contradiction, suppose that there exist $t_0 \ge T$ and $\rho \in (p^h_{-1}, p^h_1)$ such that $\big|X(\rho,t_0) - m(t_0)\big| > V(0)/4$.

Since $|\frac{\partial}{\partial t}X(\rho,t_0)| > |m'(t_0)|$, it follows that $\big|X(\rho,t) - m(t)\big| > V(0)/4$ for all $t \ge t_0$. Thus, (5.3) shows the contradiction:

- $\frac{\partial}{\partial t}X(\rho,t)$ is continuous;
- $|\frac{\partial}{\partial t}X(\rho,t)| > (V(0)/2)^2$ if $t \ge t_0$;
- $X(\rho,t)$ is bounded.

Therefore,

$$\big|X(\rho,t) - m(t)\big| \le \frac{V(0)}{4} \qquad \forall \rho \in (p^h_{-1}, p^h_{+1}), \quad t > T\,. \tag{5.4}$$

Now, let $F(x,y) = (1-x^2)(x-y)$ as in Lemma 3.2. Since F is differentiable, when $x_1 \ge x_2$, Lagrange theorem states that we can find $\xi \in (x_1, x_2)$ such that

$$F(x_1,y) - F(x_2,y) = (x_1 - x_2)\frac{\partial}{\partial x}F(x,y)\Big|_{x=\xi}.$$

Now, if $1 - x^2 \ge V(0)$ and $|x-y| \le V(0)/4$, we have

$$\frac{\partial}{\partial x}F(x,y) = 1 - 3x^2 + 2xy \ge V(0) + 2x(y-x) \ge \frac{V(0)}{2},$$

that is,

$$F(x_1,y) - F(x_2,y) \ge (x_1 - x_2)\frac{V(0)}{2}\,, \qquad x_i^2 \le 1 - V(0) \text{ and } |x_i - y| \le \frac{V(0)}{4}\,. \tag{5.5}$$

Let $p^h_{-1} < \rho_2 \le \rho_1 < p^h_{+1}$. Then both (5.3) and (5.4) are satisfied for all $t > T$, (5.5) holds, and

$$\frac{\partial}{\partial t}\Big(X(\rho_1,t) - X(\rho_2,t)\Big) \ge \big(X(\rho_1,t) - X(\rho_2,t)\big)\frac{V(0)}{2} \qquad \forall t > T\,.$$

Since the two solutions are bounded, the only possibility is that $X(\rho_1,t)=X(\rho_2,t)$ for all $t>T$ and for all (ρ_1,ρ_2): $p^h_{-1}<\rho_2\le\rho_1<p^h_{+1}$, which implies $p^h_{-1}=p^h_1$ by Lemma 4.1 and hypothesis. □

Remark. Theorem 5.1 may be read in terms of weak*-measure convergence:

$$f(x,t)\underset{t\to\infty}{\rightharpoonup} p^h_{-1}\delta_{-1}(x)+(1-p^h_{-1})\delta_1(x)\,.$$

In particular, since the support is compact, all the moments exist and will converge. We have the following

COROLLARY 5.2. *Assume* $X(\rho_1,0)=X(\rho_2,0)\iff\rho_1=\rho_2$ *or* $(X(\rho_1,0))^2=1$. *Then there exists* $\lim_{t\to+\infty}m(t)=m_\infty=1-2p^h_{-1}$.

6. Concentration of opinions. Let us recall that the stochastic partial order is naturally given on $\mathcal{K}$. Let $F(x),G(x)$ denote two probability distributions and X_F,X_G their pseudoinverse functions, respectively. We say that $F\preceq G$ if $F(x)\ge G(x)$ for all (x) or, equivalently, if $X_F(\rho)\le X_G(\rho)$ for all $\rho\in(0,1)$.

LEMMA 6.1. *The operator*

$$\phi(X)=-(X-m(X))(1-X^2)$$

is a monotone operator with respect to the stochastic ordering.

Proof. Assume that $X_1(\rho,s)\le X_2(\rho,s)$ for all $\rho\in(0,1)$. Then $m_1(s)\le m_2(s)$ (they are equal if and only if the distributions coincide). Let $\rho\in(0,1)$ be fixed. If $X_1(\rho,s)=X_2(\rho,s)$, then $X_1'(\rho,s)\le X_2'(\rho,s)$. The continuity of X' is sufficient for the remaining part of the proof. □

LEMMA 6.2. *Let* X_0 *in* (3.2) *be given. Then there exists* $\lim_{t\to+\infty}m(t)=m_\infty$.

Proof. Let $[a,b]$ be the class limit of $m(t)$. Suppose $a=-1$, i.e., $\liminf_t m(t)=-1$. Markov inequality then shows that the limit distribution is a mass concentrated in -1, and hence $b=-1$. Otherwise, we may assume that $m(t)\in[-1+\delta,\,1-\delta]$ for all $t\ge t_0$ and let p_0 be the mass not concentrated in ±1 at each time (recall Lemma 4.1). For all $\epsilon>0$, $p_0-\epsilon$ mass is in $[-1+\epsilon,\,1-\epsilon]$ at $t=t_0$. Therefore, for all $\rho\in(0,1)$, $X(\rho,t_0)\in[-1+\epsilon,\,1-\epsilon]$, and $X(\rho,t)$ decays exponentially to $m(t)$ with rate not less than $(\min(\delta,\epsilon))^2$. The large behavior of this process shows three delta masses: the initial two in ±1 and the remaining one in $m(t)$. Stationary arguments imply the existence of m_∞. □

The steady state of the process can now be defined by the following theorem.

THEOREM 6.3. *If* $(1-p_1)(1-p_{-1})<1$ *(i.e., if there are masses in* ±1 *at time* $t=0$*), then* $m_\infty=p_1-p_{-1}$. *Otherwise, if* $\log\big((1+X(\rho,0))/(1-X(\rho,0))\big)\in L^1(0,1)$, *then*

$$m_\infty=\frac{\exp\{T(0)\}-1}{\exp\{T(0)\}+1}. \tag{6.1}$$

Proof. The first part is a consequence of Lemma 6.2 and stationary arguments. The second part is a consequence of Lemma 4.2, since

$$\int_0^1\log\Big(\frac{1+X(\rho,0)}{1-X(\rho,0)}\Big)\,d\rho=\int_0^1\log\Big(\frac{1+X(\rho,t)}{1-X(\rho,t)}\Big)\,d\rho\underset{t\to\infty}{\longrightarrow}\log\Big(\frac{1+m_\infty}{1-m_\infty}\Big)\,,$$

the last limit being true by Lemma 6.2. □

Remark. Lemma 6.1 allows us to extend the previous result to cases where at least one of the two functions $\log(1\pm X(\rho,0))$ is integrable. If, for example,

$\log(1+X(\rho,0)) \in L^1(0,1)$ and $\log(1-X(\rho,0)) \notin L^1(0,1)$, if we take $X^{(n)}(\rho,0) = \min\{X(\rho,0), 1-1/n\}$, then $X^{(n)}(\rho,t) \le X(\rho,t)$ for all $t \ge 0$, for all $\rho \in (0,1)$. The monotone convergence theorem states that $\lim_n T^{(n)}(0) = +\infty$, i.e., $\lim_n m_\infty^{(n)} = 1$. Thus, by the monotonicity argument of Lemma 6.1, $m_\infty = 1$.

With this remark in mind, we now show a "counterintuitive" example. Let

$$f_0(x) = \begin{cases} \frac{1-\epsilon}{\epsilon} & \text{if } -1 < x < -1+\epsilon, \\ \frac{1}{1-x}\left(\frac{\epsilon}{1-\epsilon\log(1-x)}\right)^2 & \text{if } 0 < x < 1, \end{cases}$$

and hence

$$F_0(x) = \begin{cases} 0 & \text{if } x < -1, \\ \frac{1-\epsilon}{\epsilon}(1+x) & \text{if } -1 \le x < -1+\epsilon, \\ 1-\epsilon & \text{if } -1+\epsilon \le x < 0, \\ 1 - \frac{\epsilon}{1-\epsilon\log(1-x)} & \text{if } 0 \le x < 1, \\ 1 & \text{if } 1 \le x, \end{cases}$$

which corresponds to

$$X_0(\rho) = \begin{cases} -1 + \frac{\epsilon}{1-\epsilon}\rho & \text{if } 0 < \rho \le 1-\epsilon, \\ 1 - \exp(-\frac{1}{1-\rho} + \frac{1}{\epsilon}) & \text{if } 1-\epsilon < \rho < 1. \end{cases}$$

With this data, $\log(1+X(\rho,0)) \in L^1(0,1)$ but $\log(1-X(\rho,0)) \notin L^1(0,1)$; the $1-\epsilon$ initial mass is close to -1, while the asymptotic solution is δ_1.

7. Numerical examples. The analysis of the previous section left open the problem of the identification of the steady state in the case of the spreading of opinions. Here results can be achieved only by numerical simulation of the spreading process. To test the numerical method, we will first derive the (explicit) solution to the pure drift linear equation of spreading considered in [15] as the mean field limit of the Sznajd model [20]. This equation reads

$$\frac{\partial f}{\partial t} = \gamma \frac{\partial}{\partial x}\left(x(1-x^2)f\right), \tag{7.1}$$

namely, (1.1) without the presence of the mean $m(t)$. In terms of the quantile function $X(\rho,t)$, equation (7.1) takes the form

$$\frac{\partial X(\rho,t)}{\partial t} = -\gamma X(\rho,t)(1-X^2(\rho,t)). \tag{7.2}$$

Let us set $\gamma = -1$ (spreading), and let $X_0(\rho)$ denote the initial datum. For any given $\rho \in [0,1]$, equation (7.2) is an ordinary differential equation which can be easily integrated to give

$$X(\rho,t) = \frac{X_0(\rho)e^t}{\left(1 - X_0^2(\rho) + X_0^2(\rho)e^{2t}\right)^{1/2}}. \tag{7.3}$$

The asymptotic behavior of (7.2) can be easily deduced from the explicit solution. In fact, the solution converges exponentially in time to -1 if $X_0(\rho) < 0$, while it converges

to $+1$ if $X_0(\rho) > 0$. Solution (7.3) can be inverted by using the definition of $X(\rho, t)$. Let $F_0(x)$ $x \in \mathcal{I}$ be the initial distribution function; then, since $X_0(F_0(x)) = x$,

$$X(F_0(x), t) = \frac{xe^t}{(1 - x^2 + x^2 e^{2t})^{1/2}}. \tag{7.4}$$

Thus, since the function on the right of (7.4) is increasing with respect to the variable x, we can invert it to obtain

$$X\left(F_0\left(\frac{y}{((1-y^2)e^{2t} + y^2)^{1/2}}\right), t\right) = y. \tag{7.5}$$

Finally, (7.5) implies

$$F(y, t) = F_0\left(\frac{y}{((1-y^2)e^{2t} + y^2)^{1/2}}\right). \tag{7.6}$$

Differentiating with respect to y, we conclude that if $f_0(x)$, $x \in \mathcal{I}$, is an initial density for (7.1), the solution in time is given by

$$f(x, t) = \frac{e^{2t}}{((1-x^2)e^{2t} + x^2)^{3/2}} f_0\left(\frac{x}{((1-x^2)e^{2t} + x^2)^{1/2}}\right). \tag{7.7}$$

The behavior of (7.7) shows the formation of two peaks in correspondence to the extremal points ± 1, while in all other points of the interval $\mathcal{I}$ there is exponential decay to zero.

Using the same procedure as above, we can easily solve the problem in the opposite case of concentration, where $\gamma = 1$. In this case, if $X_0(\rho)$ denote the initial datum,

$$X(\rho, t) = \frac{X_0(\rho)e^{-t}}{(1 - X_0^2(\rho) + X_0^2(\rho)e^{-2t})^{1/2}}. \tag{7.8}$$

The solution now converges exponentially in time to zero, except for ρ values for which $X_0(\rho) = \pm 1$, where it remains constant. In the original formulation, the solution $f(x, t)$ corresponding to the initial density $f_0(x)$, $x \in \mathcal{I}$, is

$$f(x, t) = \frac{e^{-2t}}{((1-x^2)e^{-2t} + x^2)^{3/2}} f_0\left(\frac{x}{((1-x^2)e^{-2t} + x^2)^{1/2}}\right). \tag{7.9}$$

Note that except for $x = 0$, $f(x, t)$ converges exponentially to zero. If $f_0(0) > 0$, the solution shows the formation of a peak in $x = 0$.

We perform numerical simulation for different initial data in the general nonlinear case. First, we assume an initial symmetric datum as a benchmark (see Figure 7.1), where

$$f_0(x) = \begin{cases} c_0(1 - x^2)(0.64 - x^2)^{1.3} & \text{if } |x| < 0.8; \\ 0 & \text{otherwise.} \end{cases}$$

In this case we have $m(t) = 0$ for all time t. Then we know the exact solution in order to perform a comparison with numerical results. In Figure 7.1 we show the behavior

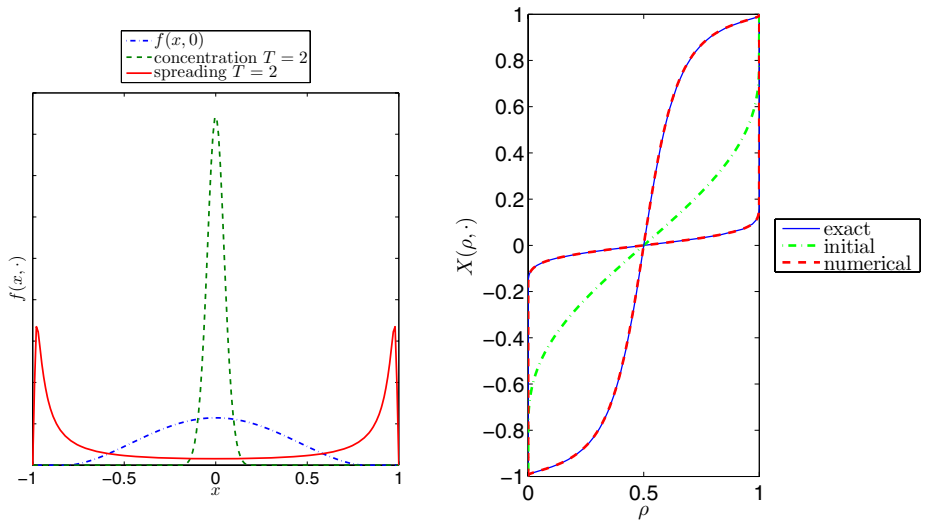


FIG. 7.1. *Benchmark case: evolution of density function (left) and comparison between analytical and numerical solutions for the quantile function (right).*

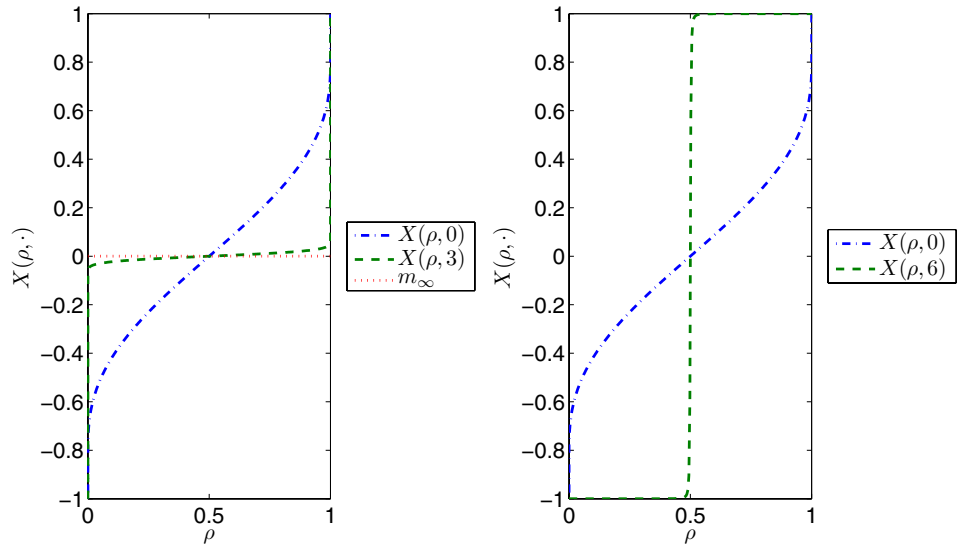


FIG. 7.2. *Plots of the behavior of a numerical solution with symmetric initial data for the concentration case (left) and for the spreading case (right).*

of an exact solution and a numerical one. The last one is obtained by using standard stiff Runge–Kutta methods, which is justified by our theoretical constructive result stated in the proof of Theorem 3.3. As one can see in Figure 7.1 we have a good agreement between the analytical and numerical solutions.

In Figure 7.2 we sketch the plot of the quantile function $X(\rho,t)$ for different times t in both the concentration and the spreading case with the same initial symmetrical data. As expected, in the concentration case, the limit values of all quantiles numerically converge to $m_\infty = 0$, while in the spreading case the quantiles converge to -1

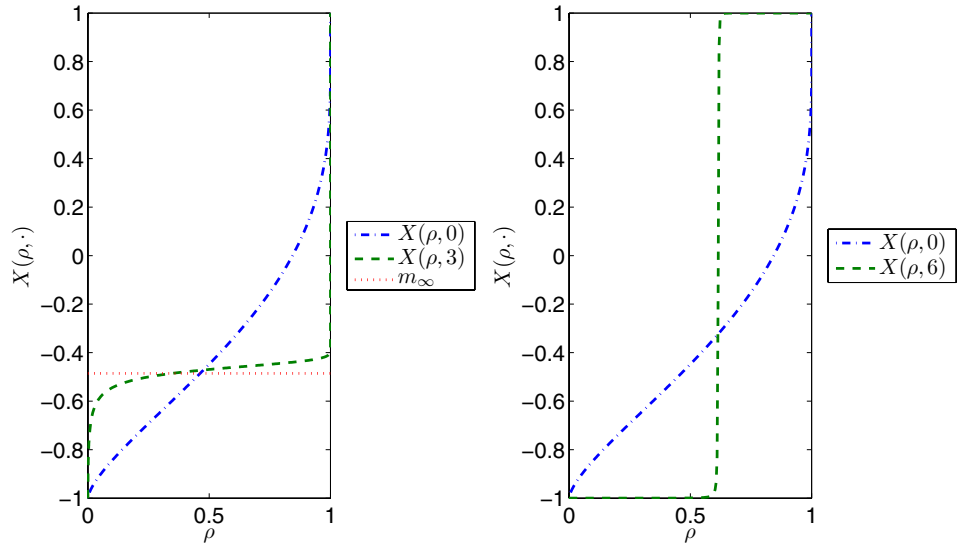


FIG. 7.3. *Plots of the behavior of a numerical solution with nonsymmetric initial data for the concentration case (left) and for the spreading case (right).*

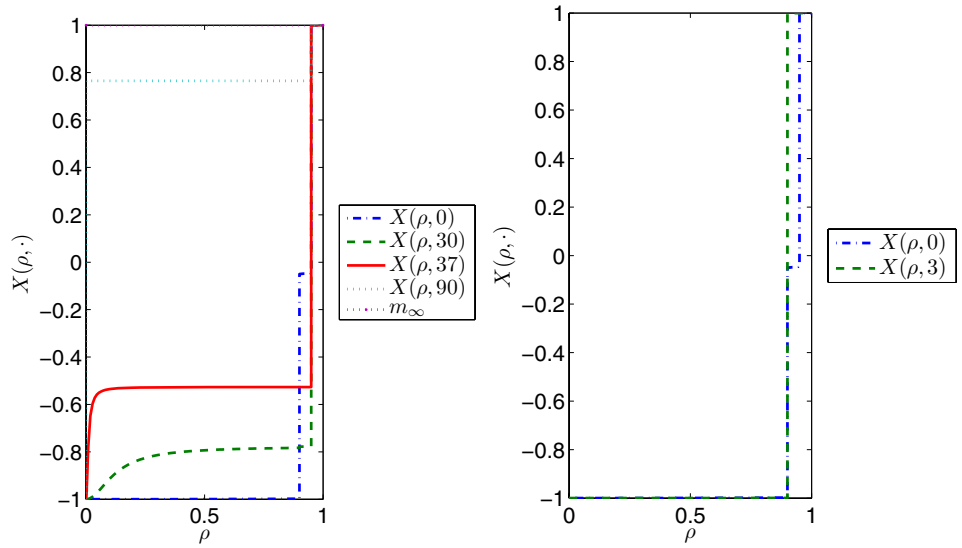


FIG. 7.4. *Plots of the behavior of a numerical "counterintuitive" solution: the concentration case tends to δ_1 (left), while the $1-2\epsilon$ mass goes to -1 in the spreading case (right).*

if $\rho < .5$, and to 1 if $\rho > .5$.

In Figure 7.3 we show an asymmetric case. Now, in the concentration case $m_\infty \neq 0$ given in (6.1) and quantiles converge to it. In the spreading case we can numerically estimate the value p^h_{-1} (see (5.1)) for which $\rho < p^h_{-1}$ implies $X(\rho,t) \to -1$, $\rho > p^h_{-1}$ implies $X(\rho,t) \to 1$.

Finally, in Figure 7.4 we show a numerical "counterintuitive example" analogous to the example given at the end of the previous section. Here, the initial datum con-

sists of the $(1-2\epsilon)$ mass very close to -1, the ϵ mass close to 0, and the remaining ϵ mass concentrated in $+1$. In the concentration case, $\gamma = 1$, the mass goes asymptotically to $+1$, while in the spreading case, $\gamma = -1$, only 2ϵ goes asymptotically to $+1$. We point out that the spreading case is really counterintuitive. In fact, if we take $X(\rho, T)$, with $T \gg 1$, as solution of the concentration case as initial data for the spreading one, the dynamic is as follows. Initially, the $(1-\epsilon)$ mass decays; then it splits into two parts: the right one goes to $+1$, the left one goes to -1. The initial data can be chosen as close as we want to the distribution δ_1. Then, starting from this data (for which $T(t) = +\infty$; see (6.1)), a $(1-2\epsilon)$ mass will reach -1. We note that the asymptotic state for spreading phenomena seems unpredictable for such concentrated initial data.

Remark. The method we used to solve (7.1) represents in various cases a possible alternative to better known methods (like the method of characteristics) able to reckon the solution to one-dimensional first-order partial differential equations of the form

$$\frac{\partial f}{\partial t} = \frac{\partial}{\partial x}\left(\phi(x) f\right). \tag{7.10}$$

Our analysis is possible in all cases where the ordinary differential equation

$$\frac{dX}{dt} = -\phi(X)$$

is explicitly solvable.

8. Conclusions. We investigated in this paper the spreading and/or the concentration of opinion in an organized society by means of a first-order nonlinear partial differential equation recently introduced in [22]. The presence of the nonlinearity renders it difficult to treat the spreading case analytically, and suitable numerical methods were discussed that are able to capture the large-time behavior of the solution in this case. This work represents a first attempt for a continuous approach to the formation of opinion in a community of agents. More complete models can be obtained by considering in addition the (linear or nonlinear) diffusion, which allows for a continuous steady state distribution function. Related problems in the presence of diffusion are presently under study.

REFERENCES

[1] F. Schweitzer, ed., *Modeling Complexity in Economic and Social Systems*, World Scientific, River Edge, NJ, 2002.

[2] T. Antal and P. L. Krapivsky, *Dynamics of social balance on networks*, Phys. Rev. E (3), 72 (2005), article 036121.

[3] E. Ben-Naim, *Opinion dynamics: Rise and fall of political parties*, Europhys. Lett., 69 (2005), pp. 671–677.

[4] J. A. Carrillo and K. Fellner, *Long-time asymptotics via entropy methods for diffusion dominated equations*, Asymptot. Anal., 42 (2005), pp. 29–54.

[5] J. A. Carrillo, M. P. Gualdani, and G. Toscani, *Finite speed of propagation in porous media by mass transportation methods*, C. R. Math. Acad. Sci. Paris, 338 (2004), pp. 815–818.

[6] G. Deffuant, F. Amblard, and G. Weisbuch, *Persuasion dynamics*, Phys. A, 353 (2005), pp. 555–575.

[7] S. Galam, Y. Gefen, and Y. Shapir, *Sociophysics: A new approach of sociological collective behavior*, J. Math. Sociology, 9 (1982), pp. 1–13.

[8] S. Galam and J.-D. Zucker, *From individual choice to group decision-making*, Phys. A, 287 (2000), pp. 644–659.

[9] L. Gosse and G. Toscani, *Identification of asymptotic decay to self-similarity for one-dimensional filtration equations*, SIAM J. Numer. Anal., 43 (2006), pp. 2590–2606.
[10] H. Li and G. Toscani, *Long-time asymptotics of kinetic models of granular flows*, Arch. Ration. Mech. Anal., 172 (2004), pp. 407–428.
[11] T. M. Liggett, *Stochastic Interacting Systems: Contact, Voter, and Exclusion Processes*, Springer-Verlag, Berlin, 1999.
[12] M. Marsili, F. Vega-Redondo, and F. Slanina, *The rise and fall of a networked society: A formal model*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 1439–1442.
[13] S. McNamara and W. R. Young, *Kinetics of a one-dimensional granular medium in the quasi–elastic limit*, Phys. Fluids A, 5 (1993), pp. 34–45.
[14] R. Ochrombel, *Simulation of Sznajd sociophysics model with convincing single opinions*, Internat. J. Modern Phys. C, 12 (2001), pp. 1091–1091.
[15] F. Slanina and H. Lavička, *Analytical results for the Sznajd model of opinion formation*, Eur. Phys. J. B, 35 (2003), pp. 279–288.
[16] D. Stauffer, *Percolation and Galam theory of minority opinion spreading*, Internat. J. Modern Phys. C, 13 (2002), pp. 975–977.
[17] D. Stauffer and P. M. C. de Oliveira, *Persistence of opinion in the Sznajd consensus model: Computer simulation*, Eur. Phys. J. B, 30 (2002), pp. 587–592.
[18] D. Stauffer, A. O. Sousa, and S. M. de Oliveira, *Generalization to square lattice of Sznajd sociophysics model*, Internat. J. Modern Phys. C, 11 (2000), pp. 1239–1245.
[19] M. D. Stiles, J. Xiao, and A. Zangwill, *Phenomenological theory of current-induced magnetization precession*, Phys. Rev. B, 69 (2004), 054408.
[20] K. Sznajd-Weron and J. Sznajd, *Opinion evolution in closed community*, Internat. J. Modern Phys. C, 11 (2000), pp. 1157–1165.
[21] G. Toscani, *One-dimensional kinetic models of granular flows*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 1277–1291.
[22] G. Toscani, *Kinetic models of opinion formation*, Commun. Math. Sci., 4 (2006), pp. 481–496.
[23] L. N. Wasserstein, *Markov processes on countable product space describing large systems of automata*, Probl. Pered. Inform., 5 (1969), pp. 64–73 (in Russian).
[24] C. Villani, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, AMS, Providence, RI, 2003.
[25] W. Weidlich, *Sociodynamics: A Systematic Approach to Mathematical Modelling in the Social Sciences*, Harwood Academic, Amsterdam, 2000.
[26] F. Wu and B. A. Huberman, *Social Structure and Opinion Formation*, http://arxiv.org/cond-mat/0407252 (2004).
[27] V. M. Zolotarev, *Probability metrics*, Theory Probab. Appl., 28 (1983), pp. 278–302.

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 854–868

ON A MODEL OF FLAME BALL WITH RADIATIVE TRANSFER*

VINCENT GUYONNE† AND PASCAL NOBLE‡

Abstract. In this paper, we derive an equation for the growth of a flame ball for a free boundary combustion model with radiative transfer. The equation for the radiative field is given by the linearized Eddington equation. We then study the mathematical properties of this equation of growth and carry out numerical computations in order to discuss the stability or instability of steady flame balls.

Key words. flame ball, radiative transfer, integro-differential equation, numerical quenching

AMS subject classifications. 80A25, 45K05, 34E05

DOI. 10.1137/060659612

1. Introduction.

Spherical flame balls have been found to exist as stable objects for small enough Lewis number during some experiments carried out at microgravity [13, 14]. Since the work of Zeldovich [19], "adiabatic" flame balls are unstable to one-dimensional radial perturbations; a stabilizing effect has to be identified. It has been argued [12] that radiation is physically important in near limit combustion at low gravity. Then, it is natural to consider a heat loss mechanism through radiation as a stabilizing effect. Moreover, it is worth noting that halon (CF_3Br) is added to experimental mixtures (to increase the luminosity of flame balls), which augments the radiation through soot formation.

Buckmaster, Joulin, and Ronney [5, 6] proposed different models to take into account the heat loss through radiation. They first considered constant heat losses in the burnt gases [5]: when the heat losses are not too large, they proved the existence of two possible steady flame balls, a small one and a large one. It was proved that they have different linear stability properties: the small flame ball, similar to the Zeldovich flame, is unstable under radial perturbations, whereas the large flame ball is stable under radial perturbations but unstable under three-dimensional perturbation if its radius is too large. Similar results have been obtained for a refined version of the previous model where linear far field heat loss is considered [6]. Using matched asymptotic expansions for large activation energy, Buckmaster, Joulin, and Ronney [5, 6] derived an integro-differential model for the nonlinear radial motion of the flame when $Le < 1$:

$$\partial_{1/2} R(\tau) = \log R(\tau) - \lambda R(\tau)^2 + \frac{Eq(\tau)}{R(\tau)}, \quad R(0) = 0,$$

with $\partial_{1/2} R = \frac{d}{dt} \int_0^t \frac{R(s)}{\sqrt{\pi(t-s)}} ds$ and $Eq(\tau)$ representing the amount of energy injected into the system. Numerical simulations of this model suggested that the small flame

*Received by the editors May 11, 2006; accepted for publication (in revised form) November 17, 2006; published electronically April 10, 2007. This work was supported by a CNRS/NWO grant and the RTN network Front-Singularities, HPRN-CT-2002-00274.
http://www.siam.org/journals/siap/67-3/65961.html
†Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands (vincent@few.vu.nl), and Université Bordeaux 1, Mathématiques Appliquées de Bordeaux, 33405 Talence cedex, France (vincent.guyonne@math.u-bordeaux1.fr).
‡Université Claude Bernard, Institut Camille Jordan, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne cedex, France (noble@math.univ-lyon1.fr).

ball is unstable and the large one is stable. Recently, Rouzaud [15] carried out a rigorous study of the long time behavior of flame balls for this equation, which confirms the qualitative behavior. Moreover, Audounet, Roquejoffre, and Rouzaud developed numerical schemes adapted for this kind of equation, which possess similar mathematical properties and exhibit the same type of asymptotic behavior [2].

More refined models have been considered with reaction taking place within an unbounded medium that contains a small volume fraction of porous solid that only exchanges heat with the gas [17]. Heat losses through radiation are modeled in two different ways: either they are constant in the burnt gas and linear in the unburnt gas (similar to the Buckmaster et al. studies), or they are a continuous dimensionless form of Stefan's law having a linear part that dominates close to ambient temperatures and a fourth power that dominates at higher temperatures. Similarly, two branches of solutions are found, the branch of large flame balls being linearly stable and the smaller one being unstable.

Recently, Guyonne, Hulshof, and Van den Berg [18] proposed another mechanism to take into account the radiation, considering a model of flame with *radiative transfer*. Indeed, the presence of particles in the mixture generates a radiation field approximated by the well-known Eddington equation

$$-\nabla\nabla\cdot q + 3\alpha^2 q + \alpha\nabla\theta^4 = 0,$$

where q represents the radiative flux, θ the temperature, and α the opacity of the medium. This system is coupled with a classical free boundary combustion model with simple chemistry $F \to B$, where F is the fresh gas and B the burnt gas. This model is derived in the high activation limit, the reaction occurring in a reaction sheet located at $r = R(t)$, and can be written

$$\partial_t Y - \frac{1}{Le}\Delta Y = 0, \quad r > R(t); \quad Y = 0, \quad r < R(t); \quad \partial_t\theta - \Delta\theta = -\beta\nabla\cdot q, \quad r \neq R(t),$$

with the jump conditions at $r = R(t)$

$$[\theta] = [y] = 0, \qquad \frac{1}{Le}[Y_r] = -[\theta_r] = F_\epsilon(\theta(R(t))),$$

where $F_\epsilon(\theta)$ is the reaction rate modeled by an Arrhenius law. It is proved that there exist steady flame balls for this model. Moreover, numerical simulations with numerical continuation software show that for the same set of parameters there exist several steady flames. Then the question of their stability arises. Even in the simpler case where a linearized Eddington law is considered,

$$-\nabla\nabla\cdot q + 3\alpha^2 q + \alpha\nabla\theta = 0,$$

the question of the linear and nonlinear stability of steady flame balls is far from being understood: we shall mention here the works of Guyonne, Hulshof, and Van den Berg [9] on the numerical analysis of the Evans function for the linear and nonlinear Eddington law, and the paper of Guyonne and Lorenzi [8], which proves that spectral instability implies nonlinear instability with semigroup techniques. Getting nonlinear stability within this framework is quite a hard problem and remains open.

The purpose of this paper is to analyze the stabilizing effect of *radiative transfer* with the viewpoint developed by Buckmaster, Joulin, and Ronney [5, 6]. Indeed, to study the nonlinear growth of radial solutions, they derived an integro-differential

equation using matched asymptotic expansions. This approach has been justified rigorously by Lederman, Roquejoffre, and Wolansky [11] for the adiabatic model with a direct derivation from the reaction diffusion system. Moreover, the asymptotic behavior of this kind of integro-differential equations is now well understood, and efficient numerical schemes are available. The aim of this paper is twofold: through the formal derivation of the same type of integro-differential equations for flame ball growth, and the mathematical and numerical analysis of this model, we want, on the one hand, to study the stabilizing effect of the *radiative transfer* in the formation of flame balls. On the other hand, for this particular model of radiative transfer (the linearized Eddington equation), we want to discuss directly the dynamic of flame balls and the stability of steady flame balls *under radial perturbations* obtained in [18]. We are not concerned here with the stability of flame balls under three-dimensional perturbations. It is worth noting that the approach proposed is very complementary to the indirect approach, which consists of studying the linear stability and then getting information for the full nonlinear free boundary problem. Moreover, if it is possible to rigorously justify this formal derivation, similarly to the paper of Lederman, Roquejoffre, and Wolansky [11], this should give a complete answer on the nonlinear stability of flame balls *under radial perturbations* in the presence of radiative transfer, but this justification is a hard issue.

The paper is organized as follows. In section 2, we derive an integro-differential equation for the nonlinear radial motion of a flame ball using matched asymptotic expansions:

$$\partial_{1/2} R(\tau) = \log R(\tau) - \lambda R(\tau) + \frac{Eq(\tau)}{R(\tau)}\,. \tag{1.1}$$

The dynamic is articulated around the two steady flame balls with radius $R_1 < R_2$ solutions of $\log R = \lambda R$, provided that $\lambda < \frac{1}{e}$. In section 3 we study mathematically the asymptotic behavior of the solutions of (1.1) and discuss the stability of the "large" flame ball with radius R_2 and instability of the "small" flame ball with radius R_1. In section 4, we carry out numerical computations on (1.1) using the numerical schemes designed by Audounet, Roquejoffre, and Rouzaud [2].

2. Growth model for the radius of the flame balls. We consider the following model of combustion with simple chemistry coupled with the *linearized* Eddington equation:

$$\begin{aligned} &\partial_t y - \frac{1}{Le}\Delta y = 0, \quad r > R(t), \qquad y = 0, \quad r < R(t), \\ &\partial_t \theta - \Delta\theta = \beta u, \quad -\Delta u + 3\alpha^2 u = \alpha\Delta\theta, \quad r \neq R(t), \end{aligned} \tag{2.1}$$

where u denotes $u = -\nabla \cdot q$, supplemented with the jump conditions at $r = R(t)$

$$\begin{aligned} &[u] = [\theta] = [y] = 0, \\ &[u_r] = -\alpha[\theta_r], \qquad \frac{1}{Le}[y_r] = -[\theta_r] = F_\epsilon(\theta(R(t))). \end{aligned} \tag{2.2}$$

Moreover, the functions (y, θ, u) must satisfy the conditions at infinity

$$\lim_{r\to\infty} (y(r), \theta(r), u(r)) = (1, 0, 0). \tag{2.3}$$

The reaction rate F_ϵ is given by an Arrhenius law $F_\epsilon(\theta) = A \exp -\frac{1}{\epsilon\theta}$: the constant A is a preexponential factor, and ϵ^{-1} is the activation energy, which is assumed to

be large ($0 < \epsilon \ll 1$). In order to derive an equation for the growth a flame ball, we follow the methodology introduced by Joulin [4] and divide the space into two concentric regions: a quasi-stationary zone, where time derivatives are neglected, in which the combustion occurs, and a far field zone, where the only phenomena that are taken into account are diffusion of the reactant and the temperature. The radiative effects are considered both in the reaction zone and the far field zone. We then obtain an equation for the radius by matching the derivatives of the inner quasi-stationary solution and the outer solution.

2.1. Steady solutions. Before computing quasi-steady solutions, let us first compute the steady solutions of (2.1), (2.2), (2.3). Let us fix the radius $R > 0$ and define η as $\eta = \eta_{\alpha\beta} = \sqrt{3\alpha^2 + \alpha\beta}$. Then there exists a unique steady solution with the following analytic expression:

$$u(r) = \begin{cases} -\dfrac{B_1\eta^2}{\beta r}\sinh(\eta r) & \text{for } r \le R, \\ -\dfrac{B_2\eta^2}{\beta r}\exp(-\eta r) & \text{for } r > R, \end{cases} \tag{2.4}$$

where the constants are given by

$$B_1 = \frac{\alpha\beta Y_f}{\mathrm{Le}\eta^3}\exp(-\eta R), \qquad B_2 = \frac{\alpha\beta Y_f}{\mathrm{Le}\eta^3}\sinh(\eta R), \qquad B_3 = \frac{3\alpha^2 Y_f}{\mathrm{Le}\eta^2}.$$

The expression for θ is

$$\theta(r) = \begin{cases} \dfrac{B_1}{r}\sinh(\eta r) + B_3 & \text{for } r \le R, \\ \dfrac{B_2}{r}\exp(-\eta r) + \dfrac{B_3 R}{r} & \text{for } r > R. \end{cases} \tag{2.5}$$

Finally the solution for the mass fraction variable is expressed by

$$y(r) = \max\left(0, 1 - \frac{R}{r}\right). \tag{2.6}$$

Then the temperature at the front is given by

$$Le\theta(R) = 1 + \frac{\alpha\beta}{\eta^2}\left(\frac{1 - \exp(-2\eta R)}{2\eta R} - 1\right).$$

Note that the dependence of the temperature at the front on the flame radius in the case where radiative transfer is taken into account is different from the case where heat loss radiative terms are considered. In the latter case, the dependence is *parabolic* [5].

It is easily seen via (2.6) that $[y_r]_{|r=R} = \frac{1}{R}$. Then the steady flame balls are the steady solutions (y, θ, u) defined by (2.4), (2.5), (2.6) such that R is the solution of

$$F_\epsilon(\theta(R)) = \frac{1}{RLe}. \tag{2.7}$$

The case where $\alpha, \beta \to 0$ as $\epsilon \to 0$ is of particular interest: as a matter of fact, Buckmaster, Joulin, and Ronney [5, 6] also considered vanishing heat loss terms as $\epsilon \to 0$. As a consequence, in some asymptotic parameter regimes, it is possible to

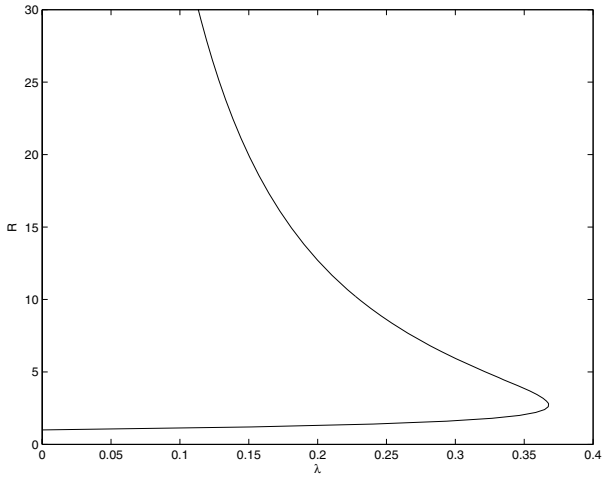


FIG. 2.1. *Diagram of the bifurcation of* (2.8) *in the* $(\lambda = \frac{\overline{\beta} Le}{\sqrt{3}}, R)$ *plane.*

simplify (2.7). Under the scaling $\beta = \overline{\beta}\epsilon$, $\alpha \gg \beta$ (we can choose $\alpha = O(\epsilon^\mu)$ with $0 < \mu < 1$), and $0 < \epsilon \ll 1$, we find that

$$\theta(R) = \frac{1}{Le} - \frac{\overline{\beta}\epsilon}{Le\sqrt{3}} R + O(\epsilon^2).$$

Inserting this relation into (2.7) and letting $\epsilon \to 0$, one finds

$$\log \frac{R}{R_{ad}} = \frac{\overline{\beta} Le}{\sqrt{3}} R, \tag{2.8}$$

where R_{ad} denotes the adiabatic radius. The set of solutions of (2.8) is plotted in Figure 2.1 in the (R, λ) plane, where $\lambda = \frac{\overline{\beta} Le}{\sqrt{3}}$.

We can see that for $\lambda > \lambda_{cr} = \frac{1}{R_{ad} e}$ no solution exists, and for $\lambda < \lambda_{cr}$ there exist two solutions $R_1 < R_2$, which correspond to steady flame balls. The smaller flame ball converges to the flame ball constructed by Zeldovich in the limit $\lambda \to 0$.

It is important to note the difference between the equation for steady flame balls (2.8) when *radiative transfer* is taken into account and the equation for steady flame balls when *radiative heat losses* are considered (which are constant in the burnt phase of order $O(\epsilon)$, linear in the fresh phase and of order $O(\epsilon^2)$). In that case, the equation for steady flame balls reads

$$\log\left(\frac{R}{R_{ad}}\right) = \Lambda R^2,$$

where Λ is a constant depending on heat loss terms. The difference between the two equations is that, considering the *constant heat loss term* in the burnt gas, the temperature at the front $\theta(R)$ is a *parabolic* function of the flame radius, whereas in the case of radiative transfer, the temperature at the front is a *linear* function of the flame radius (in the parameter regime considered previously).

2.2. Inner solutions. Now we consider the nonstationary case, and we suppose that the flame has a spherical symmetry with a flame radius at $r = R(t)$. The purpose is to compute an equation satisfied by $R(t)$. Let us denote (Y, Θ, U) the steady solution computed at the previous section and fix $\nu \in (0,1)$. We are first going to compute an approximation of the solution (y, θ, u) on $B(0, \epsilon^{-\nu})$, considering that it is a quasi-steady solution with a flame radius $r = R(t) \ll \epsilon^{-\nu}$. We write the solution (y, θ, u) as

$$(y, \theta, u) = (1 + \epsilon v(t))\big(Y, \Theta, U\big) + (0, \epsilon w(t), 0).$$

This solution satisfies the steady equations and the jump conditions, provided that

$$\frac{1 + \epsilon v(t)}{RLe} = F_\epsilon\big(\Theta(R) + \epsilon(\Theta(R)v + w)\big).$$

This equation reads, up to order $O(\epsilon)$,

$$v + \frac{w}{\Theta(R)} = -\log(RLe) - \log\Big(F_\epsilon\big(\Theta(R)\big)\Big).$$

The boundary conditions at $r = \epsilon^{-\nu}$ are given by

$$\begin{aligned}
\theta(t, \epsilon^{-\nu}) &= \frac{3\alpha^2}{Le(3\alpha^2 + \alpha\beta)} R\epsilon^\nu + \epsilon w(t) + O(\epsilon^{\nu+1}),\\
y(t, \epsilon^{-\nu}) &= 1 - \epsilon^\nu R(t) + \epsilon v(t) + O(\epsilon^{\nu+1}),\\
u(t, \epsilon^{-\nu}) &= -\frac{\alpha(1 + \epsilon v(t))\epsilon^\nu}{Le\sqrt{3\alpha^2 + \alpha\beta}} \sinh(R\sqrt{3\alpha^2 + \alpha\beta}) \exp\left(-\frac{\sqrt{3\alpha^2 + \alpha\beta}}{\epsilon^\nu}\right)\\
&= O(\epsilon^{\nu+3}).
\end{aligned} \tag{2.9}$$

The last estimate on u is valid, provided that $\sqrt{3\alpha^2 + \alpha\beta} = O(\epsilon^\mu)$ with $\mu < \nu$.

2.3. Outer solution. We compute an approximate solution (y, θ, u) outside $B(0, \epsilon^{-\nu})$ of

$$\partial_t y - \frac{1}{Le}\Delta y = 0, \quad \partial_t \theta - \Delta\theta = \beta u, \quad -\Delta u + 3\alpha^2 u = \alpha\Delta\theta,$$

supplemented with the boundary conditions (2.9); the conditions at infinity are given by $\lim_{r\to\infty}(y, \theta, u) = (1, 0, 0)$, and the initial conditions will be specified later. Following the derivation of Joulin [4], we rescale time and space—$\tau = \epsilon^2 t, \rho = \epsilon r$—and define

$$(\overline{y}, \overline{\theta}, \overline{u})(\tau, \rho) = \left(\frac{y(t,r) - 1}{\epsilon}, \frac{\theta(t,r)}{\epsilon}, \frac{u(t,r)}{\epsilon^3}\right), \qquad \overline{R}(\tau) = R(t).$$

Then $(\overline{y}, \overline{\theta}, \overline{u})$ satisfies the rescaled system

$$\partial_\tau \overline{y} - \frac{1}{Le}\Delta\overline{y} = 0, \quad \partial_\tau\overline{\theta} - \Delta\overline{\theta} = \beta\overline{u}, \quad -\epsilon^2\Delta\overline{u} + 3\alpha^2\overline{u} = \alpha\Delta\overline{\theta},$$

with the boundary conditions

$$\begin{aligned}
\overline{y}(\tau, \epsilon^{1-\nu}) &= -\epsilon^{\nu-1}\overline{R}(\tau) + \overline{v}(\tau) + O(\epsilon^\nu),\\
\overline{\theta}(\tau, \epsilon^{1-\nu}) &= \epsilon^{\nu-1}\frac{3\alpha^2}{Le(3\alpha^2 + \alpha\beta)}\overline{R}(\tau) + \overline{w}(\tau) + O(\epsilon^\nu),\\
\overline{u}(\tau, \epsilon^{1-\nu}) &= O(\epsilon^\nu).
\end{aligned}$$

We are interested only in the jumps of the radial derivatives, namely, $(\rho\overline{y})_\rho$, $(\rho\overline{\theta})_\rho$ evaluated at the point $\rho = \epsilon^{1-\nu}$. Let us start with $\frac{\partial}{\partial\rho}(\rho\overline{y})_{|\rho=\epsilon^{1-\nu}_+}$ and define $Y = \rho\overline{y}$ for all $\rho > \epsilon^{1-\nu}$. Then extend Y on $\mathbb{R}$ by an odd function $\overline{Y}$ so that

$$\overline{Y}(\tau,\rho) = \begin{cases} Y\left(\tau, \frac{1}{\sqrt{Le}}\rho + \epsilon^{1-\nu}\right) - Y(\tau,\epsilon^{1-\nu}) & \text{for } \rho > 0, \\ -Y\left(\tau, -\frac{1}{\sqrt{Le}}\rho + \epsilon^{1-\nu}\right) + Y(\tau,\epsilon^{1-\nu}) & \text{for } \rho < 0. \end{cases}$$

Define $\psi(\tau) = Y(\tau,\epsilon^{1-\nu}) = -\overline{R}(\tau) + O(\epsilon^{1-\nu})$: $\overline{Y}$ satisfies the equation

$$\overline{Y}_\tau - \overline{Y}_{\rho\rho} = \dot{\psi}(\tau)\big(1_{]-\infty,0[}(\rho) - 1_{]0,\infty[}(\rho)\big),$$

with initial condition

$$\overline{Y}(0,\rho) = \begin{cases} Y\left(0, \frac{1}{\sqrt{Le}}\rho + \epsilon^{1-\nu}\right) - \psi(0) & \text{for } \rho > 0, \\ -Y\left(0, -\frac{1}{\sqrt{Le}}\rho + \epsilon^{1-\nu}\right) + \psi(0) & \text{for } \rho < 0. \end{cases}$$

Then we find that

$$\overline{Y}(\tau,\rho) = \frac{1}{\sqrt{4\pi\tau}}\int_0^\infty \left(Y\left(0, \frac{x}{\sqrt{Le}} + \epsilon^{1-\nu}\right) - \psi(0)\right)\left(e^{-\frac{|x-\rho|^2}{4\tau}} - e^{-\frac{|x+\rho|^2}{4\tau}}\right)dx$$
$$- \int_0^\tau\int_0^\infty \dot{\psi}(s)\frac{e^{-\frac{|x-\rho|^2}{4(\tau-s)}} - e^{-\frac{|x+\rho|^2}{4(\tau-s)}}}{\sqrt{4\pi(\tau-s)}}dxds.$$

Derive $\overline{Y}$ with respect to ρ and take the value at point $\rho = 0$:

$$\frac{\partial}{\partial\rho}(\rho\overline{y})_{|\rho=\epsilon^{1-\nu}} = \sqrt{Le}\frac{\partial\overline{Y}}{\partial\rho}_{|\rho=0} = -\sqrt{Le}\partial_{1/2}\psi(\tau) + \phi_y(\tau),$$

where ϕ_y is a function of only the initial data $y(0,x)$, given by

$$\phi_y(\tau) = \frac{Le^{\frac{3}{2}}}{\sqrt{4\pi\tau}}\int_{\epsilon^{1-\nu}}^\infty \big(x\overline{y}(0,x) - \epsilon^{1-\nu}\overline{y}(0,\epsilon^{1-\nu})\big)\frac{x-\epsilon^{1-\nu}}{\tau}e^{-Le\frac{(x-\epsilon^{1-\nu})^2}{4\tau}},$$

and the fractional derivative $\partial_{1/2}\psi(\tau)$ is the function

$$\partial_{1/2}\psi(\tau) = \frac{d}{dt}\int_0^t \frac{\psi(s)}{\sqrt{\pi(t-s)}}ds.$$

As a conclusion, we find that

$$\frac{\partial}{\partial\rho}(\rho\overline{y})^+_{|\rho=\epsilon^{1-\nu}} = \sqrt{Le}\partial_{1/2}\overline{R} + \phi_y(\tau) + O(\epsilon^{1-\nu}).$$

We compute the derivative of $\rho\overline{\theta}$ at point $\rho = \epsilon^{1-\nu}$: following the analysis made previously, one introduces the functions $T = \rho\overline{\theta}$, $V = \rho\overline{u}$ and defines

$$\overline{T}(\tau,\rho) = \begin{cases} T(\tau,\rho+\epsilon^{1-\nu}) - T(\tau,\epsilon^{1-\nu}) & \text{for } \rho > 0, \\ -T(\tau,-\rho+\epsilon^{1-\nu}) + T(\tau,\epsilon^{1-\nu}) & \text{for } \rho < 0, \end{cases}$$

and

$$\overline{V}(\tau,\rho)=\begin{cases} V(\tau,\rho+\epsilon^{1-\nu})-V(\tau,\epsilon^{1-\nu}) & \text{for } \rho>0,\\ -V(\tau,-\rho+\epsilon^{1-\nu})+V(\tau,\epsilon^{1-\nu}) & \text{for } \rho<0.\end{cases}$$

The functions $\overline{T},\overline{V}$ satisfy the system

$$\begin{aligned}\partial_\tau\overline{T}-\overline{T}_{\rho\rho}&=\beta\overline{V}+\dot{\overline{T}}_\epsilon H(\rho)-\beta\overline{V}_\epsilon(\tau)H(\rho),\\ -\epsilon^2\overline{V}_{\rho\rho}+3\alpha^2\overline{V}&=\alpha\overline{T}_{\rho\rho}+3\alpha^2\overline{V}_\epsilon(\tau)H(\rho),\end{aligned}\tag{2.10}$$

with $H(\rho)=1_{]-\infty,0[}(\rho)-1_{]0,\infty[}$ and $(\overline{T}_\epsilon,\overline{V}_\epsilon)=\epsilon^{1-\nu}(\overline{\theta},\overline{u})(\tau,\epsilon^{1-\nu})$. Take the Fourier transform of the system (2.10): this yields

$$\begin{aligned}\partial_\tau\hat{\overline{T}}+\xi^2\hat{\overline{T}}&=\beta\hat{\overline{V}}+\dot{\overline{T}}_\epsilon(\tau)\hat{H}(\xi)-\beta\overline{V}_\epsilon(\tau)\hat{H}(\xi),\\ (3\alpha^2+\epsilon^2\xi^2)\hat{\overline{V}}&=-\alpha\xi^2\hat{\overline{T}}+3\alpha^2\overline{V}_\epsilon(\tau)\hat{H}(\xi).\end{aligned}\tag{2.11}$$

Eliminating $\hat{\overline{V}}$ from (2.11) yields the equation on $\hat{\overline{T}}$:

$$\partial_\tau\hat{\overline{T}}+\xi^2\left(1+\frac{\alpha\beta}{3\alpha^2+\epsilon^2\xi^2}\right)\hat{\overline{T}}=\dot{\overline{T}}_\epsilon\hat{H}(\xi)+\beta\overline{V}_\epsilon(\tau)\left(\frac{3\alpha^2}{3\alpha^2+\epsilon^2\xi^2}-1\right)\hat{H}(\xi).$$

The function $\hat{\overline{T}}$ is given by

$$\begin{aligned}\hat{\overline{T}}(\tau,\xi)=&\,e^{-(1+\frac{\alpha\beta}{3\alpha^2+\epsilon^2\xi^2}\xi^2\tau)}\hat{\overline{T}}(0,\xi)+\int_0^\tau e^{-(1+\frac{\alpha\beta}{3\alpha^2+\epsilon^2\xi^2})\xi^2(\tau-s)}\hat{H}(\xi)\dot{\overline{T}}_\epsilon(s)ds\\ &-\int_0^\tau\overline{V}_\epsilon(s)\frac{\beta\epsilon^2\xi^2}{3\alpha^2+\epsilon^2\xi^2}\hat{H}(\xi)e^{-(1+\frac{\alpha\beta}{3\alpha^2+\epsilon^2\xi^2})\xi^2(\tau-s)}ds.\end{aligned}$$

The analysis is now completely similar to the case treated previously for the derivative of $\rho\overline{y}$ at the boundary: take the inverse Fourier transform of $\hat{\overline{T}}$ and derive $\overline{T}$ with respect to ρ. There exists ϕ_θ which is a function only of $\theta(0,.)$ such that

$$\frac{\partial}{\partial\rho}(\rho\overline{\theta})_{|\rho=\epsilon^{1-\nu}}=-\frac{1}{Le(1+\frac{\alpha\beta}{3\alpha^2})^{\frac{3}{2}}}\partial_{1/2}\overline{R}+\phi_\theta(\tau)+O(\epsilon^{1-\nu}).$$

2.4. Matching of the derivatives. Recall that the analysis of the outer solution yields

$$\begin{aligned}\frac{\partial}{\partial\rho}(\rho\overline{\theta})^+_{|\rho=\epsilon^{1-\nu}}&=-\frac{1}{Le(1+\frac{\alpha\beta}{3\alpha^2})^{\frac{3}{2}}}\partial_{1/2}\overline{R}+\phi_\theta(\tau)+O(\epsilon^{1-\nu}),\\ \frac{\partial}{\partial\rho}(\rho\overline{y})^+_{|\rho=\epsilon^{1-\nu}}&=\sqrt{Le}\partial_{1/2}\overline{R}+\phi_y(\tau)+O(\epsilon^{1-\nu}).\end{aligned}\tag{2.12}$$

Moreover, it is easily proved using the expression of the inner solution that

$$\frac{\partial}{\partial\rho}(\rho\overline{\theta})^-_{|\rho=\epsilon^{1-\nu}}=w(\tau),\qquad\frac{\partial}{\partial\rho}(\rho\overline{y})^-_{|\rho=\epsilon^{1-\nu}}=v(\tau).\tag{2.13}$$

The jump conditions at the free boundary are given by

$$v(\tau)+\frac{w}{\Theta(R)}=-\log(RLe)-\log\Big(F_\epsilon\big(\Theta(R)\big)\Big).\tag{2.14}$$

Eliminating v, w from (2.12), (2.13), (2.14) and up to order $O(\epsilon^{1-\nu})$, we find the equation for the radius $\overline{R}$ of the flame ball:

$$\left(\frac{(Le\Theta(\overline{R}))^{-1}}{(1+\frac{\alpha\beta}{3\alpha^2})^{\frac{3}{2}}}-\sqrt{Le}\right)\partial_{1/2}\overline{R}=\log(RLe)+\log(F_\epsilon(\Theta(\overline{R})))+\frac{\phi_\theta(\tau)}{\Theta(\overline{R})}+\phi_y(\tau).$$

This expansion is valid, provided that we have chosen $\sqrt{3\alpha^2+\alpha\beta}=O(\epsilon^\mu)$ with $\mu<\nu$. This condition is satisfied when β,α are $O(\epsilon^\mu)$.

Let us choose the scaling $\alpha=\overline{\alpha}\epsilon^\mu$ and $\beta=\overline{\beta}\epsilon$, which clearly satisfies the hypothesis $\beta,\alpha=O(\epsilon^\mu)$. In this case, we can simplify the equation of growth. The front temperature is given by

$$\Theta(R)=\frac{1}{Le}-\frac{\overline{\beta}\epsilon}{Le\sqrt{3}}R+(\text{h.o.t.}),$$

where h.o.t. stands for higher order terms. Then the equation for the radius growth can be written

$$(1-\sqrt{Le})\partial_{1/2}\overline{R}=\log\left(\frac{\overline{R}}{R_{ad}}\right)-\frac{Le\overline{\beta}}{\sqrt{3}}R+\Phi(\tau),$$

where R_{ad} is the adiabatic radius. The function Φ is a function of only the initial data $y(0,.)$ and $\theta(0,.)$.

It is remarkable to see the influence of the radiative transfer through the term $-\lambda R$ instead of the heat loss term of radiation $-\lambda R^2$ derived in other analysis [5, 6].

Now if we consider a reaction initiated by the input of energy of order $O(\epsilon)$ at the origin, we choose initial conditions so that $\Phi(\tau)=0$, and the only difference comes from the near field equations: we have to solve the quasi-stationary equations

$$\begin{aligned}-\Delta\,\theta&=\beta\,u+\epsilon\,\mathcal{Q}(t)\delta(r=0),\\ -\Delta\,u+3\alpha^2\,u&=\alpha\Delta\,\theta,\end{aligned}\tag{2.15}$$

where $\mathcal{Q}$ represents the amount of energy input into the system at the origin (see [10] for more details). This system is completed with jump conditions detailed in section 2.2. The stationary solution of (2.15) with jump conditions is the sum of the stationary solution computed in section 2.2 and a particular solution (u_p,θ_p) which is smooth at point $r=R(t)$. It is a Fourier transform exercise to prove that a particular solution (u_p,θ_p) of this system is given by

$$u_p=\alpha\epsilon\mathcal{Q}(t)\frac{\mathrm{sh}(\sqrt{3\alpha^2+\alpha\beta})r}{4\pi\,r},\qquad \theta_p=\frac{\epsilon\mathcal{Q}(t)}{4\pi\,r}-\frac{\alpha\beta\epsilon\mathcal{Q}(t)}{4\pi(3\alpha^2+\alpha\beta)}\frac{\mathrm{sh}\sqrt{3\alpha^2+\alpha\beta}r}{r}.$$

Thus in the asymptotic $\alpha=O(\epsilon^\mu)$ with $0<\mu<1$ and $\beta=\overline{\beta}\epsilon$, the temperature at the front is given by

$$\Theta(R)=\frac{1}{Le}-\frac{\overline{\beta}\epsilon}{Le\sqrt{3}}R+\frac{\epsilon\mathcal{Q}(t)}{4\pi\,R}+(\text{h.o.t}).\tag{2.16}$$

As a consequence, we find the growth equation

$$(1-\sqrt{Le})\partial_{1/2}\overline{R}=\log\frac{\overline{R}}{R_{ad}}-\frac{Le\overline{\beta}}{\sqrt{3}}\overline{R}+\frac{Le^2}{4\pi}\frac{\mathcal{Q}(\tau)}{\overline{R}}.$$

In what follows, we put the last term concerning the energy input at the origin in the form $\frac{Eq(\tau)}{\overline{R}}$. Here E represents the intensity of the energy input and $q(\tau)$ corresponds to the time fluctuations of this energy input. In the next sections, we are going to analyze mathematically and numerically

$$(1-\sqrt{Le})\partial_{1/2}\overline{R} = \log\frac{\overline{R}}{R_{ad}} - \frac{Le\overline{\beta}}{\sqrt{3}}\overline{R} + \frac{Eq(\tau)}{\overline{R}}.$$

3. Mathematical results. In this section, we consider the more generalized equation,

$$\mu R\partial_{1/2}R = R\log R + Eq - \lambda R, \quad t\in\mathbb{R}^+,\ R(0)=0, \tag{3.1}$$

where $\mu>0$, $\lambda>0$ and

$$\partial_{1/2}R = \frac{1}{\sqrt{\pi}}\int_0^t \frac{\dot{R}(s)}{\sqrt{t-s}}\,ds = \frac{1}{\sqrt{\pi}}\frac{d}{dt}\int_0^t \frac{R(s)}{\sqrt{t-s}}\,ds. \tag{3.2}$$

This describes the evolution of a spherical flame, initiated by a point source energy input $Eq(t)$, at which are applied heat losses of radiative nature, represented by the parameter λ. The intensity of this energy input is measured by the positive constant E, and its time evolution is described by the function q. This one is a smooth, nonnegative function, with connected support and unit total mass; its initial values satisfy the assumption

$$q(t)\sim q_0 t^{\beta} \quad \text{as } t\to 0 \quad \text{with } 0\le\beta<\frac{1}{2}, \tag{3.3}$$

and as $t\to+\infty$, q tends to 0. Finally, the parameter μ can be viewed as a time rescaling (see section 4) and is assumed, in this section, to be a positive real number, fixed to 1.

Mathematical results of (3.1) are, according to minor modifications in the proofs, similar to the ones written in [1, 15]. Therefore proofs are omitted; only comments are mentioned when necessary.

Let us begin to state existence results for the Cauchy problem.

PROPOSITION 3.1. *Let us assume q positive on $[0,t_0]$, $q(0)=q_0$. Then there exists $t_1\in\,]0,t_0]$ such that* (3.1) *admits a solution in $C^{3/2}([0,t_1])$ satisfying*

$$R(t)\sim R_0 t^{1/4}, \qquad \textit{with } R_0^2 = \frac{Eq_0}{\sqrt{\pi}}\int_0^1 t^{-\frac{1}{4}}(1-t)^{-\frac{1}{2}}dt.$$

In order to prove the existence of a unique maximal solution, the flame radius is expressed as the trace at $x=0$ of a function $u(t,x)$, solution of the following parabolic equation:

$$\begin{cases} u_t - u_{xx} = 2\delta_{x=0}\left(\log u + \dfrac{Eq}{u} - \lambda u\right) & \text{for } x\in\mathbb{R}.\\ u(0,\cdot)=0. \end{cases} \tag{3.4}$$

This formulation is essential to characterizing the long time behavior of the flame. Then we consider more general Cauchy problems, such as

$$\begin{cases} u_t - u_{xx} = 2\delta_{x=0}\left(\log u + \dfrac{Eq}{u} - \lambda u\right) & \text{for } x\in\mathbb{R},\\ u(0,\cdot)=u_0(x), \end{cases} \tag{3.5}$$

where u_0 is an even, Lipschitz, square-integrable, nonnegative function. This is equivalent to solving

$$\begin{cases} u_t - u_{xx} = 0, \quad x > 0, \\ u_x(t,0) = -\left(\log u + \dfrac{Eq}{u} - \lambda u\right) \quad \text{for } x \in \mathbb{R}, \\ u(0,\cdot) = u_0(x). \end{cases}$$

Such a formulation allows us to prove the following result.

THEOREM 3.2. *Let q satisfy condition* (3.3). *We suppose there exists $t_0 > 0$ such that $q(t) > 0$ on $]0, t_0[$ and $q(t) = 0$ if $t \geq t_0$; then the following hold*:

(i) *If $t_0 = +\infty$,* (3.5) *has a unique global positive solution, except at $t = 0$. Moreover, u is C^∞ on $\mathbb{R}_+^* \times \mathbb{R}^*$ and $t \mapsto u(t,0)$ is C^∞ on $\mathbb{R}_+^*$.*

(ii) *If $t_0 < +\infty$,* (3.5) *has a unique maximal solution u defined on an interval $[0, t_{max}[$, positive, except at $t = 0$. Moreover, u is C^∞ on $]0, t_{max}[$. If $t_{max} < +\infty$, there exists $t_n \to t_{max}$ such that $\lim_{n\to+\infty} u(t_n, 0) = 0$.*

In particular, a consequence of this theorem is the existence of a solution of (3.1). The uniqueness of u is based on a comparison principle (see [15] for more details).

These results now recalled, we may discuss different cases where either quenching or stabilization of the flame occurs. For this purpose, we denote by u_E the solution of (3.4) and by $R_E(t) := u_E(t,0)$ the corresponding radius of the flame. Let us turn to the asymptotic behavior of the radius. In order to prove the following results, monotonicity methods (cf. [16]) are of major importance. Indeed, sub- or supersolutions are computed and create, therefore, an admissible range for the solutions. At this stage, a comparison principle coupled to a relevant choice of the bounds reveals to us either quenching or stabilization of the flame.

Before going further, we make a remark on the role played by the parameter λ. The stationary solutions of

$$u_t - u_{xx} = 2\delta_{x=0}(\log u - \lambda u)$$

are the constants R satisfying

$$\log R = \lambda R;$$

hence the values of λ_{cr} and the distinction we have to make between the cases $\lambda < \lambda_{cr}$ and $\lambda > \lambda_{cr}$. Please note that we do not consider the case $\lambda = \lambda_{cr}$, the study being identical to [15]. Moreover, λ is assumed nonnegative in what follows, so that u_E, the solution of (3.5), is a bounded function.

We first consider the supercritical case, $\lambda > \lambda_{cr}$, corresponding to high radiative heat losses. Then the loss of energy is too important and the flame quenches. We state the following proposition.

PROPOSITION 3.3. *Assume $\lambda > \lambda_{cr}$.*

(i) *If $q > 0$ on $\mathbb{R}_+^*$, then the solution of* (3.4) *is global and $\lim_{t\to+\infty} R_E(t) = 0$.*

(ii) *If q is compactly supported, R_E quenches in finite time.*

We now consider the subcritical case, $\lambda < \lambda_{cr}$. This situation leads to different properties, more complex because they depend on the quantities E and q, i.e., the amount of energy we input into the system and the time length of this injection. We have the following claim.

THEOREM 3.4. *Assume $\lambda < \lambda_{cr}$ and $q > 0$ on $\mathbb{R}_+^*$. Then* (3.1) *has a unique global solution $R_E(t)$ and there exists $E_{cr}(q) > 0$ such that*

(i) *if* $E < E_{cr}(q)$, $\lim_{t\to+\infty} R_E(t) = 0$,
(ii) *if* $E > E_{cr}(q)$, $\lim_{t\to+\infty} R_E(t) = R_2$,
(iii) *if* $E = E_{cr}(q)$, $\lim_{t\to+\infty} R_E(t) = R_1$.

If q *is compactly supported, with support* $]0, t_0[$, *then* (3.1) *has a unique solution* $R_E(t)$ *and there exists* $E_{cr}(q) > 0$ *such that*

(a) *if* $E < E_{cr}(q)$, R_E *quenches in finite time,*
(b) *if* $E \geq E_{cr}(q)$, *the previous result holds again.*

This theorem can be proved with minor changes in the sub- and supersolutions in the proofs developed in [15]. For more details on the theoretical study of this type of equations, we refer the reader to [1, 15].

As a conclusion, we have verified that the equation derived in section 2,

$$(1 - \sqrt{Le})\partial_{1/2}\overline{R} = \log \frac{\overline{R}}{R_{ad}} - \frac{Le\overline{\beta}}{\sqrt{3}}\overline{R} + \frac{Eq(\tau)}{\overline{R}}, \tag{3.6}$$

is well posed, provided that $Le < 1$, and the flame ball quenches if $Le\overline{\beta}$ exceed a threshold. When $Le\overline{\beta}$ is small enough, there exist two steady flame balls. The small one is unstable, and the large one is stable under radial perturbations. Since we have computed nonlinear evolutions of radial perturbations, these results shall be understood as nonlinear stability properties of the steady flame balls.

There are different explanations to justify the assumption Le < 1. From a mathematical point of view, (3.6) is ill posed for Lewis numbers greater than one. Indeed, in this case, we face a backward parabolic equation in which instabilities can occur (see, for example, [4, 3]). The special case Le $= 1$ with high activation energy has been studied in [7]. From a physical point of view, as the Lewis number is the ratio between thermal and molecular diffusion, the condition $Le < 1$ is equivalent, saying that gas molecules diffuse faster than heat. In this configuration, flame balls are able to exist, whereas for Lewis numbers greater than one, flames vanish. Considering the experiments performed by Ronney and coworkers (see, for example, [14]), flame balls are observed only for lean reactant mixtures, for which the Lewis number is between 0.06 and 0.5. Thus the restriction Le < 1 is reasonable and, in fact, is a necessary condition for stationary flame balls to exist.

4. Numerics. In this section, we follow the methods developed in [2]. We first present the numerical scheme and then turn to numerical investigations.

4.1. Presentation of the scheme. We recall that the radius R can be seen as the trace on the axis $x = 0$ of the solution $u(x, t)$ of the diffusive problem (3.5). A suitable scheme for studying long time behavior of such equations is an implicit Euler scheme in time. The scheme reads

$$\begin{cases} \dfrac{u^{n+1} - u^n}{\tau} - u_{xx}^{n+1} = 0 \quad \text{for } x > 0, \\ u_x^{n+1}(0) = -\log u^{n+1}(0) - \dfrac{Eq^{n+1}}{u^{n+1}(0)} + \lambda u^{n+1}(0), \\ u^0 = 0, \end{cases} \tag{4.1}$$

where τ denotes the time step and $q^n = q(n\tau)$. The discretized heat equations can be solved explicitly using Fourier transform, so that system (4.1) determines explicitly the quantity in which we are really interested, i.e., the sequence $R^n := u^n(0)$. Moreover, by induction and because of the maximum principle, we have $u^n \geq 0$.

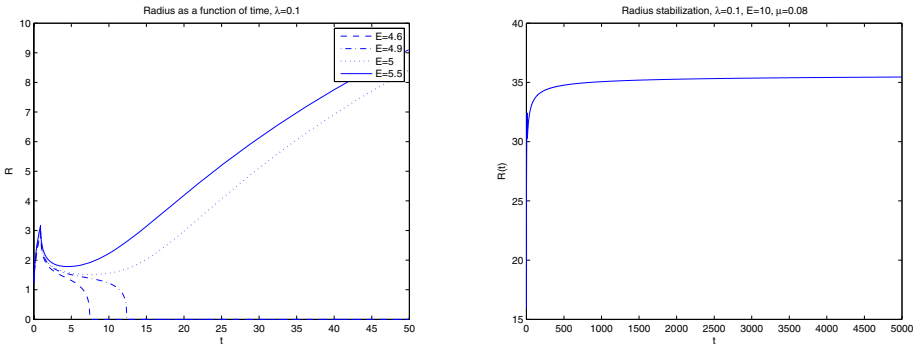


FIG. 4.1. *Evolution of the radius when $\lambda < \lambda_{cr}$ and E is variable; from extinction to stabilization.*

Before going further, we define two quantities needed later on by

$$\begin{cases} \alpha^n = \displaystyle\int_{\mathbb{R}} \frac{\hat{u}^{n-1}(\xi)d\xi}{1+4\pi^2\xi^2\tau} = \sqrt{\tau}\sum_{k=1}^{n-1}\theta_{n-k+1}g^k, \quad \hat{u}^0 = 0, \\ g^n = f^n(\alpha^n + \sqrt{\tau}g^n), \quad n \geq 1, \end{cases}$$

where

$$\theta_{p+1} = \int_{\mathbb{R}} \frac{2\sqrt{\tau}}{(1+4\pi^2\xi^2\tau)^{p+1}}d\xi = \frac{2p-1}{2p}\theta_p = \frac{C^p_{2p-1}}{2^{2p-1}}\theta_1, \qquad \text{with } \theta_1 = 1.$$

The radius R is then expressed in term of these quantities, namely $R^n = \alpha^n + \sqrt{\tau}g^n > 0$. The unknown g^n is determined by successive resolutions of the following implicit equation:

$$\Phi(g^n) := g^n - \log(\alpha^n + \sqrt{\tau}g^n) - \frac{Eq(n\tau)}{\alpha^n + \sqrt{\tau}g^n} + \lambda(\alpha^n + \sqrt{\tau}g^n) = 0. \tag{4.2}$$

In order to be consistent with (3.1), we need to introduce the parameter μ different from 1. It enters (4.2) as

$$\Phi(g^n) := \mu g^n - \log(\alpha^n + \sqrt{\tau}g^n) - \frac{Eq(n\tau)}{\alpha^n + \sqrt{\tau}g^n} + \lambda(\alpha^n + \sqrt{\tau}g^n) = 0. \tag{4.3}$$

This implicit equation is solved by a Newton method with initial data g^{n-1}.

The properties of this scheme remain unchanged so that by [2], the convergence and comparison properties still hold. The numerical scheme also sustained qualitative properties similar to those of the continuous model of flame ball growth (see [2] for more details).

4.2. Numerical results. We now turn to the numerical investigation of the problem. For this purpose, we consider an input energy $q = \chi_{[0,1]}$. In Figure 4.1, we fix a value for the parameter λ, namely $\lambda = 0.1 < \lambda_{cr} = 1/e$, and plot the different radius evolution possibilities for different energy inputs. We note that (see Figure 4.1 (left)) we recover the expected behavior of the radius: when E is small, the flame quenches, whereas when it is larger, the behavior cannot be guessed with this time scale, and numerical simulations have to be performed for longer times. For this

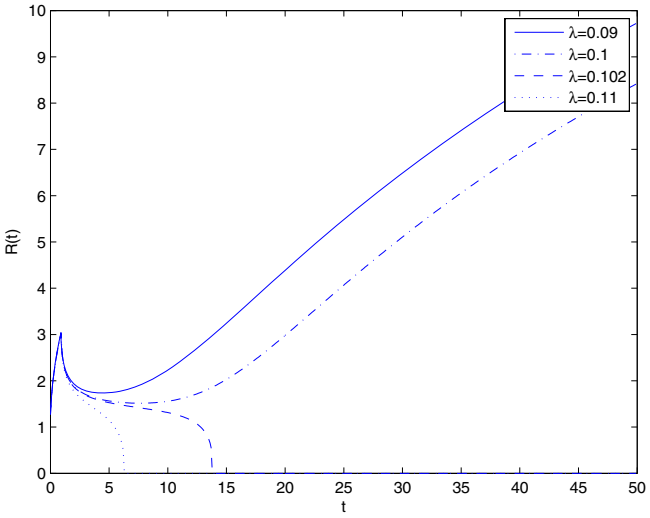


FIG. 4.2. *Evolution of the radius when E fixed and λ variable.*

purpose, we perform a time rescaling. Writing $t = \tau/\epsilon$ in formula (3.1) and dividing by R implies a new expression $\tilde{\mu}\partial_{1/2}R$, where $\tilde{\mu} = \sqrt{\epsilon}\mu$. Dropping the tilde, we are able to simulate a large time scale t via a smaller time scale τ only by taking values of μ less than one. Figure 4.1(right) shows the stabilization towards the radius R_2 in the time coordinate τ.

Finally, in Figure 4.2, the energy E is fixed, and we consider different values for λ. For important radiative heat losses, the flame quenches, whereas it can stabilize to the corresponding critical radius, depending on the value of λ.

5. Conclusion. In this paper, we have studied the stabilizing effect of *radiative transfer* on the formation of flame balls. In the formation of flame balls, radiation is an important physical effect. Instead of considering radiative heat loss, just as Buckmaster, Joulin, and Ronney did in [5, 6] through simplified versions of Stefan's law, we have considered radiative transfer using the well-known (linearized) Eddington law. In some asymptotic parameter regime, we obtain the existence of two steady flame balls. In our study, the asymptotic equation for the radius of the steady flame ball is $\log R = \lambda R$, and that is different from the one obtained in [5, 6], $\log R = \Lambda R^2$. The difference in the power of R comes from the fact that the dependence of the front temperature on R is different depending on whether we consider heat loss terms (in that case, the dependence is parabolic) or radiative transfer (in that case, it is linear). This has an influence on the size of the steady flame balls.

For the linearized Eddington equation of radiative transfer, we have derived, using an approach initiated by Buckmaster, Joulin, and Ronney, an integro-differential equation for the growth of a flame ball. This equation differs again from the one obtained by Buckmaster, Joulin, and Ronney with a loss term proportional to R instead of R^2. The equation obtained in this paper describes the *nonlinear evolution* of *radial perturbation* of steady flame balls and falls into a class of integro-differential equations that are mathematically and numerically well understood [1, 2, 15]; we have used this framework to study mathematically and numerically the asymptotic

behavior of this equation. When two steady flame balls exist, the smaller one is unstable (except for particular values of the parameters), and the larger one is stable, similar to the results obtained in [5, 6]. This gives a partial answer to the stability properties of flame balls obtained in [18] since the perturbations considered in this paper have the *radial* symmetry. Finally, we shall mention that the derivation carried out in this paper is only formal, and it would be interesting to make this derivation rigorously: for that purpose, instead of starting from the free boundary problem, we shall consider the reaction diffusion system with a singular reaction term and follow the method developed by Lederman, Roquejoffre, and Wolansky in the adiabatic case [11].

REFERENCES

[1] J. Audounet, V. Giovangigli, and J.-M. Roquejoffre, *A threshold phenomenon in the propagation of a point source initiated flame*, Phys. D, 121 (1998), pp. 295–316.

[2] J. Audounet, J.-M. Roquejoffre, and H. Rouzaud, *Numerical simulation of point-source initiated flame ball with heat losses*, Math. Model. Numer. Anal., 36 (2002), pp. 273–291.

[3] C.-M. Brauner, J. Hulshof, and A. Lunardi, *A general approach to stability in free boundary problems*, J. Differential Equations, 164 (2000), pp. 16–48.

[4] C.-M. Brauner, J. Hulshof, A. Lunardi, and C. Schmidt-Lainé, *Instabilities, bifurcations and saddle-points in some free boundary problems in combustion*, in Free Boundary Problems: Theory and Applications (Crete, 1997), Chapman & Hall/CRC Res. Notes Math. 409, Chapman & Hall/CRC, Boca Raton, FL, 1999, pp. 105–114.

[5] J. D. Buckmaster, G. Joulin, and P. D. Ronney, *The structure and stability of non adiabatic flame balls*, Combust. Flame, 79 (1990), pp. 381–392.

[6] J. D. Buckmaster, G. Joulin, and P. D. Ronney, *The structure and stability of non adiabatic flame balls:* II. *Effects of far field losses*, Combust. Flame, 84 (1991), pp. 411–422.

[7] V. A. Galaktianov, J. Hulshof, and J. L. Vazquez, *Extinction and focusing of spherical and annular flames described by a free boundary problem*, J. Math. Pures Appl. (9), 76 (1997), pp. 563–608.

[8] V. Guyonne and L. Lorenzi, *Instability in a flame ball problem*, Discrete Contin. Dyn. Syst. Ser. B, 7 (2007), pp. 315–350.

[9] V. Guyonne, J. Hulshof, and J. B. Van den Berg, *Stability of Flame Balls for a Free Combustion Model with Radiative Transfer*, preprint, 2006.

[10] G. Joulin, *Point source initiation of lean spherical flames of light reactants: An asymptotic theory*, Comb. Sci. Tech., 43 (1985), pp. 99–113.

[11] C. Lederman, J.-M. Roquejoffre, and N. Wolansky, *Mathematical justification of a nonlinear integrodifferential equation for the propagation of spherical flames*, Ann. Mat. Pura Appl., 183 (2004), pp. 173–239.

[12] P. D. Ronney, *On the mechanisms of flame propagation limits and extinction processes at microgravity*, in Proceedings of the Twenty Second Symposium (International) on Combustion, Combustion Institute, 1988, pp. 1615–1623.

[13] P. D. Ronney, *Near-limit flame structures at low Lewis number*, Combust. Flame, 82 (1990), pp. 1–14.

[14] P. D. Ronney, M. S. Wu, H. G. Pearlman, and K. J. Weiland, *Experimental study of flame balls in space: Preliminary results from STS*-83, AIAA J., 36 (1998), pp. 1361–1368.

[15] H. Rouzaud, *Long-time dynamics of an integro-differential equation describing the evolution of a spherical flame*, Rev. Mat. Comput., 16 (2003), pp. 207–232.

[16] D. Sattinger, *Monotone methods in nonlinear elliptic and parabolic boundary value problems*, Indiana Univ. Math. J., 21 (1972), pp. 979–1000.

[17] A. A. Shah, R. W. Thachter, and J. W. Dold, *Stability of a spherical flame ball in a porous medium*, Combust. Theory Model., 4 (2000), pp. 511–534.

[18] J. B. Van den Berg, V. Guyonne, and J. Hulshof, *Flame balls for a free boundary combustion model with radiative transfer*, SIAM J. Appl. Math., 67 (2006), pp. 116–137.

[19] Ya. B. Zeldovich, G. I. Barenblatt, V. B. Librovich, and G. M. Makhviladze, *The Mathematical Theory of Combustion and Explosions*, Plenum, New York, 1985.

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 869–892

THE MINIMUM FREE ENERGY FOR CONTINUOUS SPECTRUM MATERIALS*

L. DESERI† AND J. M. GOLDEN‡

Abstract. A general closed expression is given for the isothermal minimum free energy of a linear viscoelastic material with continuous spectrum response. Two quite distinct approaches are adopted, which give the same final result. The first involves expressing a positive quantity, closely related to the loss modulus of the material, defined on the frequency domain, as a product of two factors with specified analyticity properties. The second is the continuous spectrum version of a method used in [S. Breuer and E. T. Onat, *Z. Angew. Math. Phys.*, 15 (1964), pp. 13–21] for materials with relaxation function given by sums of exponentials. It is further shown that minimal energy states are uniquely related to histories and that the work function is the maximum free energy with the property that it is a function of state.

Key words. minimum free energy, linear viscoelasticity, continuous spectrum materials, materials with memory, factorization, complex frequency plane

AMS subject classifications. Primary, 74A15, 74D05; Secondary, 30E20

DOI. 10.1137/050639776

1. Introduction.

A general expression for the minimum free energy of a linear viscoelastic material under isothermal conditions was given in [1]. This was for a scalar constitutive relation. A generalization to the full tensor case has also been presented [2]. Detailed, explicit expressions for the minimum free energy and related quantities were given in [1, 2] for discrete spectrum materials, namely those for which the relaxation function is a sum of exponentials. The minimum free energy of compressible viscoelastic fluids was determined in [3], while materials with finite memory were considered in [4]. These results are used in the context of the viscoelastic Saint-Venant problem in [5].

A definition of a viscoelastic state, based on the ideas of Noll [6], has been given and explored in [7, 8, 9]. Such a state has been termed a minimal state in [10]. Further related ideas and applications are explored in [11].

Also, a formalism has been developed [10] for the scalar case, which allows expressions for a family of free energies related to a particular minimal state to be derived for discrete spectrum models, including minimum and maximum free energies. Generalization of this work to the full tensor, nonisothermal case was presented in [12]. A generalization of the formalism in [10] has been used recently to propose a closed formula for the physical free energy and rate of dissipation.

It is not clear how the formulae emerging from the the methodology developed in [1, 2] apply to materials other than those exhibiting a discrete spectrum response, in particular for materials with a continuous spectrum response, i.e., those for which

*Received by the editors September 7, 2005; accepted for publication (in revised form) December 14, 2006; published electronically April 10, 2007.
http://www.siam.org/journals/siap/67-3/63977.html

†S.A.V.A. Department, Division of Engineering, Universita degli Studi del Molise, via De Sanctis 1, 86100 Campobasso, Italy (luca.deseri@ing.unimol.it). The research of this author was partly supported by the Italian Ministry M.I.U.R. grant P.R.I.N. 2005, by the CMA - Mathematical Sciences Department, Carnegie Mellon University, Pittsburgh, PA, and the Department of Theoretical and Applied Mechanics, Cornell University, Ithaca, NY, which are gratefully acknowledged.

‡School of Mathematics Sciences, Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland (murrough.golden@dit.ie).

the relaxation function is given by an integral of a density function multiplying a decaying exponential. The object of the present work is to address this issue. We will confine the treatment to the scalar case. There is no great loss of generality in doing so because, in the general tensor case, explicit solutions have been given only if the eigenspaces of the relaxation tensor derivative are time-independent, and on each such eigenspace the explicit results are precisely those of the scalar case [2, 12].

All the above papers are based on the same methodology, which involves factorizing a quantity closely related to the loss modulus of the material, in order to solve the relevant Wiener–Hopf equation (or equivalent variational problem) for the optimal future continuation required to determine the minimum free energy. Another method was used in [13] for the discrete spectrum case. This involved making a very natural assumption on the form of the optimal future continuation and solving algebraic equations for the various parameters. The need for factorization did not arise. This method is also developed in the present work for continuous spectrum materials. The assumption involved in this case is also a very natural one, namely that the optimal future continuation has a singularity structure determined only by that of the Fourier transform of the relaxation function derivative. This is analogous to the method in [13], i.e., to restrict the class of candidate functions when seeking to maximize the recoverable work.

The layout of the paper is as follows. In section 2, fundamental relationships are written down and the basic factorization property is introduced. The Wiener–Hopf equation relating to the maximum recoverable work is derived in section 3. The factorization procedure is discussed in depth for the continuous spectrum case in section 4, and some related formulae are considered in section 5. The minimum free energy is discussed in section 6. The alternative approach referred to in the previous paragraph is discussed in detail in section 7. The concept of a minimal state for continuous spectrum materials is explored in section 8. Some examples are presented in section 9.

2. Basic relationships. We consider a linear viscoelastic solid, subject to stress in such a way that there is only one nonzero component of stress $T(t)$ and strain $E(t)$ related by

$$\begin{aligned} T(t) &= G_0 E(t) + \int_0^\infty G'(s) E^t(s)\,ds, \qquad E^t(s) = E(t-s), \quad s \in \mathcal{R}, \\ &= G_\infty E(t) + \int_0^\infty G'(s) E_r^t(s)\,ds, \quad E_r^t(s) = E^t(s) - E(t), \end{aligned} \tag{2.1}$$

where $E^t \in L^1(\mathcal{R}^+) \cap L^2(\mathcal{R}^+) \cap C^1(\mathcal{R}^+)$ and $G' \in L^1(\mathcal{R}^+) \cap L^2(\mathcal{R}^+)$, using the following notation here and below: $\mathcal{R}$ is the set of reals, $\mathcal{R}^+$ the positive reals, and $\mathcal{R}^{++}$ the strictly positive reals; similarly $\mathcal{R}^-$, $\mathcal{R}^{--}$ are the negative and strictly negative reals. The relative history E_r^t will be used extensively later.[1] The relaxation function

$$G(s) = G_0 + \int_0^s G'(u)\,du \tag{2.2}$$

is well defined, along with $G_\infty = \lim_{s\to\infty} G(s)$. We take

$$G_\infty > 0, \tag{2.3}$$

[1] Note that this notation differs from that in [1].

so that the body is a solid.

A viscoelastic state is defined in general by the current value of strain and the history $(E(t), E^t)$. The concept of a minimal state ([10], based on ideas introduced in [6, 7, 9, 8, 2, 14]) can be expressed as follows: two viscoelastic states $(E_1(t), E_1^t)$, $(E_2(t), E_2^t)$ are equivalent or in the same minimal state if

$$E_1(t) = E_2(t), \qquad \int_0^\infty G'(s+\tau)\left[E_1^t(s) - E_2^t(s)\right] ds = 0 \quad \forall\, \tau \geq 0. \tag{2.4}$$

Let Ω be the complex ω plane and

$$\begin{aligned} \Omega^+ &= \{\omega \in \Omega \mid Im(\omega) \in \mathcal{R}^+\}, \\ \Omega^{(+)} &= \{\omega \in \Omega \mid Im(\omega) \in \mathcal{R}^{++}\}. \end{aligned} \tag{2.5}$$

These define the upper half-plane including and excluding the real axis, respectively. Similarly, Ω^-, $\Omega^{(-)}$ are the lower half-planes including and excluding the real axis, respectively.

For any $f \in L^2(\mathcal{R})$, its Fourier transform $f_F \in L^2(\mathcal{R})$ is given by

$$f_F(\omega) = \int_{-\infty}^{\infty} f(\xi) e^{-i\omega\xi} d\xi. \tag{2.6}$$

If f is a real-valued function in the time domain—which will be the case for all functions of interest here—then

$$\overline{f}_F(\omega) = f_F(-\omega), \tag{2.7}$$

where the bar denotes complex conjugate.

We have

$$\begin{aligned} f_F(\omega) &= f_+(\omega) + f_-(\omega), \\ f_+(\omega) &= \int_0^\infty f(\xi) e^{-i\omega\xi} d\xi, \\ f_-(\omega) &= \int_{-\infty}^0 f(\xi) e^{-i\omega\xi} d\xi, \qquad f_\pm \in L^2(\mathcal{R}), \end{aligned} \tag{2.8}$$

where f_+ has an analytic extension to $\Omega^{(-)}$, by virtue of the unique differentiability of its definition $(2.8)_2$ in terms of an integral. For the cases of interest in the present work, we also assume that it is analytic on an open set including Ω^-, so that we include $\mathcal{R}$ in the region of analyticity. Similarly, f_- is analytic on an open set which includes Ω^+. We will abbreviate these statements in what follows as "$f_\pm$ is analytic in $\Omega^\mp$."

The fact that the singularities of $f_\pm$ are restricted to $\Omega^{(\pm)}$, which is required for the derivation of the free energy [1], means that $f(\xi)$ decays exponentially at large $|\xi|$. This is a limitation in that it excludes, for example, power law decay. However, as we will discuss later, it is in many cases possible to extrapolate final results continuously up to the real axis, thereby removing the limitation to exponential decay.

We have

$$\lim_{\omega\to\infty} i\omega f_+(\omega) = f(0^+), \qquad \lim_{\omega\to\infty} i\omega f_-(\omega) = -f(0^-). \tag{2.9}$$

Functions on $\mathcal{R}$ which vanish identically on $\mathcal{R}^{--}$ are defined as functions on $\mathcal{R}^{+}$. For such quantities, $f_F = f_c - if_s$, where f_c, f_s are the Fourier cosine and sine transforms

$$\begin{aligned} f_c(\omega) &= \int_0^\infty f(\xi)\cos\omega\xi d\xi = f_c(-\omega), \\ f_s(\omega) &= \int_0^\infty f(\xi)\sin\omega\xi d\xi = -f_s(-\omega). \end{aligned} \tag{2.10}$$

Thus

$$F(\omega) = G'_F(\omega) = \int_0^\infty G'(s)e^{-i\omega s}ds = G'_c(\omega) - iG'_s(\omega). \tag{2.11}$$

The notation F is introduced to simplify later formulae. We shall require the property of F that

$$\lim_{\omega\to\infty} i\omega F(\omega) = G'(0^+) = G'(0), \tag{2.12}$$

which is a special case of $(2.9)_1$, with the added assumption that G' is continuous from the right, at the origin. Properties of $G'_s(\omega)$ include (see [15])

$$\begin{aligned} G'_s(\omega) &\le 0 \qquad \forall\, \omega \in \mathcal{R}^{++}, \\ G'_s(-\omega) &= -G'_s(\omega) \qquad \forall\, \omega \in \mathcal{R}, \end{aligned} \tag{2.13}$$

the first relation being a consequence of the second law of thermodynamics and the second being a particular case of (2.10). It follows that $G'_s(0) = 0$. We also have [15]

$$G_\infty - G_0 = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{G'_s(\omega)}{\omega} d\omega < 0, \tag{2.14}$$

so that $G'_s(\omega)/\omega \in L^1(\mathcal{R})$. It follows from (2.3) and (2.14) that G_0 is positive.

The function F is analytic on $\Omega^{(-)}$. This is a consequence of the fact that G' vanishes on $\mathcal{R}^{--}$, which is essentially the requirement of causality [16]. As noted above, it is assumed that F is analytic in Ω^{-}. Relation $(2.11)_1$ can be used to define $F(\omega)$ where the integral converges, namely Ω^{-} and possibly a strip of $\Omega^{(+)}$. Elsewhere, it is defined by analytic continuation from the region in which the integral exists. In fact, such continuation will generally be possible to all of $\Omega^{(+)}$, excluding singular points.

We let the bar denote complex conjugate. The quantity $\overline{F}(\omega)$ is the complex conjugate of the function, leaving the argument unchanged. For $\omega \in \mathcal{R}$, we have $F(-\omega) = \overline{F}(\omega)$. The quantity $\overline{F}$ is analytic in Ω^{+}, with a mirror image, in the real axis, of the singularity structure of $F(\omega)$. Thus, $G'_s(\omega)$ has singularities in both $\Omega^{(+)}$ and $\Omega^{(-)}$, which are mirror images of one another. Similarly, its zeros will be mirror images of each other. We will be interested in the singularity structure of

$$\begin{aligned} H(\omega) &= \frac{\omega}{2i}\left(F(\omega) - \overline{F}(\omega)\right) \\ &= -\omega G'_s(\omega) = H(-\omega) \ge 0 \qquad \forall\, \omega \in \mathcal{R}, \\ H(\omega) &= H_1(\omega^2), \end{aligned} \tag{2.15}$$

where H_1 is the function H expressed in terms of ω^2. This last relation is a consequence of the analyticity of $H(\omega)$ on the real axis and its evenness property. It follows

that $H(\omega)$ goes to zero at least quadratically at the origin. It is assumed that the behavior is in fact quadratic; i.e., $H(\omega)/\omega^2$ tends to a finite nonzero quantity as ω tends to zero. Note that $H(\omega)$ is nonnegative on the real axis. For ω off the real axis, it is defined by analytic continuation from (2.15) and is in general a complex quantity. Its singularities are the same as those of F in $\Omega^{(+)}$ and of $\overline{F}$ in $\Omega^{(-)}$. We will need the following relationship:

$$\int_{-\infty}^{\infty} \frac{d}{ds} G(|s|) e^{-i\omega s} ds = -2iG'_s(\omega) = 2i\frac{H(\omega)}{\omega}, \tag{2.16}$$

giving the Fourier transform of the odd extension of G' to $\mathcal{R}$.

It will be required in later developments that $H(\omega)$ can be written in the form

$$H(\omega) = H_+(\omega) H_-(\omega), \tag{2.17}$$

where $H_+(\omega)$ has no singularities or zeros in $\Omega^{(-)}$ and is thus analytic in Ω^-. Similarly, $H_-(\omega)$ is analytic in Ω^+ with no zeros in $\Omega^{(+)}$. Therefore the singularities of F must all occur in H_+ and those of $\overline{F}$ in H_-. There may be other singularities in $H_\pm$ which cancel on multiplication. That such a factorization is always possible is shown for general tensor constitutive relations in [2].

Using (2.12) and (2.15), one can show that

$$H_\infty = \lim_{|\omega|\to\infty} H(\omega) = -G'(0) \geq 0. \tag{2.18}$$

The sign of $G'(0)$ has been deduced by various authors from thermodynamic constraints in the general three-dimensional case [17, 18, 15]. We assume for present purposes that $G'(0)$ is nonzero, so that H_∞ is a finite positive number. Then $H(\omega) \in \mathcal{R}^{++}$ $\forall\, \omega \in \mathcal{R}$, $\omega \neq 0$.

The factorization (2.17) is unique up to a constant phase factor. We set [1]

$$\begin{aligned} H_\pm(\omega) &= H_\mp(-\omega) = \overline{H}_\mp(\omega), \\ H(\omega) &= |H_\pm(\omega)|^2, \end{aligned} \tag{2.19}$$

one consequence of which is that the factorization is now unique up to a change of sign.

A general method is outlined in [1] for determining the factors of H. A modification of this method is presented here, which is more convenient for the present application. Consider the function $T(\omega)H(\omega)$ [1], where[2]

$$T(\omega) = \frac{\omega^2 + \omega_0^2}{H_\infty \omega^2}. \tag{2.20}$$

This product is nonnegative on $\mathcal{R}$, is nonsingular at the origin, and approaches unity for large ω. The frequency $\omega_0 \in \mathcal{R}^{++}$ may be chosen arbitrarily. Therefore, the function $\log(T(\omega)H(\omega))$ is well defined on $\mathcal{R}$ and approaches zero for large ω. Let [1]

$$\begin{aligned} H_+(\omega) &= \frac{\omega h_\infty}{\omega - i\omega_0} e^{-M^+(\omega)}, \\ h_\infty &= H_\infty^{1/2}, \end{aligned} \tag{2.21}$$

[2]The introduction of the parameter ω_0 represents a slight modification (improvement) of the formula in [1].

where M^+ is given by[3]

$$
\begin{aligned}
M^+(\omega) &= \lim_{\beta\to 0^-} M(\omega + i\beta), \\
M(z) &= \frac{1}{2\pi i}\int_{-\infty}^{\infty} \frac{\log[T(\omega')H(\omega')]}{\omega' - z} d\omega', \quad z \in \Omega \setminus \mathcal{R}.
\end{aligned} \tag{2.22}
$$

Using (2.15), we can write

$$
\begin{aligned}
\log\left(T(\omega)H(\omega)\right) &= \log\left[-i\frac{\omega - i\omega_0}{H_\infty}F(\omega)\right] + \log U(\omega), \\
U(\omega) &= \frac{1}{2}\left[1 - \frac{\overline{F}(\omega)}{F(\omega)}\right]\left[\frac{\omega + i\omega_0}{\omega}\right].
\end{aligned} \tag{2.23}
$$

The standard branch of the logarithm function is chosen, namely that which vanishes for argument unity. The function U is complex but nonzero on the real line and approaches unity for large ω, by virtue of (2.12). Similarly for the argument of the first term on the right of $(2.23)_1$. This term has all its singularities in $\Omega^{(+)}$ so that if we close on $\Omega^{(-)}$ for $Im\, z < 0$ then, by Cauchy's theorem, its contribution to $M(z)$ is simply the negative of itself. Thus, we have

$$
\begin{aligned}
H_+(\omega) &= \frac{-i\omega}{h_\infty}F(\omega)e^{-N^+(\omega)}, \quad \omega \in \mathcal{R}, \\
N^+(\omega) &= \lim_{\beta\to 0^-} N(\omega + i\beta), \\
N(z) &= \frac{1}{2\pi i}\int_{-\infty}^{\infty} \frac{\log U(\omega')}{\omega' - z} d\omega', \quad z \in \Omega \setminus \mathcal{R}.
\end{aligned} \tag{2.24}
$$

Using the relation $(2.19)_1$, we deduce that

$$
\begin{aligned}
H_-(\omega) &= \frac{i\omega}{h_\infty}\overline{F}(\omega)e^{-N^-(\omega)}, \quad \omega \in \mathcal{R}, \\
N^-(\omega) &= \lim_{\beta\to 0^-} \overline{N}(\omega - i\beta), \\
\overline{N}(z) &= -\frac{1}{2\pi i}\int_{-\infty}^{\infty} \frac{\log \overline{U}(\omega')}{\omega' - z} d\omega', \quad z \in \Omega \setminus \mathcal{R}.
\end{aligned} \tag{2.25}
$$

The extraction of the factor F in (2.24) and $\overline{F}$ in (2.25) has the advantage that the correct behavior of $H_\pm$ at small and large ω is assured. This is of course true of (2.21), but the apparent singularity at $\omega = i\omega_0$ must be eliminated, and the procedures for doing this after the transformations described in section 4 have been carried out is not straightforward. Using (2.24), (2.25), the parameter ω_0 drops out of the formulae in a simple manner, as we shall see later. Furthermore, the singularities of F in $\Omega^{(+)}$ must occur also in H_+ (though, in fact, H_+ may have other singularities), while a similar statement applies to H_- and $\overline{F}$. If the transformation carried out on N^+ in section 4 were instead carried out on M^+ (and this is the natural first approach), it is in fact rather difficult to ensure that the singularity structures of $H_\pm$ in $\Omega^\pm$ are correct. This is particularly true of logarithmic singularities which can, as we shall see, occur at the end points of the branch cuts.

[3]The quantity $M^+(\omega)$ was denoted by $M^-(\omega)$ in [1] and vice versa. The present usage is more consistent with the rest of the paper.

Consider now the strain history E^t. Define

$$E_+^t(\omega) = \int_0^\infty E^t(s)e^{-i\omega s}ds, \quad E_+^t \in L^2(\mathcal{R}^+). \tag{2.26}$$

It is analytic in $\Omega^{(-)}$, a property which will be assumed to extend to Ω^-. This region can be extended to include $\Omega^{(+)}$, excluding singular points, by analytic continuation. From $(2.9)_1$, it follows that

$$\lim_{\omega\to\infty} i\omega E_+^t(\omega) = E^t(0^+) = E(t). \tag{2.27}$$

We also require the Fourier transform of the relative history,

$$\begin{aligned} E_{r+}^t(\omega) &= E_+^t(\omega) - E(t)\int_0^\infty e^{-i\omega s}ds \\ &= E_+^t(\omega) - \frac{E(t)}{i\omega^-}, \qquad \omega^- = \lim_{\alpha\to 0^+}(\omega - i\alpha), \end{aligned} \tag{2.28}$$

where the limit is taken after any integration involving the quantity $(\omega^-)^{-1}$ has been carried out; for purposes of such an integration, ω^- is in $\Omega^{(-)}$. Under an assumption similar to that for E_+^t, we have that E_{r+}^t is analytic on Ω^-. Note that, by virtue of $(2.9)_1$, E_{r+}^t goes to zero at large ω as ω^{-2}.

Let us also define

$$E_-^t(\omega) = \int_{-\infty}^0 E^t(s)e^{-i\omega s}ds, \quad E_-^t \in L^2(\mathcal{R}^-). \tag{2.29}$$

It is analytic in $\Omega^{(+)}$, a property which will be assumed to extend to Ω^+. It is defined on $\Omega^{(-)}$, excluding singular points, by analytic continuation. From (2.9), it follows that

$$\lim_{\omega\to\infty} i\omega E_-^t(\omega) = -E^t(0^-) = -E(t^+). \tag{2.30}$$

We also require the Fourier transform of the relative history,

$$\begin{aligned} E_{r-}^t(\omega) &= E_-^t(\omega) - E(t)\int_{-\infty}^0 e^{-i\omega s}ds \\ &= E_-^t(\omega) + \frac{E(t^+)}{i\omega^+}, \qquad \omega^+ = \lim_{\alpha\to 0^+}(\omega + i\alpha), \end{aligned} \tag{2.31}$$

where for any integration involving the quantity $(\omega^+)^{-1}$ the singularity is in $\Omega^{(+)}$. The limit to the real axis is taken after any such integration has been carried out. Under an assumption similar to that for E_-^t, we have that E_{r-}^t is analytic in Ω^+ and E_{r+}^t goes to zero at large ω as ω^{-2}.

3. The maximum recoverable work and the Wiener–Hopf equation. The total work done on the material up to time t is given by [2]

$$\begin{aligned} W(t) &= \int_{-\infty}^t T(s)\dot{E}(s)ds = \phi(t) + \frac{1}{2}\int_0^\infty\int_0^\infty E_r^t(s)G_{12}(|s-u|)E_r^t(u)dsdu \\ &= \phi(t) + \frac{1}{2\pi}\int_{-\infty}^\infty H(\omega)\left|E_{r+}^t(\omega)\right|^2 d\omega, \end{aligned} \tag{3.1}$$

where ϕ has the form

$$\phi(t) = \frac{1}{2} G_\infty E^2(t). \tag{3.2}$$

This quantity is the equilibrium free energy. Also [19]

$$G_{12}(|s-u|) = \frac{\partial^2}{\partial s \partial u} G(|s-u|) = -2\delta(s-u)G'(|s-u|) - G''(|s-u|), \tag{3.3}$$

in terms of the singular delta function. The form $(3.1)_3$ follows from (2.16) and the convolution theorem.

The maximum recoverable work from a given state of a material with memory is equal to the minimum free energy of that state, as can be shown under very general conditions (e.g., [20] and references therein). Thus, we seek to maximize the integral

$$W_r(t) = -\int_t^\infty T(s)\dot{E}(s)ds, \tag{3.4}$$

or to minimize

$$W(\infty) = \int_{-\infty}^\infty T(s)\dot{E}(s)ds, \tag{3.5}$$

where E is varied only on $[t, \infty)$. When taking the variation, we can assume that $E(\infty)$ vanishes [2]. It follows from (3.1), on changing the integration range to $(-\infty, t]$, that

$$W(\infty) = \frac{1}{2}\int_{-\infty}^\infty \int_{-\infty}^\infty E(s)G_{12}(|s-u|)E(u)dsdu. \tag{3.6}$$

It is easily deduced that the optimization condition is

$$\int_{-\infty}^\infty G_{12}(|s-u|)E(u)du = \int_{-\infty}^\infty G_{12}(|s-u|)E_r^t(u)du = 0, \qquad s \in \mathcal{R}^-. \tag{3.7}$$

We can remove the derivative with respect to s since G_2 tends to zero as s tends to infinity. Also, the derivative with respect to u can be replaced by a derivative with respect to s. Thus, we obtain the relation

$$\int_{-\infty}^\infty \frac{\partial}{\partial s} G(|s-u|)E_r^t(u)du = 0, \qquad s \in \mathcal{R}^-, \tag{3.8}$$

or

$$\int_{-\infty}^0 \frac{\partial}{\partial s} G(|s-u|)E_o^t(u)du = \int_0^\infty G'(u-s)E_r^t(u)du, \qquad s \in \mathcal{R}^-, \tag{3.9}$$

where $E_o^t : \mathcal{R}^- \mapsto \mathcal{R}$ is the future continuation which yields the maximum recoverable work.

4. Factorization of H for continuous spectrum materials. We adopt the following continuous spectrum form for the relaxation function derivative:

$$G'(t) = \int_a^b k(\alpha)e^{-\alpha t}d\alpha, \quad t \in \mathcal{R}^+, \quad b > a > 0. \tag{4.1}$$

It is assumed that $k \in L^1([a,b])$. The upper limit b may be infinite. We take $a > 0$ because of the need to avoid singularities on the real axis. The limit $a \to 0$ is discussed in section 6. We take the Fourier transform of (4.1) to obtain

$$F(\omega) = \int_a^b \frac{k(\alpha)}{\alpha + i\omega} d\alpha, \quad \omega \in \mathcal{R}. \tag{4.2}$$

This formula can be extended by analytic continuation to Ω, excluding singular points. We restrict the density function k to be Hölder continuous on (a,b). It may be singular at the end points with a power less than unity. It is assumed that

$$k(\alpha) \leq 0, \quad \alpha \in [a,b]. \tag{4.3}$$

This assumption is not essential but is the simplest which ensures compatibility with thermodynamic constraint $(2.13)_1$. Note that it renders G completely monotonic in the sense discussed in [9]. Also, it is easily shown that F has no zeros on the finite part of $\Omega \setminus [ia, ib]$. Taking the complex conjugate of (4.2), we have

$$\overline{F}(\omega) = \int_a^b \frac{k(\alpha)}{\alpha - i\omega} d\alpha, \quad \omega \in \mathcal{R}, \tag{4.4}$$

which can similarly be continued into the complex plane.

The quantity F has a branch cut on $[ia, ib]$ and $\overline{F}$ on $[-ia, -ib]$. As ω tends to $i\alpha$, where $\alpha \in \mathcal{R} \setminus [-a, -b]$,

$$\overline{F}(i\alpha) = F(-i\alpha) = \int_a^b \frac{k(\beta)}{\beta + \alpha} d\beta = K(\alpha), \tag{4.5}$$

while if $\alpha \in (a,b)$, we have, by virtue of the Plemelj formulae [21],

$$\begin{aligned} F_R(i\alpha) &= R(\alpha) + iI(\alpha), \\ F_L(i\alpha) &= R(\alpha) - iI(\alpha), \end{aligned} \tag{4.6}$$

with

$$R(\alpha) = P \int_a^b \frac{k(\beta)}{\beta - \alpha} d\beta, \qquad I(\alpha) = -\pi k(\alpha) \geq 0, \tag{4.7}$$

where $F_R(i\alpha)$, $F_L(i\alpha)$ are the limiting values of $F(\omega)$, approaching from the right and the left, respectively, as one moves from ia to ib. Similarly,

$$\begin{aligned} \overline{F}_R(-i\alpha) &= R(\alpha) + iI(\alpha), \\ \overline{F}_L(-i\alpha) &= R(\alpha) - iI(\alpha), \end{aligned} \tag{4.8}$$

for $\alpha \in (a,b)$, where $\overline{F}_R(-i\alpha)$, $\overline{F}_L(-i\alpha)$ are the limiting values of $\overline{F}(\omega)$ from the right and left, respectively, as one moves from $-ia$ to $-ib$. The symbol P in (4.7) indicates a principal value.

From (2.15), we have

$$H(\omega) = -\omega^2 \int_a^b \frac{k(\alpha)}{\alpha^2 + \omega^2} d\alpha. \tag{4.9}$$

Let us consider the behavior of $F(\omega)$ at the end points ia and ib for various limiting behaviors of $k(\alpha)$ as α approaches a or b [21]. If $k(a) = 0$, then $F(\omega)$ has a definite finite nonzero limit as $\omega \to ia$. A similar statement applies to the limit $\omega \to ib$ if $k(b) = 0$.

If

$$k(a) = k_a < 0 \tag{4.10}$$

and k is Hölder continuous near and at a, then $F(\omega)$ has a logarithmic singularity at $\omega = ia$. As ω approaches this end point along any path off $[ia, ib]$, then

$$F(\omega) = k_a \log \frac{1}{a + i\omega} + F_1(\omega), \tag{4.11}$$

where $F_1(a)$ is well defined. Similarly, if

$$k(b) = k_b < 0 \tag{4.12}$$

and k is Hölder continuous near and at b, then, as ω approaches ib, not along $[ia, ib]$, we have

$$F(\omega) = -k_b \log \frac{1}{b + i\omega} + F_2(\omega), \tag{4.13}$$

where $F_2(b)$ is well defined. For points on (ia, ib), relations (4.11) and (4.13) are replaced by

$$\begin{aligned} R(\alpha) &\underset{\alpha \to ia^+}{\longrightarrow} k_a \log \frac{1}{\alpha - a} \\ &\underset{\alpha \to ib^-}{\longrightarrow} -k_b \log \frac{1}{b - \alpha}, \end{aligned} \tag{4.14}$$

where $R(\alpha)$ is given by (4.7). If $k(\alpha)$ has dominant behavior as $\alpha \to a^+$ along (a, b) of the form

$$k(\alpha) \underset{\alpha \to a^+}{\longrightarrow} \frac{k_1}{(\alpha - a)^\gamma}, \quad 0 < \gamma < 1, \quad k_1 < 0, \tag{4.15}$$

then for $\omega \notin (ia, ib)$

$$F(\omega) \underset{\omega \to ia}{\longrightarrow} \frac{Ak_1}{(a + i\omega)^\gamma}. \tag{4.16}$$

The detailed form of A is given in [21]. A similar observation applies to the case where k has such behavior at b. For points on (ia, ib), relation (4.16) is replaced by

$$R(\alpha) \underset{\alpha \to a^+}{\longrightarrow} \frac{A_1 k_1}{(\alpha - a)^\gamma}, \tag{4.17}$$

where again the form of A_1 may be found in [21]. A similar observation applies at b.

We return our attention to (2.24). The function $U(i\alpha)$, $\alpha > 0$, is real for $\alpha \notin [a, b]$. It is discontinuous across $[a, b]$. We define, for $\alpha \in [a, b]$,

$$\begin{aligned} U_R(i\alpha) &= \lim_{\omega \to \omega_R} U(\omega), \qquad \omega_R = \alpha e^{\frac{i\pi}{2}}, \\ U_L(i\alpha) &= \lim_{\omega \to \omega_L} U(\omega), \qquad \omega_L = \alpha e^{\frac{-3i\pi}{2}}. \end{aligned} \tag{4.18}$$

As noted earlier, the function $U(\omega)$ is nonzero on $\Omega^{(+)}$ and approaches unity as $\omega \to \infty$. Thus, $\log U(\omega)$ has a branch cut on $[ia, ib]$ and no other singularity in $\Omega^{(+)}$. The factor $\log\left[\frac{\omega+i\omega_0}{\omega}\right]$ is assigned a branch cut on $[0, -i\omega_0]$. Moving the line of integration in $(2.24)_3$ to the infinite half-circle in $\Omega^{(+)}$ while going around the branch cut, we obtain

$$\begin{aligned} N(z) &= \frac{1}{2\pi i}\int_a^b \frac{\Delta(\alpha)}{\alpha + iz} d\alpha, \\ \Delta(\alpha) &= \log U_R(i\alpha) - \log U_L(i\alpha), \end{aligned} \tag{4.19}$$

where the branch of the logarithm function is as specified earlier. Its imaginary part lies in $[-\pi, \pi]$. Note that the factor $\left[\frac{\omega+i\omega}{\omega}\right]$ in $U(\omega)$ cancels out of $\Delta(\alpha)$; it can henceforth be omitted. Thus, we set

$$\begin{aligned} Y(\omega) &= \frac{1}{2}\left(1 - \frac{\overline{F}(\omega)}{F(\omega)}\right), \\ \Delta(\alpha) &= \log Y_R(i\alpha) - \log Y_L(i\alpha), \end{aligned} \tag{4.20}$$

where, from (4.5) and (4.6),

$$\begin{aligned} Y_R(i\alpha) &= \frac{1}{2}\left[1 - \frac{K(\alpha)}{R(\alpha) + iI(\alpha)}\right], \\ Y_L(i\alpha) &= \frac{1}{2}\left[1 - \frac{K(\alpha)}{R(\alpha) - iI(\alpha)}\right] = \overline{Y_R(i\alpha)}. \end{aligned} \tag{4.21}$$

We can write

$$\Delta(\alpha) = 2iA(\alpha), \quad A(\alpha) = \arg Y_R(\alpha), \quad -\pi \le A(\alpha) \le \pi, \tag{4.22}$$

and

$$\begin{aligned} H_+(\omega) &= -\frac{i\omega}{h_\infty} F(\omega) e^{-N^+(\omega)}, \\ N^+(\omega) &= \frac{1}{\pi}\int_a^b \frac{A(\alpha)}{\alpha + i\omega} d\alpha, \end{aligned} \tag{4.23}$$

while

$$\begin{aligned} H_-(\omega) &= \frac{i\omega}{h_\infty} \overline{F}(\omega) e^{-N^-(\omega)}, \\ N^-(\omega) &= \frac{1}{\pi}\int_a^b \frac{A(\alpha)}{\alpha - i\omega} d\alpha. \end{aligned} \tag{4.24}$$

In the notation of (4.7), we have

$$V(\alpha) = 2[R(\alpha)^2 + I(\alpha)^2] Re Y_R(i\alpha) = R(\alpha)^2 + I(\alpha)^2 - K(\alpha)R(\alpha), \tag{4.25}$$

where $K(\alpha)$, given by (4.5), is real and negative for $\alpha > -a$. Also,

$$W(\alpha) = 2[R(\alpha)^2 + I(\alpha)^2] Im Y_R(i\alpha) = K(\alpha)I(\alpha) \le 0, \tag{4.26}$$

from which it follows that $-\pi \le A(\alpha) \le 0$. Then

$$\begin{aligned} A(\alpha) &= -B(\alpha), \quad V(\alpha) \ge 0, \\ &= -\pi + B(\alpha), \quad V(\alpha) < 0; \\ B(\alpha) &= \arctan\left|\frac{W(\alpha)}{V(\alpha)}\right|, \quad 0 \le B(\alpha) \le \frac{\pi}{2}. \end{aligned} \tag{4.27}$$

Note the following result.

PROPOSITION 4.1. *The quantity*

$$V(\alpha) = R(\alpha)^2 + I(\alpha)^2 - K(\alpha)R(\alpha), \quad \alpha \in (a,b), \tag{4.28}$$

is nonnegative in the vicinity of the end points a and b. It is also nonnegative when $R(\alpha) \in \mathcal{R}^+$.

Proof. The latter statement follows immediately from the fact that $K(\alpha) \le 0$ for $\alpha \in (a,b)$. The statement is trivially true when $R(\alpha)$ vanishes. Nonnegativity near a given end point is manifestly true if R is unbounded at that end point, which is true even if k is finite but nonzero at the end points (see (4.14)). Thus, we must consider only the case where the density function k vanishes at the end point. Consider first the lower end point a. We have

$$R(a) = P\int_a^b \frac{k(\beta)}{\beta - a}d\beta \le 0. \tag{4.29}$$

Then

$$V(a) \ge R(a)^2 - K(a)R(a) \ge 0 \tag{4.30}$$

if

$$-R(a) \ge -K(a). \tag{4.31}$$

Observing that

$$K(a) = \int_a^b \frac{k(\beta)}{\beta + a}d\beta, \tag{4.32}$$

we see that (4.31) is true. Also,

$$R(b) = \int_a^b \frac{k(\beta)}{\beta - b}d\beta \ge 0, \tag{4.33}$$

so that $V(b) \ge 0$. □

If $V \ge 0$ on (a,b), then

$$\begin{aligned} H_+(\omega) &= -\frac{i\omega}{h_\infty}F(\omega)\exp\left\{\frac{1}{\pi}\int_a^b \frac{B(\alpha)d\alpha}{\alpha + i\omega}\right\}, \\ H_-(\omega) &= \frac{i\omega}{h_\infty}\overline{F}(\omega)\exp\left\{\frac{1}{\pi}\int_a^b \frac{B(\alpha)d\alpha}{\alpha - i\omega}\right\}, \end{aligned} \tag{4.34}$$

where B is defined by $(4.27)_3$.

5. Some consequences of the factorization formulae. It is of interest to consider the limits of $H_{\pm}$, given by (4.23) and (4.24), as ω approaches the branch cuts on $[ia, ib]$ and $[-ia, -ib]$. Consider (4.24) as $\omega \to -i\alpha$, $\alpha \in (a, b)$, from the left, i.e., from the fourth quadrant. Noting (4.8), we obtain

$$\begin{aligned} H_{-L}(-i\alpha) &= \frac{\alpha}{h_\infty}(R(\alpha) - iI(\alpha))P(\alpha)e^{-iA(\alpha)}, \\ P(\alpha) &= \exp\left\{-\frac{1}{\pi}P\int_a^b \frac{A(\beta)}{\beta - \alpha}d\beta\right\}, \end{aligned} \tag{5.1}$$

where the Plemelj formulae have been used. Also, from (4.5) and (4.23),

$$\begin{aligned} H_{+L}(-i\alpha) &= -\frac{\alpha}{h_\infty}K(\alpha)Q(\alpha) = H_{+R}(-i\alpha), \\ Q(\alpha) &= \exp\left\{-\frac{1}{\pi}\int_a^b \frac{A(\beta)}{\beta + \alpha}d\beta\right\}. \end{aligned} \tag{5.2}$$

Multiplying $H_{\pm L}$ together, we obtain the limit of $H(\omega)$ as $\omega \to -i\alpha$, $\alpha \in (a, b)$, namely

$$H_L(-i\alpha) = -\frac{\alpha^2}{H_\infty}(R(\alpha) - iI(\alpha))K(\alpha)P(\alpha)Q(\alpha)e^{-iA(\alpha)}. \tag{5.3}$$

Also, from (2.15), we have

$$H_L(-i\alpha) = \frac{\alpha}{2}(R(\alpha) - K(\alpha) - iI(\alpha)). \tag{5.4}$$

Equating the arguments of these two expressions for $H_L(-i\alpha)$ gives

$$\arg(R(\alpha) - K(\alpha) - iI(\alpha)) = -A(\alpha) + \arg(R(\alpha) - iI(\alpha)), \tag{5.5}$$

or, taking complex conjugates,

$$A(\alpha) = \arg\left[1 - \frac{K\alpha)}{R(\alpha) + iI(\alpha)}\right], \tag{5.6}$$

which is, of course, simply $(4.22)_2$. Equating the magnitudes of the two expressions given by (5.3) and (5.4), we obtain

$$-2\alpha K(\alpha)P(\alpha)Q(\alpha) = H_\infty\sqrt{\frac{(R(\alpha) - K(\alpha))^2 + I^2(\alpha)}{R^2(\alpha) + I^2(\alpha)}}. \tag{5.7}$$

With the aid of (5.5), we can write (5.1) in the form

$$\begin{aligned} H_{-L}(-i\alpha) &= \frac{\alpha}{h_\infty}(R(\alpha) - K(\alpha) - iI(\alpha))\sqrt{\frac{R^2(\alpha) + I^2(\alpha)}{(R(\alpha) - K(\alpha))^2 + I^2(\alpha)}}P(\alpha) \\ &= -\frac{h_\infty}{2}\frac{(R(\alpha) - K(\alpha) - iI(\alpha))}{K(\alpha)Q(\alpha)}. \end{aligned} \tag{5.8}$$

The second form follows from (5.7).

Finally, we observe that $(2.15)_1$, (2.17), (4.23), and (4.24) give

$$Z(\omega) = \frac{H_\infty}{2i\omega}\left(\frac{1}{\overline{F}(\omega)} - \frac{1}{F(\omega)}\right) = \exp\left\{-\frac{1}{\pi}\int_a^b \frac{A(\alpha)d\alpha}{\alpha + i\omega} - \frac{1}{\pi}\int_a^b \frac{A(\alpha)d\alpha}{\alpha - i\omega}\right\}. \tag{5.9}$$

Let us show this directly, noting that the left-hand side does not vanish at the origin and is unity at infinity. Consider the contour C, taken clockwise at infinity except that it excludes the positive imaginary axis above ia and the negative imaginary axis below $-ia$. The quantity Z is finite and nonzero within C. Then we see that

$$Z(\omega) = \exp\left\{-\frac{1}{2\pi i}\int_C \frac{\log(Z(u))du}{u - \omega}\right\}, \tag{5.10}$$

where ω is in the interior of C. Invoking an argument similar to that leading to (4.19) and (4.20), the result follows on noting that

$$\log Z_R(i\alpha) - \log Z_L(i\alpha) = \log Y_R(i\alpha) - \log Y_L(i\alpha) \tag{5.11}$$

for $\alpha \in (a, b)$, since real positive factors in the arguments of the logarithms cancel.

6. The minimum free energy. First, we derive the expression for the continuation that yields the maximum recoverable work—which is equal to the minimum free energy—when F is given by (4.2), from the Wiener–Hopf equation (3.8) or (3.9). The unique factorization of H is given by (4.23), (4.24), and (4.27). We shall use the formalism for relative histories, defined by $(2.1)_2$, as in [3, 4, 12, 5] rather than in [1, 2]. Let us replace the right-hand side of (3.8) by $R^t(s)$, where this function vanishes on $\mathcal{R}^-$. Taking the Fourier transform of (3.8) yields

$$\frac{2i}{\omega}H(\omega)(E^t_{r+}(\omega) + E^t_m(\omega)) = R^t_+(\omega), \tag{6.1}$$

where E^t_m is the Fourier transform of the optimum relative continuation E^t_o in (3.9) and R^t_+ is an unknown function, analytic in Ω^-. Equation (6.1) is an immediate consequence of (2.16) and (2.15). Using the factorization property of H, we can write (6.1) as

$$H_-(\omega)E^t_{r+}(\omega) + H_-(\omega)E^t_m(\omega) = \frac{\omega R^t_+(\omega)}{2iH_+(\omega)}. \tag{6.2}$$

Let us define

$$\begin{aligned} p^t(z) &= \frac{1}{2\pi i}\int_{-\infty}^{\infty} d\omega' \frac{H_-(\omega')E^t_{r+}(\omega')}{\omega' - z}, \\ p^t_\pm(\omega) &= \lim_{\alpha \to 0^+} p^t(\omega \mp i\alpha). \end{aligned} \tag{6.3}$$

By the Plemelj formulae [21],

$$H_-(\omega)E^t_{r+}(\omega) = p^t_-(\omega) - p^t_+(\omega). \tag{6.4}$$

Then (6.2) can be written in the form

$$p^t_-(\omega) + H_-(\omega)E^t_m(\omega) = p^t_+(\omega) + \frac{\omega R^t_+(\omega)}{2iH_+(\omega)}. \tag{6.5}$$

Recalling that E_m^t is analytic in Ω^+ (see after (2.29)), we see that the left-hand side of this relation is analytic in Ω^+ and the right-hand side is analytic in Ω^-. Also, the left-hand side goes to zero as ω^{-1}, as we see by applying (2.30) to E_m^t. Thus, both sides are analytic in Ω and vanish at infinity, and are therefore individually zero. Therefore

$$E_m^t(\omega) = -\frac{p_-^t(\omega)}{H_-(\omega)}, \tag{6.6}$$

and the minimum free energy is given by (see [1, 2, 4], for example)

$$\begin{aligned} \psi_m(t) &= \phi(t) + \frac{1}{2\pi}\int_{-\infty}^{\infty} H(\omega)\left|E_m^t(\omega)\right|^2 d\omega \\ &= \phi(t) + \frac{1}{2\pi}\int_{-\infty}^{\infty} \left|p_-^t(\omega)\right|^2 d\omega, \end{aligned} \tag{6.7}$$

where ϕ is given by (3.2). The quantity $\psi_m(t)$ was shown in [1, 2] to be a free energy by the Graffi definition [22, 23] and in [2] by the Coleman–Owen definition [24, 25] for the general tensor case.

From $(3.1)_3$ and (6.4) we have

$$\begin{aligned} W(t) &= \phi(t) + \frac{1}{2\pi}\int_{-\infty}^{\infty} \left[\left|p_+^t(\omega)\right|^2 + \left|p_-^t(\omega)\right|^2\right] d\omega \\ &= \psi_m(t) + \frac{1}{2\pi}\int_{-\infty}^{\infty} \left|p_+^t(\omega)\right|^2 d\omega, \end{aligned} \tag{6.8}$$

where the orthogonality property [1]

$$\int_{-\infty}^{\infty} p_-^t(\omega)\bar{p}_+^t(\omega)d\omega = \int_{-\infty}^{\infty} p_+^t(\omega)\bar{p}_-^t(\omega)d\omega \tag{6.9}$$

was used in writing $(6.8)_1$. This follows from Cauchy's theorem, since $\bar{p}_\pm^t$ are analytic in $\Omega^\pm$ and go to zero as ω^{-1} at large ω.

Note that p_-^t can be written in the form

$$\begin{aligned} p_-^t(\omega) &= \frac{1}{2\pi}\int_a^b \frac{\Delta_h(\alpha)E_{r+}^t(-i\alpha)}{\alpha - i\omega}d\alpha, \\ \Delta_h(\alpha) &= -i(H_{-L}(-i\alpha) - H_{-R}(-i\alpha)), \end{aligned} \tag{6.10}$$

by closing the contour on $\Omega^{(-)}$ around the branch cut and changing variables. The quantity H_{-L} is given by (5.8), while H_{-R} is its complex conjugate. Thus, we have

$$\begin{aligned} \Delta_h(\alpha) &= -\frac{2\alpha}{h_\infty} I(\alpha)P(\alpha)\sqrt{\frac{R^2(\alpha) + I^2(\alpha)}{(R(\alpha) - K(\alpha))^2 + I^2(\alpha)}} \\ &= h_\infty \frac{I(\alpha)}{K(\alpha)Q(\alpha)} \le 0, \qquad \alpha \in [a, b]. \end{aligned} \tag{6.11}$$

The second form has the advantage that the need to evaluate a principal value integral is avoided. The quantities involved are also free of end point singularities.

The definitions of the various quantities in these relationships are summarized for convenience in Table 1.

TABLE 1
Definitions of the various quantities in the formula (6.11).

Formula	Equation reference
$F(\omega) = \int_a^b \frac{k(\alpha)}{\alpha + i\omega} d\alpha, \quad \omega \in \mathcal{R}$	(4.2)
$K(\alpha) = \int_a^b \frac{k(\beta)}{\beta + \alpha} d\beta, \quad \alpha \in \mathcal{R} \backslash [-a, -b]$	(4.5)
$R(\alpha) = P\int_a^b \frac{k(\beta)}{\beta - \alpha} d\beta, \quad I(\alpha) = -\pi k(\alpha), \quad \alpha \in (a, b)$	(4.7)
$A(\alpha) = \arg\left[1 - \frac{K(\alpha)}{R(\alpha) + iI(\alpha)}\right], \quad -\pi \le A(\alpha) \le 0$	(4.21), (4.22), (4.27)
$P(\alpha) = \exp\left\{-\frac{1}{\pi} P\int_a^b \frac{A(\beta)}{\beta - \alpha} d\beta\right\}$	(5.1)
$Q(\alpha) = \exp\left\{-\frac{1}{\pi} \int_a^b \frac{A(\beta)}{\beta + \alpha} d\beta\right\}$	(5.2)

Using (6.7) and (6.10), we can write the minimum free energy in the form (cf. (3.1))

$$\psi_m(t) = \phi(t) + \frac{1}{2}\int_0^\infty \int_0^\infty E_r^t(s) G_{12}(s,u) E_r^t(u) ds du, \tag{6.12}$$

where

$$G_{12}(s,u) = \frac{1}{2\pi^2}\int_a^b \int_a^b \frac{\Delta_h(\alpha) e^{-\alpha s} \Delta_h(\beta) e^{-\beta u}}{\alpha + \beta} d\alpha d\beta, \tag{6.13}$$

and we understand the subscripts to mean differentiation with respect to the first and second variable. It follows that

$$G(s,u) = G(\infty,\infty) + \frac{1}{2\pi^2}\int_a^b \int_a^b \frac{\Delta_h(\alpha) e^{-\alpha s} \Delta_h(\beta) e^{-\beta u}}{(\alpha + \beta)\alpha\beta} d\alpha d\beta \tag{6.14}$$

if we require that [1]

$$G(\infty,\infty) = G(s,\infty) = G(\infty,s), \quad s \in \mathcal{R}^+, \tag{6.15}$$

yielding $G_1(s,\infty) = G_2(\infty,s) = 0$. It is also required that [1]

$$G(s,0) = G(0,s) = G(s), \quad s \in \mathcal{R}^+, \tag{6.16}$$

where $G(s)$ is defined by (2.2). We deduce from (6.15) and (6.16) that

$$G(\infty,\infty) = G(\infty) = G_\infty \tag{6.17}$$

in the notation of (2.1). To show that (6.16) holds, observe that for $z \in \Omega^-$,

$$\frac{1}{2\pi i}\int_{-\infty}^\infty \frac{H_-(\omega')}{(\omega' - z)\omega'} d\omega' = -\frac{H_-(z)}{z} = \frac{i}{2\pi}\int_a^b \frac{\Delta_h(\beta)}{(\beta - iz)\beta} d\beta, \tag{6.18}$$

where the first relation follows by closing the contour in $\Omega^{(-)}$, and the second results in the same manner as (6.10). It follows from (5.2) and $H_-(i\alpha) = H_+(-i\alpha)$ that

$$\frac{1}{2\pi}\int_a^b \frac{\Delta_h(\beta)}{(\beta+\alpha)\beta}d\beta = \frac{1}{h_\infty}K(\alpha)Q(\alpha). \tag{6.19}$$

Noting that

$$G(s) = G_\infty - \int_a^b \frac{k(\alpha)}{\alpha}e^{-\alpha s}d\alpha, \tag{6.20}$$

we deduce from $(6.11)_2$ that (6.16) holds. Observe that both G and G_{12} are positive quantities.

The isothermal energy balance equation can be written as

$$\dot{\psi}_m(t) + D_m(t) = T(t)\dot{E}(t), \tag{6.21}$$

where D_m is the rate of dissipation associated with the minimum free energy. This quantity must be nonnegative by the second law. It is given by [1]

$$\begin{aligned} D_m(t) &= \left\{\frac{1}{2\pi}\int_{-\infty}^{\infty} H_-(\omega)E_{r+}^t(\omega)d\omega\right\}^2 \\ &= \left\{\frac{1}{2\pi}\int_a^b \Delta_h(\alpha)E_{r+}^t(-i\alpha)d\alpha\right\}^2 \\ &= \left\{\frac{1}{2\pi}\int_0^\infty\int_a^b \Delta_h(\alpha)E_r^t(u)e^{-\alpha u}d\alpha du\right\}^2 \geq 0. \end{aligned} \tag{6.22}$$

Integrating (6.21), we obtain the relation

$$\psi_m(t) + \mathcal{D}_m(t) = W(t), \tag{6.23}$$

where $\mathcal{D}_m$ is the total dissipation, defined by

$$\mathcal{D}_m(t) = \int_{-\infty}^t D(s)ds. \tag{6.24}$$

This quantity is assumed to be finite.

It is through the rate of dissipation and the total dissipation that we make the most direct connection with measurable physical quantities. In particular, $\mathcal{D}_m(t)$ is the least upper bound on the total dissipation, under isothermal conditions, which actually occurs in the material. This is clear from (6.23), since $\psi_m(t)$ is the greatest lower bound on the physical free energy.

It follows from (6.8), (6.23) that

$$\mathcal{D}_m(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left|p_+^t(\omega)\right|^2 d\omega. \tag{6.25}$$

This is not, however, particularly convenient for deriving a useful formula for $\mathcal{D}_m$. Instead, we take another, more direct approach. From (3.1), $(3.3)_2$, (4.1), and (6.12),

we see that

$$\begin{aligned}\mathcal{D}_m(t) = W(t) - \psi_m(t) &= -G'(0)\int_0^\infty [E_r^t(s)]^2 ds \\ &\quad - \frac{1}{2}\int_0^\infty\int_0^\infty E_r^t(s)L(s,u)E_r^t(u)dsdu,\\ L(s,u) &= -\int_a^b \alpha k(\alpha)e^{-\alpha|s-u|}d\alpha \\ &\quad + \frac{1}{2\pi^2}\int_a^b\int_a^b \frac{\Delta_h(\alpha)e^{-\alpha s}\Delta_h(\beta)e^{-\beta u}}{\alpha+\beta}d\alpha d\beta.\end{aligned} \tag{6.26}$$

A point of interest is whether we can take the limit $a \to 0$ in the above formulae, which would extend the class of relaxation functions beyond those with exponential decay at large times. In light of (4.11), we see that if $k(0) = 0$, then there is no singularity at the real axis, and it should always be possible to do so. In all other cases considered in section 4, there will be an integrable singularity. However, the form $(6.11)_2$ is free of these singularities (see Table 1), and the integrals in (6.13), (6.22), and (6.26) exist, so the formulae may be accepted as valid in the limit $a \to 0$.

7. An alternative approach. Another approach to finding the minimum free energy of a continuous spectrum material is outlined in this section. Its most remarkable feature is that it does not require explicit factorization of the function H. It was motivated initially by the method of Breuer and Onat [13] who propose an ansatz for the optimal continuation in the discrete spectrum case and solve the problem by this means. A similar ansatz can be written down without difficulty for the continuous spectrum case. However, it turns out that no such explicit assumption is required.

We start from the form (6.1) of the Wiener–Hopf equation, absorbing the factor $2i$ in R_+ and seeking not $E_m^t(\omega)$ but

$$\Xi_m^t(\omega) = i\omega\overline{F}(\omega)E_m^t(\omega), \tag{7.1}$$

which is also analytic in Ω^+. The reason for this change of unknown is so that we end up with formulae that are directly comparable with earlier results, in particular (6.6), based on the factorization of H with factors F and $\overline{F}$ extracted, as in (2.24), (2.25), and later formulae. The quantity Ξ_m^t is related to the memory-dependent part of the Fourier transform of the stress associated with the optimal continuation E_m^t and a zero history before time t.

Thus, recalling (2.15), we consider the relation

$$H(\omega)\left[E_{r+}^t(\omega) + \frac{\Xi_m^t(\omega)}{i\omega\overline{F}(\omega)}\right] = H(\omega)E_{r+}^t(\omega) + \overline{Y}(\omega)\Xi_m^t(\omega) = R_+(\omega), \tag{7.2}$$

where Y is defined by (4.20). We consider the discontinuity of both sides across the cut $(-ia, -ib)$. The quantities E_{r+}^t and R_+ are analytic in Ω^- and therefore have no discontinuity across the cut. Using (5.4) and its complex conjugate which gives $H_{-R}(-i\alpha)$, we obtain

$$\begin{aligned}\overline{Y}_L(-i\alpha)\Xi_L^t(-i\alpha) - \overline{Y}_R(-i\alpha)\Xi_R^t(-i\alpha) &= i\alpha I(\alpha)E_{r+}^t(-i\alpha), \quad \alpha \in (a,b),\\ &= 0, \quad \alpha \notin (a,b),\end{aligned} \tag{7.3}$$

where Ξ_L^t, Ξ_R^t are the limits of Ξ_m^t on $[-ia, -ib]$ from the left and right, respectively. If it were assumed that Ξ_m^t could be written as a Cauchy integral over $[-ia, -ib]$, which

amounts to the continuous version of the Breuer–Onat ansatz, then (7.3) could be put in the form of a singular integral equation. As remarked earlier, this is unnecessary. The only and very natural assumption needed is that the only singularity of E_m^t is a branch cut on $[-ia, -ib]$. Note that

$$\Xi_m^t(\omega) \approx \frac{1}{\omega} \tag{7.4}$$

for large frequencies, which follows from (2.9) and (7.1). Relation (7.3) is a Hilbert problem, which we can write in the form

$$\begin{aligned} &\Xi^{t+}(\alpha) = C_1(\alpha)\Xi^{t-}(\alpha) + C_2(\alpha), \\ &\Xi^{t+}(\alpha) = \Xi_L^t(-i\alpha), \qquad \Xi^{t-}(\alpha) = \Xi_R^t(-i\alpha), \\ &C_1(\alpha) = \frac{\overline{Y}_R(-i\alpha)}{\overline{Y}_L(-i\alpha)}, \qquad C_2(\alpha) = \frac{i\alpha I(\alpha) E_{r+}(-i\alpha)}{\overline{Y}_L(-i\alpha)}. \end{aligned} \tag{7.5}$$

Note that, from the complex conjugate of (4.21),

$$C_1(a) = C_1(b) = 1. \tag{7.6}$$

This is clear for singular end points as given by (4.14) and (4.17). For the nonsingular case, $I(a)$ and $I(b)$ vanish.

Equation (7.5) will now be solved for $\Xi^t(z) = \Xi_m^t(-iz)$, which has a branch cut on $[a, b]$ and where $\Xi^{t\pm}(\alpha)$ are the limits of this function from the left and the right of the cut. The solution is subject to (7.4) and to the condition that it is bounded except possibly at a or b, where it may diverge logarithmically or as a power less than unity. This latter property reflects the assumptions made relating to the density function k. The general solution is (see [21, p. 237])

$$\begin{aligned} \Xi^t(z) &= \frac{X(z)}{2\pi i} \int_a^b \frac{C_2(\beta)}{X^+(\beta)(\beta - z)} d\beta + X(z)P(z), \\ X(z) &= \Pi(z) e^{N(iz)}, \\ N(iz) &= \frac{1}{2\pi i} \int_a^b \frac{\log C_1(\lambda)}{\lambda - z} d\lambda, \\ \Pi(z) &= (z-a)^{\lambda_1}(z-b)^{\lambda_2}, \end{aligned} \tag{7.7}$$

where λ_1, λ_2 are integers and $P(z)$ is an arbitrary polynomial of degree not less than $\kappa - 1$ with

$$\kappa = -\lambda_1 - \lambda_2.$$

Observe that $N(iz)$ is the quantity defined by (4.19) and (4.20) since

$$\overline{Y}_R(-i\alpha) = Y_R(i\alpha), \qquad \overline{Y}_L(-i\alpha) = Y_L(i\alpha), \tag{7.8}$$

by virtue of (4.6) and (4.8). The quantity $X^+(\beta)$ is the limit of $X(z)$ as $z \to \beta \in (a, b)$ from the positive half-plane. Near $z = a, b$ the quantity N is finite because of (7.6), so that

$$\begin{aligned} X(z)) \underset{z \to a}{\approx} K_1(z-a)^{\lambda_1} \\ \underset{z \to b}{\approx} K_2(z-b)^{\lambda_2}, \end{aligned} \tag{7.9}$$

where K_1, K_2 are constants. To ensure no divergence in Ξ of order unity or stronger, we must have $\lambda_1, \lambda_2 \geq 0$ and $\kappa \leq 0$. For $\kappa < 0$, solutions vanishing at infinity are possible only if restrictions are placed on C_2, which depends only on given physical parameters [21]. Thus, we must have $\kappa = 0$ and $\lambda_1 = \lambda_2 = 0$. The polynomial P is zero. Therefore

$$X(z) = e^{N(iz)}$$

and

$$\Xi^t(i\omega) = \Xi^t_m(\omega) = \frac{X(i\omega)}{2\pi i} \int_a^b \frac{C_2(\beta)}{X^+(\beta)(\beta - i\omega)} d\beta. \tag{7.10}$$

Observe that, from (4.24),

$$X(i\omega) = \frac{i\omega \overline{F}(\omega)}{h_\infty H_-(\omega)} \tag{7.11}$$

and

$$X^+(\beta) = \frac{\beta \overline{F}_L(-i\beta)}{h_\infty H_{-L}(-i\beta)} = \frac{1}{P(\beta)} e^{iA(\beta)}, \qquad \beta \in (a, b), \tag{7.12}$$

where (4.8) and (5.1) have been used. Now, from (4.21) and (7.8),

$$\overline{Y}_L(-i\beta) = \frac{1}{2}\left[1 - \frac{K(\beta)}{R(\beta) - iI(\beta)}\right] = \frac{1}{2}\sqrt{\frac{(R(\beta) - K(\beta))^2 + I^2(\beta)}{R^2(\beta) + I^2(\beta)}} e^{-iA(\beta)}, \tag{7.13}$$

by virtue of (5.5). Thus

$$\frac{C_2(\beta)}{X^+(\beta)} = 2i\beta P(\beta) I(\beta) \sqrt{\frac{R^2(\beta) + I^2(\beta)}{(R(\beta) - K(\beta))^2 + I^2(\beta)}} E^t_{r+}(-i\beta) = -ih_\infty \Delta_h(\beta) E^t_{r+}(-i\beta) \tag{7.14}$$

in the notation of (6.11). Then, we finally obtain from (7.1), (7.10), and (7.14)

$$E^t_m(\omega) = -\frac{1}{2\pi H_-(\omega)} \int_a^b \frac{\Delta_h(\beta) E^t_{r+}(-i\beta)}{\beta - i\omega} d\beta,$$

which agrees with (6.6) and (6.10).

Note that the quantity X, given by (7.11), is closely related to the factor H_-. This is how the factors of H enter the formulae. The quantity X is the solution of the homogeneous part of the Hilbert problem (7.5). We note that the factorization problem of H can be expressed as a homogeneous Hilbert problem on the real axis:

$$H_-(w) = H(\omega) \left[H_+(w)\right]^{-1}.$$

It is straightforward to show that this is equivalent to the homogeneous problem associated with $(7.5)_1$ by taking the limit of this relation on both sides of the branch cut on $[-ia, -ib]$ in H_- and H, and using (7.11).

8. Minimal states. Finally, let us explore the concept of minimal states, defined by (2.4), in the context of continuous spectrum materials.

PROPOSITION 8.1. *For the relaxation function derivative given by* (4.1)*, where* k *is negative on* (a,b)*, except possibly at a finite number of isolated points, and for histories with* E^t_+ *analytic on* $\mathcal{R}$ *(see the remark after* (2.26)*) the minimal states are singletons. In other words,* $(E(t),E^t)$ *is the minimal state.*

Proof. We define $(E_d(t),E^t_d)$ as

$$\begin{aligned} E_d(t) &= E_1(t)-E_2(t),\\ E^t_d(s) &= E^t_1(s)-E^t_2(s), \quad s\in\mathcal{R}^+. \end{aligned} \tag{8.1}$$

Then (2.4) becomes

$$\begin{aligned} &E_d(t)=0,\\ &\int_0^\infty G'(s+\tau)E^t_d(s)ds=\int_a^b k(\alpha)e^{-\alpha\tau}E^t_{d+}(-i\alpha)d\alpha=0 \quad \forall\,\tau\ge 0. \end{aligned} \tag{8.2}$$

The function

$$Z(\tau)=\int_a^b k(\alpha)e^{-\alpha\tau}E^t_{d+}(-i\alpha)d\alpha \tag{8.3}$$

can be extended to the complex τ plane. It is analytic (and therefore zero) for $Re\ \tau>0$. Taking the inverse Laplace transform, we deduce that $k(\alpha)E^t_+(-i\alpha)$ vanishes for $\alpha\in\mathcal{R}^+$. Thus, since $k(\alpha)$ does not vanish for $\alpha\in(a,b)$, except at most at a finite number of isolated points, we have

$$E^t_{d+}(-i\alpha)=0, \tag{8.4}$$

over (a,b) or some open subinterval of this region, which in turn implies that $E^t_{d+}(\omega)$ vanishes in the region of analyticity connected to $(-ia,-ib)$. This certainly includes Ω^- and in particular the real axis. We conclude that

$$E_d(t)=0, \qquad E^t_d(s)=0, \quad s\in\mathcal{R}^{++}. \qquad \square \tag{8.5}$$

This result is in sharp contrast with the situation prevailing for discrete spectrum materials [10, 26].

It follows from Proposition 8.1 that the work function is a function of state and is the maximum free energy, for relaxation functions obeying a strong dissipativity condition [8].

A generalization of Proposition 8.1 is given in [26, 11]. Also, it follows from a more general result proved in [9, Proposition 7.3].

9. Particular cases and approximations. Explicit expressions for $F(\omega)$, $R(\alpha)$, $K(\alpha)$, and $G'(0)=-h^2_\infty$ corresponding to a number of choices of $k(\alpha)$ are presented in Table 2. These are the quantities required to determine $H_\pm$ in (4.23) and (4.24) or indeed the free energy functional (6.12). A multiplying positive constant may of course be included in k in all cases. In addition, we note the following formulae which allow simple generalizations of those tabulated. If $k(\alpha)$ yields $F(\omega)$, then $\alpha k(\alpha)$ yields $F_1(\omega)$, where

$$\begin{aligned} F_1(\omega) &= G'(0)-i\omega F(\omega),\\ G'(0) &= \int_a^b k(\alpha)d\alpha. \end{aligned} \tag{9.1}$$

TABLE 2

The quantities F, K, R, and $G'(0)$ required to determine the factors of H and the minimum free energy for various choices of the density function k. The function $I(\alpha) = -\pi k(\alpha)$. The quantity $c = (a+b)/2$. The function Ei *is the exponential-integral function. The fourth and fifth rows are, of course, special cases ($\theta = 1/2$) of the sixth and seventh rows. The branch of $[(a-z)(b-z)]^{1/2}$ is chosen to be the one that approaches $-z$ as $|z|$ becomes large; similarly for $(a-z)^{\theta}(b-z)^{1-\theta}$. The quantity $\theta \in (0,1)$.*

	$k(\alpha),\ \alpha \in (a,b)$	$F(\omega)$	$K(\alpha),\ \alpha > -a$	$R(\alpha),\ \alpha \in (a,b)$	$G'(0)$
(1)	-1	$\log\left(\dfrac{\omega - ia}{\omega - ib}\right)$	$\log\left(\dfrac{\alpha + a}{\alpha + b}\right)$	$\log\left(\dfrac{\alpha - a}{b - \alpha}\right)$	$-(b-a)$
(2)	$-(\alpha - a)(b - \alpha)$	$-(b-a)(c+i\omega)$ $+ (a+i\omega)(b+i\omega)\log\left(\dfrac{b+i\omega}{a+i\omega}\right)$	$-(b-a)(c+\alpha)$ $+ (a+\alpha)(b+\alpha)\log\left(\dfrac{b+\alpha}{a+\alpha}\right)$	$-(b-a)(c-\alpha)$ $+ (\alpha-a)(b-\alpha)\log\left(\dfrac{\alpha-a}{b-\alpha}\right)$	$-\dfrac{(b-a)^3}{6}$
(3)	$-e^{-\alpha t_0},\quad \alpha \geq a > 0$	$e^{i\omega t_0}\,\mathrm{Ei}(-(a+i\omega)t_0)$	$e^{\alpha t_0}\,\mathrm{Ei}(-(a+\alpha)t_0)$	$e^{-\alpha t_0}\,\mathrm{Ei}((\alpha-a)t_0)$	$-\dfrac{e^{-at_0}}{t_0}$
(4)	$-\dfrac{1}{\sqrt{(\alpha-a)(b-\alpha)}}$	$-\dfrac{\pi}{[(a+i\omega)(b+i\omega)]^{1/2}}$	$-\dfrac{\pi}{\sqrt{(a+\alpha)(b+\alpha)}}$	0	$-\pi$
(5)	$-\sqrt{(\alpha-a)(b-\alpha)}$	$-\pi\left\{c+i\omega - [(a+i\omega)(b+i\omega)]^{1/2}\right\}$	$-\pi\left\{c+\alpha - \sqrt{(a+\alpha)(b+\alpha)}\right\}$	$-\pi(c-\alpha)$	$-\dfrac{\pi(b-a)^2}{8}$
(6)	$-\dfrac{1}{(\alpha-a)^{1-\theta}(b-\alpha)^{\theta}}$	$-\dfrac{\pi}{\sin\pi\theta}\dfrac{1}{(a+i\omega)^{1-\theta}(b+i\omega)^{\theta}}$	$-\dfrac{\pi}{\sin\pi\theta}\dfrac{1}{(a+\alpha)^{1-\theta}(b+\alpha)^{\theta}}$	$\dfrac{\pi\cot\pi\theta}{(\alpha-a)^{1-\theta}(b-\alpha)^{\theta}}$	$-\dfrac{\pi}{\sin\pi\theta}$
(7)	$-(\alpha-a)^{1-\theta}(b-\alpha)^{\theta}$	$-\dfrac{\pi}{\sin\pi\theta}\{a(1-\theta)+b\theta+i\omega$ $-(a+i\omega)^{1-\theta}(b+i\omega)^{\theta}\}$	$-\dfrac{\pi}{\sin\pi\theta}\{a(1-\theta)+b\theta+\alpha$ $-(a+\alpha)^{1-\theta}(b+\alpha)^{\theta}\}$	$-\dfrac{\pi}{\sin\pi\theta}\{a(1-\theta)+b\theta-\alpha$ $+\cos\pi\theta(\alpha-a)^{1-\theta}(b-\alpha)^{\theta}\}$	$-\dfrac{\pi(b-a)^2\theta(1-\theta)}{2\sin\pi\theta}$

Also

$$
\begin{aligned}
R_1(\alpha) &= G'(0) + \alpha R(\alpha), \\
K_1(\alpha) &= G'(0) - \alpha K(\alpha), \\
G_1'(0) &= \int_a^b \alpha k(\alpha) d\alpha,
\end{aligned}
\tag{9.2}
$$

where the subscript "1" indicates the various quantities corresponding to $\alpha k(\alpha)$. Furthermore, all quantities are linear in k, so that formulae for linear combinations of density functions may be constructed without difficulty.

A choice of k which is of some physical interest is (see [27])

$$
k(\alpha) = -A\alpha^{\lambda}, \quad 0 < \lambda \leq 1, \quad 0 < \alpha < \infty, \quad A > 0,
\tag{9.3}
$$

particularly for $\lambda = 0.5$. However, $G'(0)$ is infinite for this relaxation spectrum, a problem that is easily remedied in principle by taking the range of integrations to be finite. In this case, it would probably be simpler to evaluate all the quantities of interest by numerical methods. Note, however, that the seventh and fifth rows of Table 2 provide a good approximation to (9.3) if a is set equal to zero and b is taken to be large. In any case, the relevance of (9.3) is weakened by the fact that the power λ may depend on the value of α.

Note that the behavior of $H_{\pm}$, given by (4.23) and (4.24), for large ω, is approximated by

$$
\begin{aligned}
H_+(\omega) &\approx -\frac{i\omega}{h_\infty} F(\omega), \\
H_-(\omega) &\approx \frac{i\omega}{h_\infty} \overline{F}(\omega).
\end{aligned}
\tag{9.4}
$$

At small ω, they are approximately given by a real constant times the quantities on the right-hand side of (9.4).

This suggests that (9.4), perhaps with multiplying constants, may provide a reasonable approximation for the factors at all frequencies. However, the functional (6.7) has the properties of a free energy (see [1]) only approximately, if (9.4) is used.

Acknowledgments. Our thanks to M. Fabrizio and G. Gentili for useful conversations and to M. Fabrizio for several opportunities afforded to J. M. Golden to visit the University of Bologna, where this work was initiated.

REFERENCES

[1] J. M. Golden, *Free energies in the frequency domain: The scalar case*, Quart. Appl. Math., 58 (2000), pp. 127–150.

[2] L. Deseri, G. Gentili, and J. M. Golden, *An explicit formula for the minimum free energy in linear viscoelasticity*, J. Elasticity, 54 (1999), pp. 141–185.

[3] M. Fabrizio, G. Gentili, and J. M. Golden, *The minimum free energy for a class of compressible viscoelastic fluids*, Adv. Differential Equations, 7 (2002), pp. 319–342.

[4] M. Fabrizio and J. M. Golden, *Minimum free energies for materials with finite memory*, J. Elasticity, 72 (2003), pp. 121–143.

[5] L. Deseri, G. Gentili, and J. M. Golden, *Free energies and Saint-Venant's principle in linear viscoelasticity*, J. Elasticity, submitted.

[6] W. Noll, *A new mathematical theory of simple materials*, Arch. Ration. Mech. Anal., 48 (1972), pp. 1–50.

[7] C. Banfi, *Su una nuova impostazione per l'analisi dei sistemi ereditari*, Ann. Univ. Ferrara Sez. VII (N.S.), 23 (1977), pp. 29–38.
[8] G. Del Piero and L. Deseri, *On the analytic expression of the free energy in linear viscoelasticity*, J. Elasticity, 43 (1996), pp. 247–278.
[9] G. Del Piero and L. Deseri, *On the concepts of state and free energy in linear viscoelasticity*, Arch. Ration. Mech. Anal., 138 (1997), pp. 1–35.
[10] M. Fabrizio and J. M. Golden, *Maximum and minimum free energies for a linear viscoelastic material*, Quart. Appl. Math., 60 (2002), pp. 341–381.
[11] L. Deseri, M. Fabrizio, and J. M. Golden, *On the concept of a minimal state in viscoelasticity: New free energies and applications to PDEs*, Arch. Ration. Mech. Anal., 181 (2006), pp. 43–96.
[12] M. Fabrizio and J. M. Golden, *Non-isothermal free energies for linear materials with memory*, Math. Comput. Modelling, 39 (2004), pp. 219–253.
[13] S. Breuer and E. T. Onat, *On recoverable work in linear viscoelasticity*, Z. Angew. Math. Phys., 15 (1964), pp. 13–21.
[14] D. Graffi and M. Fabrizio, *Sulla nozione di stato materiali viscoelastici di tipo 'rate,'* Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8), 83 (1990), pp. 201–208.
[15] M. Fabrizio and A. Morro, *Mathematical Problems in Linear Viscoelasticity*, SIAM Stud. Appl. Math. 12, SIAM, Philadelphia, 1992.
[16] J. M. Golden and G. A. C. Graham, *Boundary Value Problems in Linear Viscoelasticity*, Springer, Berlin, 1988.
[17] M. E. Gurtin and I. Herrera, *On dissipation inequalities and linear viscoelasticity*, Quart. Appl. Math., 23 (1988), pp. 235–245.
[18] W. A. Day, *Thermodynamics based on a work axiom*, Arch. Ration. Mech. Anal., 31 (1968), pp. 1–34.
[19] M. Fabrizio, C. Giorgi, and A. Morro, *Relaxation property and free energy in materials with fading memory*, J. Elasticity, 40 (1995), pp. 107–122.
[20] M. Fabrizio, C. Giorgi, and A. Morro, *Free energies and dissipation properties for systems with memory*, Arch. Ration. Mech. Anal., 125 (1994), pp. 341–373.
[21] N. I. Muskhelishvili, *Singular Integral Equations*, Noordhoff, Groningen, The Netherlands, 1953.
[22] D. Graffi, *Sull'expressione analitica di alcune grandezze termodinamiche nei materiali con memoria*, Rend. Sem. Mat. Univ. Padova, 68 (1982), pp. 17–29.
[23] D. Graffi, *Ancora sull'expressione analitica dell'energia libera nei materiali con memoria*, Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur., 120 (1986), pp. 111–124.
[24] B. D. Coleman and D. R. Owen, *A mathematical foundation for thermodynamics*, Arch. Ration. Mech. Anal., 54 (1974), pp. 1–104.
[25] B. D. Coleman and D. R. Owen, *On thermodynamics and elastic-plastic materials*, Arch. Ration. Mech. Anal., 59 (1975), pp. 25–51.
[26] J. M. Golden, *A proposal concerning the physical rate of dissipation in materials with memory*, Quart. Appl. Math., 63 (2005), pp. 117–155.
[27] J. D. Ferry, *Viscoelastic Properties of Polymers*, John Wiley, New York, 1980.

SIAM J. APPL. MATH.
Vol. 67, No. 3, pp. 893–913

D-BAR METHOD FOR ELECTRICAL IMPEDANCE TOMOGRAPHY WITH DISCONTINUOUS CONDUCTIVITIES*

KIM KNUDSEN†, MATTI LASSAS‡, JENNIFER L. MUELLER§, AND SAMULI SILTANEN¶

Abstract. The effects of truncating the (approximate) scattering transform in the D-bar reconstruction method for two-dimensional electrical impedance tomography are studied. The method is based on the uniqueness proof of Nachman [*Ann. of Math.* (2), 143 (1996), pp. 71–96] that applies to twice differentiable conductivities. However, the reconstruction algorithm has been successfully applied to experimental data, which can be characterized as piecewise smooth conductivities. The truncation is shown to stabilize the method against measurement noise and to have a smoothing effect on the reconstructed conductivity. Thus the truncation can be interpreted as regularization of the D-bar method. Numerical reconstructions are presented demonstrating that features of discontinuous high contrast conductivities can be recovered using the D-bar method. Further, a new connection between Calderón's linearization method and the D-bar method is established, and the two methods are compared numerically and analytically.

Key words. inverse conductivity problem, electrical impedance tomography, exponentially growing solution, Faddeev's Green's function

AMS subject classifications. 35R30, 65N21, 92C55

DOI. 10.1137/060656930

1. Introduction.

The two-dimensional (2-D) inverse conductivity problem is to determine and reconstruct an unknown conductivity distribution γ in an open, bounded, and smooth domain $\Omega \subset \mathbb{R}^2$ from voltage-to-current measurements on the boundary $\partial\Omega$. We assume that there is a $C > 0$ such that

$$C^{-1} < \gamma(x) < C, \quad x \in \Omega. \tag{1}$$

The boundary measurements are modeled by the Dirichlet-to-Neumann (DN) map

$$\Lambda_\gamma f = \gamma \frac{\partial u}{\partial \nu}\Big|_{\partial\Omega},$$

where u is the solution to the generalized Laplace equation

$$\nabla \cdot \gamma \nabla u = 0 \quad \text{in } \Omega, \quad u|_{\partial\Omega} = f. \tag{2}$$

Mathematically, the problem is to show that the map $\gamma \mapsto \Lambda_\gamma$ is injective and find an algorithm for the inversion of the map. Physically, u is the electric potential in Ω, and Λ_γ represents knowledge of the current flux through $\partial\Omega$ resulting from the voltage distribution f applied on $\partial\Omega$.

*Received by the editors April 10, 2006; accepted for publication (in revised form) December 28, 2006; published electronically April 13, 2007.
http://www.siam.org/journals/siap/67-3/65693.html

†Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, DK-9220 Aalborg Ø, Denmark (kim@math.aau.dk). This author was supported by the Carlsberg Foundation.
‡Institute of Mathematics, Helsinki University of Technology, P.O. Box 1100, 02015 TKK, Finland (matti.lassas@hut.fi). This author was supported by the Academy of Finland.
§Department of Mathematics, Colorado State University, Fort Collins, CO 80523 (mueller@math.colostate.edu). This author was supported by the National Science Foundation under grant 0513509.
¶Institute of Mathematics, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland (samuli.siltanen@tut.fi). This author was supported by the Academy of Finland.

The inverse conductivity problem has applications in subsurface flow monitoring and remediation [30, 31], underground contaminant detection [10, 17], geophysics [9, 27], nondestructive evaluation [11, 36, 39, 37], and a medical imaging technique known as electrical impedance tomography (EIT) (see [8, 5] for a review article on EIT). Conductivity distributions appearing in applications are typically piecewise continuous. This is the case, for example, in medical EIT, since various tissues in the body have different conductivities, and there are discontinuities at organ boundaries. Here we consider 2-D reconstructions. These can be used to image cross-sections of a three-dimensional (3-D) region, such as a patient's torso. In the case of patients receiving mechanical ventilation, for example, 2-D cross-sections are useful for obtaining regional ventilation information in the lungs, which is valuable for setting and controlling the airflow and pressure settings on the ventilator [34, 1]. Real-time imaging of cross-sectional lung activity can also be used for diagnostic purposes, such as detecting a lung collapse, a pulmonary embolism, pulmonary edema, or a pneumothorax.

Let us briefly outline the history of D-bar solution methods for EIT. Recently, Astala and Päivärinta [2] showed that knowledge of the DN map uniquely determines the conductivity $\gamma(x) \in L^\infty(\Omega)$, $0 < c \le \gamma$. This result has been generalized to anisotropic conductivities in [3]. In this work, we will refer to the 2-D uniqueness result by Nachman [25] for $\gamma \in W^{2,p}(\Omega)$, $p > 1$, and by Brown and Uhlmann [6] for $\gamma \in W^{1,p}(\Omega)$, $p > 2$. The proof in [25] is constructive; that is, it outlines a direct method for reconstructing the conductivity γ from knowledge of Λ_γ. This method was realized as a numerical algorithm for C^2 conductivities in [29, 24, 15]. The uniqueness result of Brown and Uhlmann in [6] was formulated as a reconstruction algorithm in [20], which has been implemented in [18, 19]. There are many similarities between the two methods. In fact, it was shown in [18] using the Brown–Uhlmann approach that the reconstruction method of Nachman [25] can be extended to the class of conductivities $\gamma \in W^{1+\epsilon,p}(\Omega)$, $p > 2$, $\epsilon > 0$. We refer the reader to [24, 5, 38] for discussions of uniqueness results for γ in other spaces and $\Omega \subset \mathbb{R}^n$, $n \ge 2$.

Nachman's D-bar approach in [25] is based on the evaluation of the scattering transform $\mathbf{t}(k)$ by the formula

$$\mathbf{t}(k) = \int_{\partial\Omega} e^{i\overline{k}\bar{x}}(\Lambda_\gamma - \Lambda_1)\psi(\cdot,k)d\sigma(x), \quad k \in \mathbb{C},\ x = x_1 + ix_2, \tag{3}$$

where Λ_1 denotes the DN map corresponding to the homogeneous conductivity 1. Then γ can be recovered by solving a D-bar equation containing $\mathbf{t}(k)$. The functions $\psi(\cdot,k)$ in (3) are traces of certain exponentially growing solutions to (2), i.e., solutions that behave like e^{ikx} asymptotically as either $|x|$ or $|k|$ tends to infinity. These traces can, in principle, be found by solving a particular boundary integral equation. However, as solving such an equation is quite sensitive to measurement noise, the following approximation to $\mathbf{t}(k)$ was introduced in [29]:

$$\mathbf{t}^{\exp}(k) = \int_{\partial\Omega} e^{i\overline{k}\bar{x}}(\Lambda_\gamma - \Lambda_1)e^{ikx}d\sigma(x). \tag{4}$$

This approximation can be viewed as a linearizing assumption, since the approximation $\psi|_{\partial\Omega} \approx e^{ikx}$ is used instead of solving for ψ on the boundary using the nonlinear equation (7).

Formula (4) allows the evaluation of $\mathbf{t}^{\exp}(k)$ for L^∞ conductivities, and the D-bar method is found to be effective even when the conductivity does not satisfy the assumptions of the original reconstruction theorem. In [15], quite accurate reconstructions are computed from experimental data collected on a phantom chest consisting

of agar heart and lungs in a saline-filled tank. They are the first reconstructions using the D-bar method on a discontinuous conductivity and on measured data. In [16], the D-bar algorithm with a differencing $\mathbf{t}^{\exp}$ approximation is used to reconstruct conductivity changes in a human chest, particularly pulmonary perfusion.

Our aim is to better understand the reconstruction of realistic conductivities from noisy EIT data using the D-bar method by studying its application to piecewise smooth conductivities. Section 2 gives necessary background on the method and its variants. In section 3, we prove that reconstructions from any truncated scattering data are smooth. In section 4, we show that the reconstructions from noisy data using truncated $\mathbf{t}^{\exp}$ are stable. We remark that previous work [23, 4] shows that the exact reconstruction algorithm is stable in a restricted sense, i.e., as a map defined on the range of the forward operator $\Lambda\colon \gamma \mapsto \Lambda_\gamma$. In contrast, we show that the approximate reconstruction is continuously defined on the entire data space $\mathcal{L}(H^{1/2}(\partial\Omega), H^{-1/2}(\partial\Omega))$. As an application of the stability, we consider in section 5 mollified versions γ_λ of a piecewise continuous conductivity distribution γ and show that reconstructions of γ_λ converge to reconstructions of γ as $\lambda \to 0$. This means that no systematic artifacts are introduced when the reconstruction method is applied to conductivities outside the assumptions of the theory.

In section 6, a connection between the linearization method of Calderón [7] and the D-bar method is established. Calderón's method is written in terms of $\mathbf{t}^{\exp}$ and is revealed to be a low-order approximation to the D-bar method. The simple example of the unit disk containing one concentric ring of constant conductivity with a discontinuity at the interface is studied in depth in section 7. We write $\mathbf{t}^{\exp}$ as a series showing the asymptotic growth rate. Reconstructions by Calderón's method and the D-bar method with the $\mathbf{t}^{\exp}$ approximation are expressed in explicit formulas.

In section 8, we illustrate our theoretical findings by numerical examples. We find that both the D-bar method and Calderón's method can approximately recover the location of a discontinuity. Also, both methods yield good reconstructions of low-contrast conductivities but have difficulties in recovering the actual conductivity values in the presence of high contrast features near the boundary.

2. The D-bar reconstruction method. In this section, we briefly review the reconstruction method based on the proof by Nachman [25]. We will describe both the exact mathematical algorithm and an approximate numerical algorithm.

2.1. Exact reconstruction from infinite precision data. The reconstruction method uses exponentially growing solutions to the conductivity equation. Suppose $\gamma - 1 \in W^{1+\epsilon,p}(\mathbb{R}^2)$ with $p > 2$ and $\gamma \equiv 1$ in $\mathbb{R}^2 \setminus \overline{\Omega}$. Then the equation

$$\nabla \cdot \gamma \nabla u = 0 \text{ in } \mathbb{R}^2 \tag{5}$$

has a unique exponentially growing solution u that behaves like e^{ixk}, where x is understood as $x = x_1 + ix_2 \in \mathbb{C}$ and the parameter $k = k_1 + ik_2 \in \mathbb{C}$. The construction of exponentially growing solutions is done by reducing the conductivity equation to either a Schrödinger equation (requires two derivatives on the conductivity) or a first-order system (requires one derivative). Consider $\psi := u\sqrt{\gamma}$ satisfying $(e^{-ixk}\psi(x,k) - 1) \in W^{1,p}(\mathbb{R}^2)$ with $p > 2$, and note that since $\gamma = 1$ at $\partial\Omega$ we have $u|_{\partial\Omega} = \psi|_{\partial\Omega}$. The intermediate object in the reconstruction method is the scattering transform defined in terms of the DN map by (3).

The reconstruction algorithm consists of the two steps

$$\Lambda_\gamma \to \mathbf{t} \to \gamma. \tag{6}$$

In order to compute $\mathbf{t}$ from Λ_γ by (3), one needs to find the trace of $\psi(\cdot,k)$ on $\partial\Omega$. It turns out that $\psi|_{\partial\Omega}$ satisfies

$$\psi(\cdot,k)|_{\partial\Omega} = e^{ikx} - S_k(\Lambda_\gamma - \Lambda_1)\psi(\cdot,k). \tag{7}$$

Here S_k is the single-layer operator

$$(S_k\phi)(x) := \int_{\partial\Omega} G_k(x-y)\phi(y)d\sigma(y), \quad k \in \mathbb{C}\setminus 0, \tag{8}$$

where the Faddeev's Green's function G_k is defined by

$$G_k(x) := \frac{e^{ikx}}{(2\pi)^2}\int_{\mathbb{R}^2} \frac{e^{ix\cdot\xi}}{|\xi|^2 + 2k(\xi_1 + i\xi_2)}d\xi, \quad -\Delta G_k = \delta. \tag{9}$$

Here the dot product is computed with real vectors $x = (x_1, x_2)$ and $\xi = (\xi_1, \xi_2)$. The Fredholm equation (7) is uniquely solvable in $H^{1/2}(\partial\Omega)$; see [25].

To compute γ from $\mathbf{t}$, the key observation is that with respect to the parameter k, the function $\mu(x,k) = e^{-ixk}\psi(x,k)$ satisfies a differential equation where $\mathbf{t}$ enters as a coefficient. More precisely, μ satisfies for fixed $x \in \mathbb{C}$ the D-bar equation

$$\overline{\partial}_k\mu(x,k) = \frac{1}{4\pi\overline{k}}\mathbf{t}(k)e_{-x}(k)\overline{\mu(x,k)}, \quad k \in \mathbb{C}, \tag{10}$$

where the unimodular function e_k is defined by

$$e_x(k) := e^{i(kx+\overline{k}\bar{x})} = e^{-i(-2k_1,2k_2)\cdot x}. \tag{11}$$

It is shown in [25] (see also [6, 20]) that $\mu(x,\cdot)$ is, in fact, the unique solution to (10) defined by the asymptotic condition $\mu(x,\cdot) - 1 \in L^r(\mathbb{R}^2), r > 2/\epsilon$. Moreover, the solution belongs to $C^\alpha(\mathbb{R}^2)$ with $\alpha < 1$; see section 3. Hence $\mu(x,k)$ can be computed from $\mathbf{t}$ by solving (10) or, equivalently, the Fredholm integral equation

$$\mu(x,s) = 1 + \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} \frac{\mathbf{t}(k)}{(s-k)\overline{k}}e_{-x}(k)\overline{\mu(x,k)}dk_1dk_2. \tag{12}$$

Finally, the conductivity can be recovered from μ using the formula

$$\gamma(x) = \mu(x,0)^2, \quad x \in \Omega. \tag{13}$$

2.2. Truncation of scattering transform. Note that in (12) the integral is over the whole plane. We will define a regularized D-bar algorithm by truncating the scattering transform to a disk of radius R. Then

$$\mathbf{t}_{\mathrm{R}}(k) \equiv \begin{cases} \mathbf{t}(k) & \text{for } |k| \le R, \\ 0 & \text{for } |k| > R, \end{cases} \tag{14}$$

and $\mu_{\mathrm{R}}(x,k)$ is the solution of

$$\mu_{\mathrm{R}}(x,s) = 1 + \frac{1}{(2\pi)^2}\int_{|k|\le R} \frac{\mathbf{t}_{\mathrm{R}}(k)}{(s-k)\overline{k}}e_{-x}(k)\overline{\mu_{\mathrm{R}}(x,k)}dk_1dk_2. \tag{15}$$

This defines a modified D-bar algorithm consisting of the following steps:

1. Solve (7) for $\psi|_{\partial\Omega}$.
2. Compute $\mathbf{t}_{\mathrm{R}}$ by (3) and (14).
3. Solve the integral equation (15) for μ_{R}.
4. Compute the reconstruction $\gamma_{\mathrm{R}}(x) = \mu_{\mathrm{R}}(x,0)^2$.

According to [24], this algorithm gives correct results at the asymptotic limit $R \to \infty$.

2.3. Approximate reconstruction from finite precision data. In the presence of noise in the data, solving (7) is difficult, and therefore the approximate scattering transform $\mathbf{t}^{\exp}(k)$ defined by (4) was introduced. An advantage of $\mathbf{t}^{\exp}(k)$ is that the definition applies just as well to discontinuous conductivities as to smooth ones. However, it can be shown in certain cases that $\mathbf{t}^{\exp}(k)$ grows so fast as $|k|$ tends to infinity (see (64) below) that the corresponding D-bar equation is not solvable. In addition, the practical computation is stable only for $|k| \le R$, where the radius R depends on the noise level. Thus the scattering transform needs to be truncated. Set

$$\mathbf{t}_{\mathrm{R}}^{\exp}(k) = \begin{cases} \mathbf{t}^{\exp}(k) & \text{for } |k| \le R, \\ 0 & \text{for } |k| > R, \end{cases} \tag{16}$$

and write the corresponding D-bar equation:

$$\mu_{\mathrm{R}}^{\exp}(x,s) = 1 + \frac{1}{(2\pi)^2} \int_{|k|\le R} \frac{\mathbf{t}_{\mathrm{R}}^{\exp}(k)}{(s-k)\overline{k}} e_{-x}(k) \overline{\mu_{\mathrm{R}}^{\exp}(x,k)} dk_1 dk_2. \tag{17}$$

We arrive at the following reconstruction algorithm:

1. Compute $\mathbf{t}_{\mathrm{R}}^{\exp}$ by (4) and (16).
2. Solve (17) for $\mu_{\mathrm{R}}^{\exp}$.
3. Compute $\gamma_{\mathrm{R}}^{\exp}(x) = \mu_{\mathrm{R}}^{\exp}(x,0)^2$.

We will show in section 4 that this reconstruction algorithm is robust against noise. In the numerical implementation, the challenge is to solve (17); see [29, 21].

3. Smoothness of reconstructions from truncated scattering data. We first investigate the x-smoothness of the solution to the D-bar equation

$$\mu(x,s) = 1 + \frac{1}{\pi} \int_{\mathbb{R}^2} \frac{\phi(k)}{s-k} e_{-x}(k) \overline{\mu(x,k)} dk_1 dk_2 \tag{18}$$

under various assumptions on the coefficient ϕ. Then we will show that the reconstructions γ_{R} and $\gamma_{\mathrm{R}}^{\exp}$ computed from truncated scattering data are smooth functions.

The analysis of (18) makes heavy use of the solid Cauchy transform

$$\mathcal{C}f(s) := \frac{1}{\pi} \int_{\mathbb{R}^2} \frac{f(k)}{s-k} dk_1 dk_2. \tag{19}$$

The following result is essentially [25, Lemma 1.2] and [35, Theorem 1.21].

LEMMA 3.1. *Suppose $f \in L^{p_1}(\mathbb{R}^2)$, where $1 < p_1 < 2$. Then*

$$\|\mathcal{C}f\|_{L^{\tilde{p}_1}(\mathbb{R}^2)} \le C \|f\|_{L^{p_1}(\mathbb{R}^2)}, \quad \frac{1}{\tilde{p}_1} = \frac{1}{p_1} - \frac{1}{2}. \tag{20}$$

Suppose further that $f \in L^{p_1}(\mathbb{R}^2) \cap L^{p_2}(\mathbb{R}^2)$, where $1 < p_1 < 2 < p_2 < \infty$. Then

$$\|\mathcal{C}f\|_{C^\alpha(\mathbb{R}^2)} \le C(\|f\|_{L^{p_1}(\mathbb{R}^2)} + \|f\|_{L^{p_2}(\mathbb{R}^2)}), \quad \alpha = 1 - \frac{2}{p_2}. \tag{21}$$

In the next lemma, we consider the continuity of the Cauchy transform applied to functions depending on a parameter. To simplify notation, we introduce for $x \in \mathbb{R}^2$ the real-linear operator Φ_x by $\Phi_x f = \phi(k) e_{-k}(x) \overline{f(k)}$.

LEMMA 3.2. *Let $\phi \in L^{p_1} \cap L^{p_2}(\mathbb{R}^2)$ with $1 < p_1 < 2 < p_2 < \infty$. Then the map*

$$x \mapsto \mathcal{C}(\phi e_{-x}) \tag{22}$$

is continuous from $\mathbb{R}^2$ *into* $L^{\widetilde{p}_1}(\mathbb{R}^2) \cap C^\alpha(\mathbb{R}^2)$, $\alpha = 1 - 2/p_2$. *Further,* $\mathcal{C}\Phi_x$ *is bounded on* $L^r(\mathbb{R}^2)$, $r > 2$, *and the map* $x \mapsto \mathcal{C}\Phi_x$ *is continuous from* $\mathbb{R}^2$ *into* $\mathcal{L}(L^r(\mathbb{R}^2))$.

Proof. Using the Lebesgue dominated convergence theorem, it is straightforward to see that the map $x \mapsto \phi e_{-x}$ is continuous from $\mathbb{R}^2$ into $L^{p_1}(\mathbb{R}^2) \cap L^{p_2}(\mathbb{R}^2)$. The continuity of the map (22) then follows from the linearity of $\mathcal{C}$ and (20)–(21).

The assumption on ϕ implies by the Hölder inequality that $\phi(k)e_{-k}(x) \in L^2(\mathbb{R}^2)$, and, as before, we can argue that $\phi(k)e_{-k}(x)$ is continuous with respect to $x \in \mathbb{R}^2$. It follows then by the Hölder inequality that $\Phi_x \in \mathcal{L}(L^r(\mathbb{R}^2), L^{2r/(2+r)}(\mathbb{R}^2))$. Moreover, Φ_x is continuous with respect to $x \in \mathbb{R}^2$. The claim for $\mathcal{C}\Phi_x$ then follows from (20). ∎

We are now ready to prove the unique solvability of (18) in the case where ϕ is in certain L^p-spaces and analyze how the solution depends on the parameter x.

LEMMA 3.3. *Let* $\phi \in L^{p_1} \cap L^{p_2}(\mathbb{R}^2)$ *with* $1 < p_1 < 2 < p_2 < \infty$. *Then* (18) *has a unique solution* μ *with* $\mu(x,\cdot) - 1 \in L^r \cap C^\alpha(\mathbb{R}^2)$ *for any* $r \geq \widetilde{p}_1$ *and* $\alpha < 1 - 2/p_2$. *Moreover, the map*

$$x \mapsto \mu(x,\cdot) \tag{23}$$

is continuous from $\mathbb{R}^2$ *into* $L^r \cap C^\alpha(\mathbb{R}^2)$.

Proof. Equation (18) is equivalent to the integral equation

$$(I - \mathcal{C}\Phi_x)(\mu - 1) = \mathcal{C}\Phi_x(1). \tag{24}$$

Note that $\mathcal{C}\Phi_x(1) \in L^r(\mathbb{R}^2)$ for any $r \geq \widetilde{p}_1$. Now, from Lemma 3.2, we know that $\mathcal{C}\Phi_x$ is bounded on $L^r(\mathbb{R}^2)$, $r > 2$. Moreover, the operator is compact (see [26, Lemma 4.2]), and hence (24) is a Fredholm equation of the second kind. Since the associated homogeneous equation has only the trivial solution (see, for instance, [6]), we can define

$$\mu - 1 = [I - \mathcal{C}\Phi_x]^{-1}(\mathcal{C}\Phi_x(1)) \in L^r(\mathbb{R}^2), \quad r \geq \widetilde{p}_1. \tag{25}$$

By (21),

$$\mathcal{C}\Phi_x(1) \in C^\alpha(\mathbb{R}^2), \quad \alpha = 1 - \frac{2}{p_2}, \tag{26}$$

$$\mathcal{C}\Phi_x(\mu - 1) \in C^\alpha(\mathbb{R}^2), \quad \alpha < 1 - \frac{2}{p_2}, \tag{27}$$

and then the Hölder regularity of $\mu - 1$ is obtained from (24).

Next, we show continuity of the map $x \mapsto (\mu(x,\cdot) - 1)$ from $\mathbb{R}^2$ into $L^r(\mathbb{R}) \cap C^\alpha(\mathbb{R}^2)$. By Lemma 3.2, we know that $\mathcal{C}\Phi_x 1 \in L^r(\mathbb{R}^2) \cap C^\alpha(\mathbb{R}^2)$ depends continuously on x.

Also, by Lemma 3.2, the map $x \mapsto \mathcal{C}\Phi_x$ is continuous from $\mathbb{R}^2$ to $\mathcal{L}(L^r(\mathbb{R}^2))$. Since the operator $I - \mathcal{C}\Phi_x$ is invertible for all $x \in \mathbb{R}^2$, the map $x \mapsto [I - \mathcal{C}\Phi_x]^{-1}$ is continuous from $\mathbb{R}^2$ to $\mathcal{L}(L^r(\mathbb{R}^2))$ as well. Hence the right-hand side of (25) depends continuously on x as a map from $\mathbb{R}^2$ into $L^r(\mathbb{R}^2)$. The continuity into $C^\alpha(\mathbb{R}^2)$ now follows as before from (21) and (24). ∎

Next, we consider the solvability of (18) in the case where ϕ is compactly supported.

LEMMA 3.4. *Suppose* $\phi \in L^p(\mathbb{R}^2)$, $p > 2$, *is compactly supported. Then* (18) *has a unique solution* μ *with* $\mu - 1 \in L^r \cap C^\alpha(\mathbb{R}^2), r > 2, \alpha < 1 - 2/p$. *Moreover, the map*

$$x \mapsto \mu(x,\cdot) - 1 \tag{28}$$

is smooth from $\mathbb{R}^2$ *into* $L^r \cap C^\alpha(\mathbb{R}^2)$.

Proof. Since $\phi \in L^{p_1} \cap L^p(\mathbb{R}^2)$ for $1 < p_1 < 2$, by Lemma 3.3, (18) has a unique solution μ with $\mu - 1 \in L^r \cap C^\alpha(\mathbb{R}^2), r > 2, \alpha < 1 - 2/p$, depending continuously on x.

To prove that μ is smooth, we will show first that $\partial_{x_1}\mu$ is continuous. By applying the differential operator ∂_{x_1} to (18), it follows that ∂_{x_1} satisfies the equation

$$\overline{\partial}_k \partial_{x_1}\mu = \phi(k)e_{-k}\overline{\partial_{x_1}\mu} - \phi(k)k_1 e_{-k}\overline{\mu}. \tag{29}$$

Since $\mu - 1 \in L^r(\mathbb{R}^2)$ for any $r > 2$, we have $\phi(k)k_1 e_{-k}\overline{\mu} \in L^q(\mathbb{R}^2)$ for any $q < p$. Hence $\mathcal{C}(\phi(k)k_1 e_{-k}\overline{\mu}) \in L^r(\mathbb{R}^2)$ for any $r > 2$. Equation (29) then has the unique solution

$$\partial_{x_1}\mu = -(I - \mathcal{C}\Phi_x)^{-1}(\phi(k)k_1 e_{-k}\overline{\mu}) \in L^r(\mathbb{R}^2) \cap C^\alpha(\mathbb{R}^2), \quad r > 2, \quad \alpha < 1 - 2/p.$$

Since $(\phi(k)k_1 e_{-k}\overline{\mu})$ and $(I - \mathcal{C}\Phi_x)^{-1}$ are continuous with respect to x, so is $\partial_{x_1}\mu$. Using induction, this argument can easily be extended to show that all x-derivatives of μ are continuous, i.e., that μ is smooth. ∎

We can now show using Lemma 3.3 that (12) admits a unique solution. Moreover, by using Lemma 3.4, we can show that (15) and (17) are uniquely solvable and that the solutions are smooth functions of the x-variable.

PROPOSITION 3.5. *Let* $\Omega \subset \mathbb{R}^2$ *be the unit disc, and suppose* γ *satisfies* (1) *with* $\gamma = 1$ *near* $\partial\Omega$.

(a) *Suppose further that* $\gamma \in W^{1+\epsilon,p}(\Omega)$ *with* $2 < p$. *Then for each* $x \in \mathbb{R}^2$ *and* $R > 0$, (12) *and* (15) *have unique solutions* μ, μ_R, *respectively, which satisfy* $\mu(x, \cdot) - 1 \in L^r \cap \mathbb{C}^\alpha(\mathbb{R}^2), r > 2/\epsilon, \alpha < 1$, *and* $(\mu_R(x, \cdot) - 1) \in L^r \cap C^\alpha(\mathbb{R}^2), r > 2, \alpha < 1$. *Furthermore,* $\mu_R(x, \cdot)$ *is smooth with respect to* x.

(b) *For* $x \in \mathbb{R}^2$ *and* $R > 0$, (17) *has a unique solution* $\mu_R^{\exp}(x, \cdot)$ *with* $\mu_R^{\exp}(x, \cdot) - 1 \in L^r \cap C^\alpha(\mathbb{R}^2), r > 2, \alpha < 1$, *which is smooth with respect to* x.

Proof. To prove (a), we use the fact from [18, 20] that $\mathbf{t}(k)/\overline{k} \in L^p(\mathbb{R}^2), 2 - \epsilon < p < \infty$. Hence $\phi(k) = \mathbf{t}(k)/(4\pi\overline{k})$ satisfies the assumptions in Lemma 3.3, and the claim follows. Furthermore, $\phi = \mathbf{t}_{\mathrm{R}}/(4\pi\overline{k})$ satisfies the assumptions in Lemma 3.4. It follows that (15) has a unique solution μ_{R} with the indicated properties.

To prove (b), we note that $\mathbf{t}_{\mathrm{R}}^{\exp}$ is a bounded function with compact support. Then again we use Lemma 3.4. ∎

As a consequence of this proposition, it follows that the reconstructions $\gamma_{\mathrm{R}}(x) = (\mu_{\mathrm{R}}(x, 0))^2$ and $\gamma_{\mathrm{R}}^{\exp} = (\mu_{\mathrm{R}}^{\exp}(x, 0))^2$ based on truncated scattering data are smooth functions.

4. Stability of the approximate reconstruction method. In this section, we show that the reconstruction method using the truncated $\mathbf{t}^{\exp}$ is stable. We will start by formulating the reconstruction procedure as an operator. Let $L_c^p(\mathbb{R}^2)$ denote the space of $L^p(\mathbb{R}^2)$ functions with compact support, and define for $k \in \mathbb{C}$ the linear operator $\mathcal{T}_{\mathrm{R}}^{\exp} : \mathcal{L}(H^{1/2}(\partial\Omega), H^{-1/2}(\partial\Omega)) \to L_c^\infty(\mathbb{R}^2)$ by

$$(\mathcal{T}_{\mathrm{R}}^{\exp} L)(k) = \chi_{|k|<R} \frac{1}{4\pi\overline{k}} \int_{\partial\Omega} (e^{i\overline{kx}} - 1)L(e^{ikx} - 1)d\sigma(x). \tag{30}$$

Define further for $p > 2$ the nonlinear operator

$$\mathcal{S} : L_c^p(\mathbb{R}^2) \to C^\infty(\overline{\Omega}), \quad \phi \mapsto \mu(x, 0),$$

where $\mu(x,\cdot)$ is the unique solution to (10) (see Lemma 3.4). By composition, we then define $\mathcal{M}_{\mathrm{R}}^{\exp} : \mathcal{L}(H^{1/2}(\partial\Omega), H^{-1/2}(\partial\Omega)) \to C^\infty(\overline{\Omega})$ by

$$\mathcal{M}_{\mathrm{R}}^{\exp} = \mathcal{S} \circ \mathcal{T}_{\mathrm{R}}^{\exp}. \tag{31}$$

Using this notation, it is clear that

$$(\gamma_{\mathrm{R}}^{\exp}(x))^{1/2} = \mu_{\mathrm{R}}^{\exp}(x,0) = \mathcal{M}_{\mathrm{R}}^{\exp}(\Lambda_\gamma - \Lambda_1), \tag{32}$$

since $(\Lambda_\gamma - \Lambda_1)1 = 0$ and $\int_{\partial\Omega}(\Lambda_\gamma - \Lambda_1) f d\sigma(x) = 0$ for all $f \in H^{1/2}(\partial\Omega)$. Thus $\mathcal{M}_{\mathrm{R}}^{\exp}$ is an operator that implements the reconstruction algorithm based on the truncated approximate scattering data.

The main goal of this section is to show that $\mathcal{M}_{\mathrm{R}}^{\exp}$ is continuous as an operator from $\mathcal{L}(H^{1/2}(\partial\Omega), H^{-1/2}(\partial\Omega))$ into $C^\infty(\overline{\Omega})$. This will show that the reconstruction algorithm using $\mathbf{t}_{\mathrm{R}}^{\exp}$ is stable.

LEMMA 4.1. *The operator $\mathcal{T}_R^{\exp}$ is bounded from $\mathcal{L}(H^{1/2}(\partial\Omega), H^{-1/2}(\partial\Omega))$ into $L_c^\infty(\mathbb{R}^2)$ and satisfies*

$$\|\mathcal{T}_{\mathrm{R}}^{\exp} L\|_{L^\infty(\mathbb{R}^2)} \le C e^{2R} \|L\|_{\mathcal{L}(H^{1/2}(\partial\Omega), H^{-1/2}(\partial\Omega))}. \tag{33}$$

Proof. For $|k| < R$, it is straightforward to obtain the estimate

$$|\mathcal{T}_{\mathrm{R}}^{\exp} L(k)| \le C \frac{1}{|k|} \|e^{ikx} - 1\|^2_{H^{1/2}(\partial\Omega)} \|L\|_{\mathcal{L}(H^{1/2}(\partial\Omega), H^{-1/2}(\partial\Omega))}.$$

Hence (33) follows from the uniform estimate $\|e^{ikx} - 1\|_{H^{1/2}(\partial\Omega)} \le C|k|^{1/2} e^{|k|}$. ∎

Next, we consider the solution operator $\mathcal{S}$. A stability estimate for this operator was given in [18, Lemma 3.1.5]; we will generalize this result slightly. The aim is to show that the solution μ to (18) depends continuously on the coefficient ϕ. Let $\mu_j, j = 1, 2$, be the solution to

$$\begin{aligned} \mu_j(x,s) - 1 = {} & \frac{1}{\pi} \int_{\mathbb{R}^2} \frac{\phi_j(k)}{s-k} e_{-x}(k) (\overline{\mu_j(x,k)} - 1) dk_1 dk_2 \\ & + \frac{1}{\pi} \int_{\mathbb{R}^2} \frac{\phi_j(k)}{s-k} e_{-x}(k) dk_1 dk_2, \quad j = 1, 2. \end{aligned} \tag{34}$$

Then we have the following result.

LEMMA 4.2. *Let $1 < p_1 < 2 < p_2 < \infty$ with $0 < 1/p_1 + 1/p_2 - 1/2 < 1/2$, and suppose $\phi_j \in L^{p_1}(\mathbb{R}^2) \cap L^{p_2}(\mathbb{R}^2)$, $j = 1, 2$. Further, let $x \in \overline{\Omega}$. Then, for the solution $\mu_j(x,\cdot)$ to (34), we have the estimate*

$$\|\mu_1(x,\cdot) - \mu_2(x,\cdot)\|_{C^\alpha(\mathbb{R}^2)} \le C K_1 K_2 \|\phi_1 - \phi_2\|_{L^{p_1}(\mathbb{R}^2) \cap L^{p_2}(\mathbb{R}^2) \cap L^2(\mathbb{R}^2) \cap L^q(\mathbb{R}^2)}, \tag{35}$$

where $\alpha < 1 - 2/q$, $1/q = 1/p_2 + 1/p_1 - 1/2$, and $K_j = \exp(C\|\phi_j\|_{L^{p_1}(\mathbb{R}^2) \cap L^{p_2}(\mathbb{R}^2)})$. If $\phi_1, \phi_2 \in L_c^p(\mathbb{R}^2)$, $p > 2$, we have the estimate

$$\|\mu_1(x,\cdot) - \mu_2(x,\cdot)\|_{C^\alpha(\mathbb{R}^2)} \le C K_1 K_2 \|\phi_1 - \phi_2\|_{L^p(\mathbb{R}^2)} \tag{36}$$

for $\alpha < 1 - 2/p$.

Proof. From [4, Lemma 2.6], we know that if $a \in L^{p_1}(\mathbb{R}^2) \cap L^{p_2}(\mathbb{R}^2)$ for $1 < p_1 < 2 < p_2 < \infty$ and $b \in L^p(^2)$ for $1 < p < 2$, then the solution to the integral equation

$$m = \mathcal{C}(a\overline{m}) + \mathcal{C}(b)$$

satisfies the estimate

$$\|m\|_{L^{\tilde{p}}(\mathbb{R}^2)} \le C\exp(C\|a\|_{L^{p_1}(\mathbb{R}^2)\cap L^{p_2}(\mathbb{R}^2)})\|b\|_{L^p(\mathbb{R}^2)}. \tag{37}$$

Applied to (34), the estimate reads

$$\|\mu_1(x,\cdot)-1\|_{L^{\widetilde{p_1}}(\mathbb{R}^2)} \le CK_1\|\phi_1\|_{L^{p_1}(\mathbb{R}^2)}. \tag{38}$$

Since

$$\begin{aligned}\mu_1(x,s)-\mu_2(x,s) &= \frac{1}{\pi}\int_{\mathbb{R}^2}\frac{\phi_2(k)}{s-k}e_{-x}(k)(\overline{\mu_1-\mu_2})dk_1dk_2\\ &\quad+\frac{1}{\pi}\int_{\mathbb{R}^2}\frac{\phi_1(k)-\phi_2(k)}{s-k}e_{-x}(k)\overline{\mu_1(x,k)}dk_1dk_2,\end{aligned} \tag{39}$$

the estimate (37) applied to $\mu_1-\mu_2$ then gives

$$\begin{aligned}\|\mu_1-\mu_2\|_{L^{\widetilde{p_1}}(\mathbb{R}^2)} &\le CK_2\|(\phi_1-\phi_2)\mu_1\|_{L^{p_1}(\mathbb{R}^2)}\\ &\le CK_2\big(\|\phi_1-\phi_2\|_{L^{p_1}(\mathbb{R}^2)}+\|\phi_1-\phi_2\|_{L^2(\mathbb{R}^2)}\|\mu_1-1\|_{L^{\widetilde{p_1}}(\mathbb{R}^2)}\big)\\ &\le CK_2\left(\|\phi_1-\phi_2\|_{L^{p_1}(\mathbb{R}^2)}+\|\phi_1-\phi_2\|_{L^2(\mathbb{R}^2)}\Phi_1\|\phi_1\|_{L^{p_1}(\mathbb{R}^2)}\right),\end{aligned} \tag{40}$$

where we have used (38). To get the Hölder estimate, we use (39) and (21) to obtain

$$\|\mu_1-\mu_2\|_{C^\alpha(\mathbb{R}^2)} \le \|\phi_2(\mu_1-\mu_2)\|_{L^q(\mathbb{R}^2)}+\|(\phi_1-\phi_2)\mu_1\|_{L^q(\mathbb{R}^2)} \tag{41}$$

for $q>2$ and $\alpha=1-2/q$. By choosing $1/q=1/p_2+1/\widetilde{p_1}$ ($<1/2$ by assumption), we have by (40) and (38)

$$\begin{aligned}\|\phi_2(\mu_1-\mu_2)\|_{L^q(\mathbb{R}^2)} &\le \|\phi_2\|_{L^{p_2}(\mathbb{R}^2)}\|\mu_1-\mu_2\|_{L^{\widetilde{p_1}}(\mathbb{R}^2)}\\ &\le \|\phi_2\|_{L^{p_2}(\mathbb{R}^2)}CK_2\big(\|\phi_1-\phi_2\|_{L^{p_1}(\mathbb{R}^2)}\\ &\quad+\|\phi_1-\phi_2\|_{L^2(\mathbb{R}^2)}K_1\|\phi_1\|_{L^{p_1}(\mathbb{R}^2)}\big)\end{aligned} \tag{42}$$

$$\begin{aligned}\|(\phi_1-\phi_2)\mu_1\|_{L^q(\mathbb{R}^2)} &\le \|\phi_1-\phi_2\|_{L^{p_2}(\mathbb{R}^2)}\|\mu_1-1\|_{L^{\widetilde{p_1}}(\mathbb{R}^2)}+\|\phi_1-\phi_2\|_{L^q(\mathbb{R}^2)}\\ &\le \|\phi_1-\phi_2\|_{L^{p_2}(\mathbb{R}^2)}CK_1\|\phi_1\|_{L^{p_1}(\mathbb{R}^2)}+\|\phi_1-\phi_2\|_{L^q(\mathbb{R}^2)}.\end{aligned} \tag{43}$$

Combining (41) with (42) and (43) gives (35).

Finally, (36) follows from (35) by using the compact support of ϕ_1,ϕ_2. ∎

As a direct consequence of the lemma, we obtain the following result.

COROLLARY 4.3. *The operator $\mathcal{S}$ is bounded from $L^p_c(\mathbb{R}^2)$, $p>2$, into $L^\infty(\Omega)$ and*

$$\|\mathcal{S}(\phi_1)-\mathcal{S}(\phi_2)\|_{L^\infty(\Omega)} \le C\|\phi_1-\phi_2\|_{L^p(\mathbb{R}^2)}, \tag{44}$$

where C depends on p, the support of ϕ_1,ϕ_2, and $\|\phi_1\|_{L^p(\mathbb{R}^2)},\|\phi_2\|_{L^p(\mathbb{R}^2)}$.

Proof. For fixed $x\in\overline{\Omega}$, (36) implies that

$$|\mathcal{S}(\phi_1)-\mathcal{S}(\phi_2)| = |\mu_1(x,0)-\mu_2(x,0)| \le C\Phi_1\Phi_2\|\phi_1-\phi_2\|_{L^p(\mathbb{R}^2)}.$$

This proves the result. ∎

We have now seen that the linear operator $\mathcal{T}_{\mathrm{R}}^{\exp}$ is bounded and the operator $\mathcal{S}$ is continuous. This enables us to conclude that $\mathcal{M}_{\mathrm{R}}^{\exp}=\mathcal{S}\circ\mathcal{T}_{\mathrm{R}}^{\exp}$ is continuous.

5. Convergence of reconstructions of mollified conductivities. Conductivities in practical applications of EIT are often piecewise smooth, but the theory of D-bar method covers only differentiable conductivities. We exclude the possibility of systematic artifacts introduced by discontinuities by proving the following: smooth approximations to nonsmooth conductivities yield almost the same reconstructions.

Let $\Omega = B(0,1)$. Let $c_0 > 0$ and $0 < R < 1$, and define $X = X(c_0, R) \subset L^\infty(\Omega)$ by

$$X = \{\gamma \in L^\infty(\Omega) \mid c_0^{-1} \le \gamma \le c_0,\ \mathrm{supp}(\gamma - 1) \subset B(0,R)\}.$$

The following lemma contains a continuity result for the operator $\gamma \mapsto \Lambda_\gamma$.

LEMMA 5.1. *Let $\gamma, \gamma_j \in X, j \in \mathbb{N}$, and suppose $\gamma_j \to \gamma$ a.e. Then, for any $s \in \mathbb{R}$, $\Lambda_{\gamma_j} - \Lambda_\gamma \longrightarrow 0$ in the strong topology of $\mathcal{L}(H^{1/2}(\partial\Omega), H^s(\partial\Omega))$.*

Proof. Let $f \in H^{1/2}(\partial\Omega)$. Let $\gamma_0 = \gamma$, and define $u_j, j = 0, 1, 2, \ldots$, as the unique solution to $\nabla \cdot \gamma_j \nabla u_j = 0$ in Ω, $u_j|_{\partial\Omega} = f$. Then

$$\|u_0\|_{H^1(\Omega)} \le C_2 \|f\|_{H^{1/2}}, \tag{45}$$

where the constant C_2 depends only on the uniform ellipticity constant c_0. Since

$$\nabla \cdot \gamma_j \nabla (u_j - u_0) = \nabla \cdot (\gamma - \gamma_j) \nabla u_0, \quad (u_j - u_0)|_{\partial\Omega} = 0,$$

there is a constant C_3 (depending only on c_0) such that the estimate

$$\|(u_j - u_0)\|_{H^1(\Omega)} \le C_3 \|(\gamma - \gamma_j)\nabla u_0\|_{L^2(\Omega)} \tag{46}$$

holds. Furthermore, $\Delta(u_j - u_0) = 0$ and $\partial_\nu (u_j - u_0)|_{\partial\Omega} = 0$ in the region $\Omega \setminus B(0,R)$. Therefore we can extend (46) to

$$\|(u_j - u_0)\|_{H^s(\Omega \setminus B(0,R_1))} \le C_4 \|(\gamma - \gamma_j)\nabla u_0\|_{L^2(\Omega)},$$

for any $s \in \mathbb{R}$, where C_4 depends on s, c_0, and $R_1 \in (R, 1)$. By taking normal derivative at the boundary, we then obtain, for any $s \in \mathbb{R}$,

$$\|(\Lambda_{\gamma_j} - \Lambda_\gamma) f\|_{H^s(\partial\Omega)} \le C_5 \|(\gamma - \gamma_j)\nabla u_0\|_{L^2(\Omega)}, \tag{47}$$

where C_5 depends on s, c_0, and R.

To consider convergence in the strong topology, let us fix f and γ, implying that u_0 can be considered as a fixed function. Then using Lebesgue dominated convergence, it follows that $\lim_{j\to\infty} \|(\gamma - \gamma_j)\nabla u_0\|_{L^2(\Omega)} = 0$. By (47), this implies the claim. ∎

The next lemma shows that a strongly convergent sequence of operators is norm convergent when composed with a compact operator defined on a Hilbert space.

LEMMA 5.2. *Let X, Y be Banach spaces and H be a separable Hilbert space. Suppose $T, T_j \in \mathcal{L}(X,Y)$, $j = 1, 2, \ldots$, and $K \in \mathcal{L}(H,X)$ is compact. If $T_j \to T$ as $j \to \infty$ in the strong topology of $\mathcal{L}(X,Y)$, then $T_j K \to TK$ as $j \to \infty$ in the norm topology of $\mathcal{L}(H,Y)$.*

Proof. By the principle of uniform boundedness, there is a constant C_0 such that $\|T\|_{\mathcal{L}(X,Y)} < C_0$ and $\|T_j\|_{\mathcal{L}(X,Y)} < C_0$ for all j.

Since the compact operator K maps from a separable Hilbert space into a Banach space, there is a sequence of finite rank operators $K_n, n \in \mathbb{N}$, with $\mathrm{rank}(K_n) = n$ that converges in norm to K (see [28, Theorem 6.13] for a proof of this fact in the Hilbert space case; the proof is the same in our case). Fix $\epsilon > 0$, and take n such that

$$\|K - K_n\|_{\mathcal{L}(H,X)} < \epsilon.$$

Further, since K_n has finite rank, there is a $J = J(\epsilon) \geq 0$ such that, for $j \geq J$,

$$\|(T_j - T)K_n\|_{\mathcal{L}(H,Y)} \leq \epsilon.$$

Hence, for $j \geq J$,

$$\|(T_j - T)K\|_{\mathcal{L}(H,Y)} \leq \|T_j(K - K_n) + (T_j - T)K_n + T(K_n - K)\|_{\mathcal{L}(H,Y)} = 3C_0\epsilon.$$

This proves the result. ∎

We will now use the preceding lemma to prove norm convergence of the sequence of DN maps in Lemma 5.1.

LEMMA 5.3. *Let γ and γ_j be as in Lemma* 5.1. *Then*

$$\lim_{j\to\infty} \|\Lambda_{\gamma_j} - \Lambda_\gamma\|_{\mathcal{L}(H^{1/2}(\partial\Omega),H^{-1/2}(\partial\Omega))} = 0.$$

Proof. Since the inclusion operator $J\colon H^{1/2+r}(\partial\Omega) \to H^{1/2}(\partial\Omega)$ is compact for any $r > 0$, it follows from Lemmas 5.1 and 5.2 that

$$\lim_{j\to\infty} \|\Lambda_{\gamma_j} - \Lambda_\gamma\|_{\mathcal{L}(H^{\frac{1}{2}+r}(\partial\Omega),H^s(\partial\Omega))} = \lim_{j\to\infty} \|(\Lambda_{\gamma_j} - \Lambda_\gamma)J\|_{\mathcal{L}(H^{\frac{1}{2}+r}(\partial\Omega),H^s(\partial\Omega))} = 0 \tag{48}$$

for $r > 0, s \in \mathbb{R}$. Further, since $\gamma = \gamma_j = 1$ near $\partial\Omega$, $\Lambda_{\gamma_j} - \Lambda_\gamma$ is a smoothing pseudo-differential operator that can be extended to an operator $\mathcal{D}'(\partial\Omega) \to C^\infty(\partial\Omega)$. An application of Green's formula implies that the operator $\Lambda_{\gamma_j} - \Lambda_\gamma : \mathcal{D}'(\partial\Omega) \to C^\infty(\partial\Omega)$ and its transpose $(\Lambda_{\gamma_j} - \Lambda_\gamma)' : \mathcal{D}'(\partial\Omega) \to C^\infty(\partial\Omega)$ coincide. Thus (48) implies

$$\lim_{j\to\infty} \|\Lambda_{\gamma_j} - \Lambda_\gamma\|_{\mathcal{L}(H^{-s'}(\partial\Omega),H^{-1/2-r'}(\partial\Omega))} = 0 \tag{49}$$

for $r' > 0, s' \in \mathbb{R}$. Interpolation of (48) and (49) gives the result; see, for example, Theorem 5.1 and section 7.3 of [22] or formula (3.2) on p. 282 of [33]. ∎

Let $\gamma \in X(c_0, R)$ for some $c_0, R > 0$, and suppose that γ is continuous a.e. Let $\eta \in C_0^\infty(D(0, \alpha/2))$ be nonnegative and $\int_{\mathbb{R}^2} \eta = 1$. Define $\eta_\lambda(x) := \lambda^{-2}\eta(x/\lambda)$ for any $0 < \lambda < 1$, and set $\gamma_\lambda := \eta_\lambda * \gamma$. We then have the following result.

THEOREM 5.4. *Let $\gamma \in X(c_0, R)$ for some $c_0, R > 0$, and let γ_λ be defined as above. Let $\mathcal{M}_R^{\exp}$ be defined as in* (31). *Then we have*

$$\lim_{\lambda\to 0} \|\mathcal{M}_{\mathrm{R}}^{\exp}(\Lambda_{\gamma_\lambda} - \Lambda_\gamma)\|_{\mathcal{L}(L^\infty(\Omega),L^\infty(\Omega))} = 0.$$

Proof. As a consequence of the definition, there exist $\widetilde{c}_0, \widetilde{R} > 0$ such that $\gamma, \gamma_\lambda \in X(\widetilde{c}_0, \widetilde{R})$ for λ sufficiently small. Also $\gamma_\lambda \to \gamma$ a.e. Using Lemma 5.3, it follows that Λ_{γ_λ} converges to Λ_γ in the norm topology of $\mathcal{L}(H^{1/2}(\partial\Omega), H^{-1/2}(\partial\Omega))$. Finally, using the continuity of $\mathcal{M}_{\mathrm{R}}^{\exp}$ (see section 4), we conclude that the reconstruction $\mathcal{M}_{\mathrm{R}}^{\exp}(\Lambda_{\gamma_\lambda} - \Lambda_1)$ of $\gamma_\lambda^{1/2}$ converges to $\mathcal{M}_{\mathrm{R}}^{\exp}(\Lambda_\gamma - \Lambda_1)$. ∎

6. Connection to Calderón's linearization method. In the seminal paper [7], Calderón gave an algorithm for the reconstruction of conductivities close to constant (see also [38]). We write Calderón's method in the context of the approximate scattering transform $\mathbf{t}^{\exp}$ and compare it to the D-bar method.

6.1. Calderón's linearization method. Integrating by parts in (4) gives

$$\mathbf{t}^{\exp}(k) = \int_\Omega (\gamma - 1)\nabla u(x,k) \cdot \nabla(e^{i\overline{kx}})dx, \tag{50}$$

where

$$\nabla \cdot (\gamma - 1)\nabla u = 0 \text{ in } \Omega, \quad u|_{\partial\Omega} = e^{ikx}.$$

When $\|\gamma - 1\|_{L^\infty(\Omega)}$ is small, then u is close to e^{ikx} inside Ω. Indeed, if we write $u = e^{ikx} + \delta u$ for $\delta u \in H_0^1(\Omega)$ satisfying $\nabla \cdot \gamma \nabla \delta u = -\nabla \cdot (\gamma - 1)\nabla(e^{ikx})$, we have the estimate

$$\|\delta u\|_{H^1(\Omega)} \leq C\|\gamma - 1\|_{L^\infty(\Omega)} e^{|k|r}, \tag{51}$$

where r is the radius of the smallest ball containing Ω. Substituting $u = e^{ikx} + \delta u$ into (50) and dividing by $-2|k|^2$, we obtain

$$\begin{aligned} -\frac{\mathbf{t}^{\exp}(k)}{2|k|^2} &= -\frac{1}{2|k|^2}\int_\Omega (\gamma - 1)\nabla(e^{ikx} + \delta u) \cdot \nabla(e^{i\overline{kx}})dx \\ &= \int_\Omega (\gamma - 1)e_k(x)dx + R(k) \\ &= 2\pi \mathcal{F}(\chi_\Omega(\gamma - 1))(-2k_1, 2k_2) + R(k), \end{aligned} \tag{52}$$

where $\mathcal{F}$ denotes the Fourier transform and

$$R(k) = -\frac{1}{2|k|^2}\int_\Omega (\gamma - 1)\nabla \delta u \cdot \nabla(e^{i\overline{kx}})dx.$$

Using (51), it is not hard to obtain

$$|R(k)| \leq C\|\gamma - 1\|^2_{L^\infty(\Omega)} e^{2|k|r}. \tag{53}$$

The idea behind Calderón's method is to multiply (52) by a smooth cut-off function and then apply the inverse Fourier transform. Let $\hat{\eta} \in C_0^\infty(\mathbb{R}^2)$ be a nonnegative function supported in the unit ball with $\hat{\eta} = 1$ near $x = 0$, and let σ be a positive parameter determining the cut-off radius. Then from (52) we obtain

$$\mathcal{F}(\chi_\Omega(\gamma - 1))(-2k_1, 2k_2)\hat{\eta}\left(\frac{k}{\sigma}\right) = -\frac{\mathbf{t}^{\exp}(k)}{4\pi|k|^2}\hat{\eta}(k/\sigma) - R(k)\hat{\eta}\left(\frac{k}{\sigma}\right).$$

Changing variables $s = (s_1, s_2) = 2(-k_1, k_2)$ gives

$$\begin{aligned} &\mathcal{F}(\chi_\Omega(\gamma - 1))(s_1, s_2)\hat{\eta}\left(\frac{(-s_1, s_2)}{(2\sigma)}\right) \\ &= -\frac{\mathbf{t}^{\exp}((-s_1, s_2)/2)}{\pi|s|^2}\hat{\eta}\left(\frac{(-s_1, s_2)}{(2\sigma)}\right) - R\left(\frac{(-s_1, s_2)}{2}\right)\hat{\eta}\left(\frac{(-s_1, s_2)}{(2\sigma)}\right). \end{aligned} \tag{54}$$

Inverting $\mathcal{F}$ and neglecting the second term in (54) yields an approximation to γ:

$$\gamma^{\text{app}}(x) - 1 = -\frac{1}{2\pi}\int_{\mathbb{R}^2} e^{ix\cdot s}\frac{\mathbf{t}^{\exp}((-s_1, s_2)/2)}{\pi|s|^2}\hat{\eta}\left(\frac{(-s_1, s_2)}{(2\sigma)}\right)ds_1 ds_2.$$

Changing back the variables in the integral to $(k_1, k_2) = (-s_1, s_2)/2$ yields the formula

$$\begin{aligned}\gamma^{\text{app}}(x) - 1 &= -\frac{1}{2\pi^2}\int_{\mathbb{R}^2} e^{2i(-x_1k_1+x_2k_2)}\frac{\mathbf{t}^{\exp}(k)}{4|k|^2}\hat{\eta}\left(\frac{k}{\sigma}\right)4dk_1dk_2 \\ &= -\frac{2}{(2\pi)^2}\int_{\mathbb{R}^2} e_{-x}(k)\frac{\mathbf{t}^{\exp}(k)}{|k|^2}\hat{\eta}\left(\frac{k}{\sigma}\right)dk_1dk_2. \qquad (55)\end{aligned}$$

The reconstruction γ^{app} is an approximation of a low-pass filtered version of γ. Choosing the parameter σ as in [7] with $0 < \alpha < 1$ to be

$$\sigma = \frac{1-\alpha}{2r}\log\frac{1}{\|\gamma - 1\|_{L^\infty(\Omega)}} \qquad (56)$$

yields due to (53) the error estimate

$$\|\gamma^{\text{app}}(x) - \eta_\sigma * \gamma\|_{L^\infty} \le \|R(k)\hat{\eta}(k/\sigma)\|_{L^1(\mathbb{R}^2)} \le C\|\gamma - 1\|^{1+\alpha}_{L^\infty(\Omega)}(\log(\|\gamma - 1\|_{L^\infty(\Omega)}))^2.$$

Note that when $\|\gamma - 1\|_{L^\infty(\Omega)}$ is sufficiently small this error is much smaller than $\|\gamma - 1\|_{L^\infty(\Omega)}$. Since in most applications the approximate magnitudes of the conductivities comprising $\gamma(x)$ are known, an estimate to σ in (56) can be computed.

In summary, the algorithm proposed in [7] is tantamount to the following:

1. Compute $\mathbf{t}^{\exp}(k)$ by (4).
2. Construct a low-pass filter $\hat{\eta}(k/\sigma)$.
3. Compute the approximation γ^{app} by (55).

6.2. Calderón's method as an approximation of the D-bar method. Calderón's method using (55) can be seen as a three-step approximation of the D-bar method using (3) and (12)–(13):

1. In (12), $\mathbf{t}(k)$ is approximated by $\mathbf{t}^{\exp}(k)\hat{\eta}(k/\sigma)$, where $\hat{\eta}(k/\sigma)$ is a smooth cut-off function.
2. The function μ in the integral in the right-hand side of (12) is approximated by its asymptotic value $\mu \sim 1$.
3. The square function in (13) is linearized: $(1-h)^2 \sim 1 - 2h$.

In contrast, the D-bar method using $\mathbf{t}^{\exp}_{\text{R}}$ makes only the first approximation (with sharp cut-off).

7. Analysis of a simple radial conductivity distribution. In this section, we consider the simple example of a piecewise constant radial conductivity defined in the unit disc Ω. We will show that in this case $\mathbf{t}^{\exp}$ can be expanded conveniently using Bessel functions. Furthermore, such an expansion leads to a result concerning the asymptotic behavior of $\mathbf{t}^{\exp}(k)$ as $|k|$ tends to infinity. We will write out explicit formulas for γ^{app}, the Calderón reconstruction, and $\gamma^{\exp}$, the D-bar reconstruction of γ with the $\mathbf{t}^{\exp}$ approximation.

Consider the radial conductivity

$$\gamma(x) = \begin{cases} \sigma & \text{for } |x| \le r, \\ 1 & \text{for } |x| > r, \end{cases} \qquad (57)$$

where $0 < r < 1$ and $\sigma > 0, \sigma \ne 1$. Define

$$\alpha = \frac{\sigma - 1}{\sigma + 1}. \qquad (58)$$

According to [32], trigonometric basis functions are eigenfunctions for Λ_γ when γ is radial. More precisely, $\Lambda_\gamma \varphi_n = \lambda_n \varphi_n$ with $\varphi_n(\alpha) := (2\pi)^{-1/2} e^{in\alpha}$ for $n \in \mathbb{Z}$. It is well known [12] that the eigenvalues of Λ_γ corresponding to (57) are given by

$$\lambda_n = n\left(1 + \frac{2\alpha r^{2n}}{1-\alpha r^{2n}}\right), \quad n = 1, 2, 3, \ldots. \tag{59}$$

Hence $\delta\Lambda \equiv \Lambda_\gamma - \Lambda_1$ has eigenvalues

$$\delta\lambda_n = \frac{2n\alpha r^{2n}}{1-\alpha r^{2n}}, \quad n = 1, 2, 3, \ldots. \tag{60}$$

As in [29], one can derive a series representation of $\mathbf{t}^{\exp}(k)$. For the case of the conductivity distribution (57), this leads to a particularly simple representation of $\mathbf{t}^{\exp}$ in terms of Bessel functions, which we derive here. Expanding e^{ikx} in a Fourier series on the circle $x = e^{i\theta}$ yields [13]

$$e^{ikx} = \sum_{n=-\infty}^{\infty} a_n(k) e^{in\theta} \quad \text{with} \quad a_n(k) = \begin{cases} \frac{(ik)^n}{n!}, & n \geq 0, \\ 0, & n < 0. \end{cases}$$

Substituting this series into formula (4) and using (59) gives a series for $\mathbf{t}^{\exp}(k)$:

$$\mathbf{t}^{\exp}(k) = 2\pi \sum_{n=1}^{\infty} (\lambda_n - n) \frac{(-1)^n |k|^{2n}}{(n!)^2} = 4\pi\alpha \sum_{n=1}^{\infty} \frac{n r^{2n}}{1-\alpha r^{2n}} \frac{(-1)^n |k|^{2n}}{(n!)^2}. \tag{61}$$

Write $(1-\alpha r^{2n})^{-1} = \sum_{m=0}^{\infty} (\alpha r^{2n})^m$ so that

$$\mathbf{t}^{\exp}(k) = 4\pi\alpha \sum_{m=0}^{\infty} \alpha^m \sum_{n=1}^{\infty} \frac{n(-1)^n}{(n!)^2} (|k| r^{m+1})^{2n}. \tag{62}$$

Note that the Bessel function $J_1(t) = -J_0'(t) = -(2/t) \sum_{j=1}^{\infty} j(-1)^j (j!)^{-2} (t/2)^{2j}$. Thus, with $t = 2r^{m+1}|k|$,

$$\mathbf{t}^{\exp}(k) = -4\pi\alpha |k| r \sum_{m=0}^{\infty} (\alpha r)^m J_1(2r^{m+1}|k|). \tag{63}$$

This formula gives an accurate way for computing $\mathbf{t}^{\exp}$ numerically. Furthermore, we can derive the asymptotic behavior of $\mathbf{t}^{\exp}$ from (63).

For small z, $J_1(z) \sim z/2$. So using $\sum_{m=0}^{\infty} (\alpha r^2)^m = 1/(1-\alpha r^2)$ yields

$$\mathbf{t}^{\exp}(k) \sim -4\pi\alpha |k|^2 r \sum_{m=0}^{\infty} \alpha^m r^{2m+1} = \frac{-4\pi\alpha (|k| r)^2}{1-\alpha r^2} = O(|k|^2) \quad \text{for small } |k|.$$

For large $|z|$, we have $J_1(z) \sim (2/(\pi z))^{1/2} \cos(z - 3\pi/4) + e^{|\mathrm{Im}\, z|} O(1/|z|)$, and so

$$\mathbf{t}^{\exp}(k) \sim -4\pi\alpha |k| r \sum_{m=0}^{\infty} (\alpha r)^m \sqrt{\frac{1}{\pi r^{m+1} |k|}} \cos\left(2r^{m+1}|k| - \frac{3\pi}{4}\right)$$

for large $|k|$. After some simplification, this results in the asymptotic formula

$$\mathbf{t}^{\exp}(k) \sim \frac{-4\pi\alpha |k|^{1/2} r^{1/2}}{\sqrt{\pi}} \sum_{m=0}^{\infty} (\alpha)^m r^{m/2} \cos\left(2r^{m+1}|k| - \frac{3\pi}{4}\right) = O(|k|^{1/2}). \tag{64}$$

Note that (64) shows the importance of truncation of $\mathbf{t}^{\exp}$: the solvability of the D-bar equation is not proven for $\mathbf{t}^{\exp}$ with asymptotic behavior (64).

7.1. Calderón's method for a simple radial conductivity. Let γ be of the form (57). Note from [14], for example, that the Fourier transform of the characteristic function χ_r is given by

$$\begin{aligned}\mathcal{F}^{-1}(\chi_r)(k) = \check{\chi}_r(k) &= \frac{1}{2\pi}\int_{\mathbb{R}^2}\chi_r(p)e^{ik\cdot p}dp = \frac{1}{2\pi}\int_0^r\int_0^{2\pi} e^{i|k|\rho\cos(\theta)}\rho d\rho d\theta\\ &= \frac{1}{2\pi}\sum_{j=0}^{\infty}\frac{(i|k|)^j}{j!}\int_0^r\rho^{j+1}d\rho\int_0^{2\pi}\cos^j(\theta)d\theta\\ &= \begin{cases}\sum_{j=0}^{\infty}\frac{(i|k|)^j}{j!}\frac{r^{j+2}}{j+2}\frac{1\cdot3\cdot5\cdots(j-1)}{2\cdot4\cdot6\cdots j}, & j \text{ even},\\ 0, & j \text{ odd}\end{cases}\\ &= \sum_{m=0}^{\infty}\frac{(m+1)(-1)^m}{((m+1)!)^2}\left(\frac{r|k|}{2}\right)^{2m}\frac{r^2}{2}\\ &= -\frac{2}{|k|^2}\sum_{n=1}^{\infty}\frac{n(-1)^n}{(n!)^2}\left(\frac{r|k|}{2}\right)^{2n}.\end{aligned}$$

Thus, from (62),

$$\mathbf{t}^{\exp}(k) = -8\pi|k|^2\alpha\sum_{m=0}^{\infty}\alpha^m\check{\chi}_{r^{m+1}}(2k). \tag{65}$$

For this simple case, we can substitute $\mathbf{t}^{\exp}$ directly into (55) and compute explicitly without multiplying by the cut-off function $\hat{\eta}$ (or, equivalently, take $\sigma=\infty$, implying $\hat{\eta}\equiv 1$). Thus Calderón's reconstruction γ^{app} is given by

$$\begin{aligned}\gamma^{\text{app}}(x) &= 1 - \frac{8}{(2\pi)^2}\int_{\mathbb{R}^2}e_{-x}(k)\frac{\mathbf{t}^{\exp}(k)}{|k|^2}dk_1dk_2\\ &= 1 + \frac{4}{\pi}\int_{\mathbb{R}^2}e_{-x}(k)\alpha\sum_{m=0}^{\infty}\alpha^m\check{\chi}_{r^{m+1}}(2|k|)dk_1dk_2\\ &= 1 + 2\alpha\sum_{m=0}^{\infty}\alpha^m\frac{1}{2\pi}\int_{\mathbb{R}^2}e^{-i(x_1w_1-x_2w_2)}\check{\chi}_{r^{m+1}}(|w|)dw_1dw_2\\ &= 1 + 2\alpha\sum_{m=0}^{\infty}\alpha^m\frac{1}{2\pi}\int_{\mathbb{R}^2}e^{-ix\cdot\bar{w}}\check{\chi}_{r^{m+1}}(|w|)dw_1dw_2\\ &= 1 + 2\alpha\sum_{m=0}^{\infty}\alpha^m\mathcal{F}(\check{\chi}_{r^{m+1}}(|s|)) \qquad (66)\\ &= 1 + 2\alpha\sum_{m=0}^{\infty}\alpha^m\chi_{r^{m+1}}(x). \qquad (67)\end{aligned}$$

Note that (67) preserves the location of the jump in the actual conductivity distribution γ. Furthermore, elementary calculations show $\gamma^{\text{app}}(0)=\sigma$; i.e., the correct value of the conductivity is attained at $x=0$.

Similar computations were done in [14] (see also [13]), where the starting point, however, was the Neumann-to-Dirichlet (ND) map. They found the approximation

$$\gamma(x) \approx 1 + 2\alpha\sum_{m=0}^{\infty}(-\alpha)^m\chi_{r^{m+1}}(x). \tag{68}$$

7.2. The D-bar method for a simple radial conductivity. The series (65) can be used in the analysis of the truncated D-bar method. We have

$$\begin{aligned}
\gamma_{\mathrm{R}}^{\exp}(x)^{1/2} \equiv \mu_{\mathrm{R}}^{\exp}(x,0) &= 1 - \frac{1}{4\pi^2}\int_{\mathbb{R}^2} \frac{\mathbf{t}^{\exp}(k)}{|k|^2} e_{-x}(k)\chi_R(|k|)\overline{\mu_{\mathrm{R}}^{\exp}(x,k)}dk_1dk_2 \\
&= 1 + \frac{2\alpha}{\pi}\sum_{m=0}^{\infty}\alpha^m \int_{\mathbb{R}^2} \check{\chi}_{r^{m+1}}(2|k|)e_{-x}(k)\chi_R(|k|)\overline{\mu_{\mathrm{R}}^{\exp}(x,k)}dk_1dk_2 \\
&= 1 + \frac{\alpha}{2\pi}\sum_{m=0}^{\infty}\alpha^m \int_{\mathbb{R}^2} \check{\chi}_{r^{m+1}}(|w|)e^{-ix\cdot\overline{w}}\chi_R(2|w|)\overline{\mu_{\mathrm{R}}^{\exp}\left(x,\frac{w}{2}\right)}dw_1dw_2 \\
&= 1 + \frac{\alpha}{2\pi}\sum_{m=0}^{\infty}\alpha^m \int_{\mathbb{R}^2} \check{\chi}_{r^{m+1}}(|s|)e^{-ix\cdot s}\chi_R(2|s|)\overline{\mu_{\mathrm{R}}^{\exp}\left(x,\frac{\overline{s}}{2}\right)}ds_1ds_2 \\
&= 1 + \alpha\sum_{m=0}^{\infty}\alpha^m \mathcal{F}\left(\check{\chi}_{r^{m+1}}(|\cdot|)\chi_R(2\cdot)\overline{\mu_{\mathrm{R}}^{\exp}\left(x,\frac{\cdot}{2}\right)}\right)(x) \\
&= 1 + \frac{\alpha}{2\pi}\sum_{m=0}^{\infty}\alpha^m \left(\chi_{r^{m+1}}(\cdot) * \mathcal{F}\left(\chi_R(2\cdot)\overline{\mu_{\mathrm{R}}^{\exp}\left(x,\frac{\cdot}{2}\right)}\right)\right)(x). \qquad (69)
\end{aligned}$$

It is evident from this formula that a ringing effect will appear in the reconstruction, but the effect will be somewhat blurred by the convolution of the characteristic functions with the Fourier transform of $\overline{\mu^{\exp}}$.

8. Numerical experiments. In this section, numerical examples are computed that offer intuition and illustrate the results of the previous sections.

8.1. Example conductivities. We consider discontinuous conductivities defined by (57) with all nine possible combinations of the choices $r \in \{0.2, 0.55, 0.9\}$ and $\sigma \in \{1.1, 2, 8\}$. See Figure 8.2 for profiles of the conductivities.

Radially symmetric examples are chosen for their ease of computation and display (it is sufficient to display profiles of the reconstructed conductivities and scattering transforms, which are real-valued and radially symmetric for radially symmetric examples), as well as to illustrate the results of sections 6 and 7.1. However, all of our computational methods apply equally well to nonsymmetric conductivities.

8.2. Results. The computed scattering transform is denoted by $\mathbf{t}_R^{\exp}$, where R indicates the truncation radius. We compute $\mathbf{t}_R^{\exp}$ from the Bessel-series formula (63) with 10 terms in the expansion, which was verified to be in very good agreement with computations of (61) with 45 eigenvalues. Figure 8.1 contains plots of the approximate scattering transforms $\mathbf{t}_R^{\exp}$ with constant multiples of $\sqrt{|k|}$ superimposed to illustrate the growth of $\mathbf{t}^{\exp}(k)$ as demonstrated in (64).

Plots of the reconstructed conductivities are found in Figures 8.2, 8.3, 8.4, and 8.5. Figure 8.2 contains plots of the reconstructed conductivities from the approximate scattering transform $\mathbf{t}_R^{\exp}$ with truncation radius $R = 15$ for rows 1 and 2 and $R = 12$ for row 3. Figure 8.3 illustrates the dependence of the reconstructions $\gamma_{\mathrm{R}}^{\exp}$ on R. Profiles of the reconstructed discontinuous conductivities with contrast 0.1 and a jump at $|x| = 0.2, 0.55$, and 0.9 are plotted for $R = 4, 5, 6, 7, 8$, and 15. The reconstructions from Calderón's linearization method are found in Figure 8.4. Finally, 2-D plots of three conductivities are included in Figure 8.5.

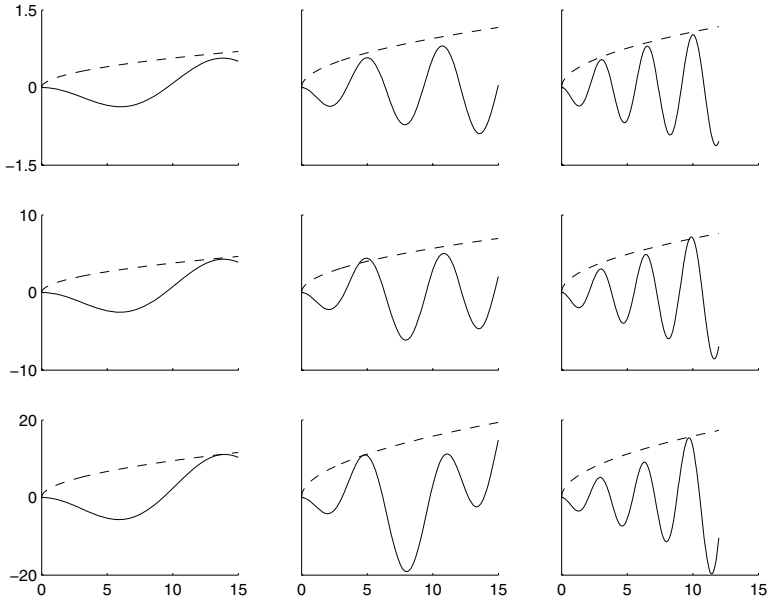


FIG. 8.1. *Profiles of approximate scattering transforms* $\mathbf{t}^{\exp}$ *(solid) for the discontinuous conductivity distributions given by* (57) *with constant multiples of* $\sqrt{|k|}$ *superimposed (dashed) to illustrate the growth of* $\mathbf{t}^{\exp}$*. The top row corresponds to* $\sigma = 1.1$ *in* (57)*, the middle row to* $\sigma = 2.0$*, and the bottom row to* $\sigma = 8.0$*. The first column corresponds to* $r = 0.2$ *in* (57)*, the second column to* $r = 0.55$*, and the third column to* $r = 0.9$*. Note that the vertical axis limits are the same in each row of plots.*

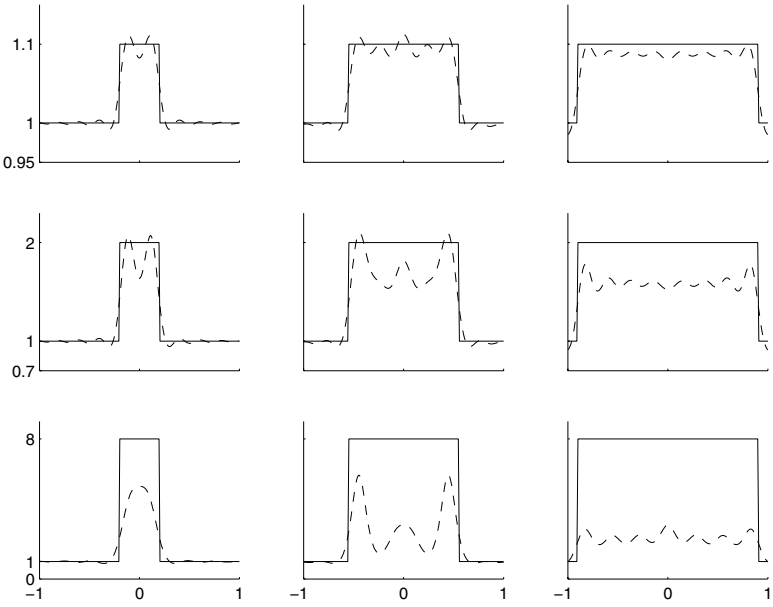


FIG. 8.2. *Actual (solid) and reconstructed (dashed) conductivity profiles* $\gamma_R^{\exp}$ *(*$R = 15$ *for the first two rows,* $R = 12$ *for the last row) for the discontinuous examples. Note that the vertical axis limits are the same in each row of plots.*

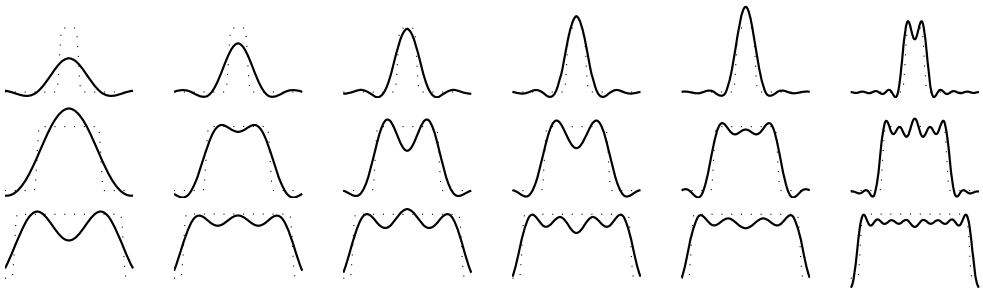


FIG. 8.3. *The discontinuous conductivity distributions (dotted lines) with a jump of* 0.1 *at* $|x| = 0.2$ *(top row),* 0.55 *(center row), and* 0.9 *(bottom row) reconstructed (solid line) from* $\mathbf{t}_R^{\exp}$ *with truncation radii* $R = 4, 5, 6, 7, 8, 15$ *(left to right). Note that the vertical axis limits are the same in each row of plots.*

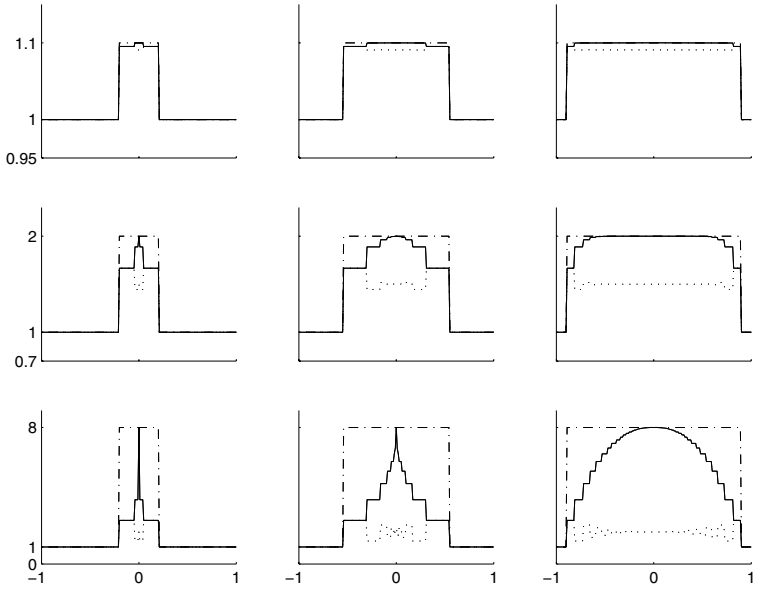


FIG. 8.4. *Actual (dash-dotted) and reconstructed conductivity profiles for the discontinuous examples. The dotted reconstructions are from Calderón's linearization formula* (68) *from the ND map, and the solid reconstructions are from Calderón's linearization formula* (67) *from the DN map. Note that the vertical axis limits are the same in each row of plots.*

8.3. Discussion. From Figure 8.1, we see that the scattering transforms demonstrate the expected asymptotic growth $\mathbf{t}^{\exp} \sim O(|k|^{1/2})$. The magnitude of the scattering transform increases with the amplitude of γ, and $\mathbf{t}^{\exp}$ becomes more oscillatory as $\operatorname{supp}(\gamma - 1)$ increases. This implies that conductivity distributions with high contrast near the boundary should be particularly difficult to reconstruct, because such a scattering transform is more sensitive to errors in $\partial\Lambda_\gamma$ and more difficult to represent on a discrete mesh.

We see from the corresponding reconstructions in Figure 8.2 that in all cases the location of the jump is reconstructed equally well, but a loss in accuracy in the amplitude becomes apparent as the contrast increases and as the support of $\gamma -$

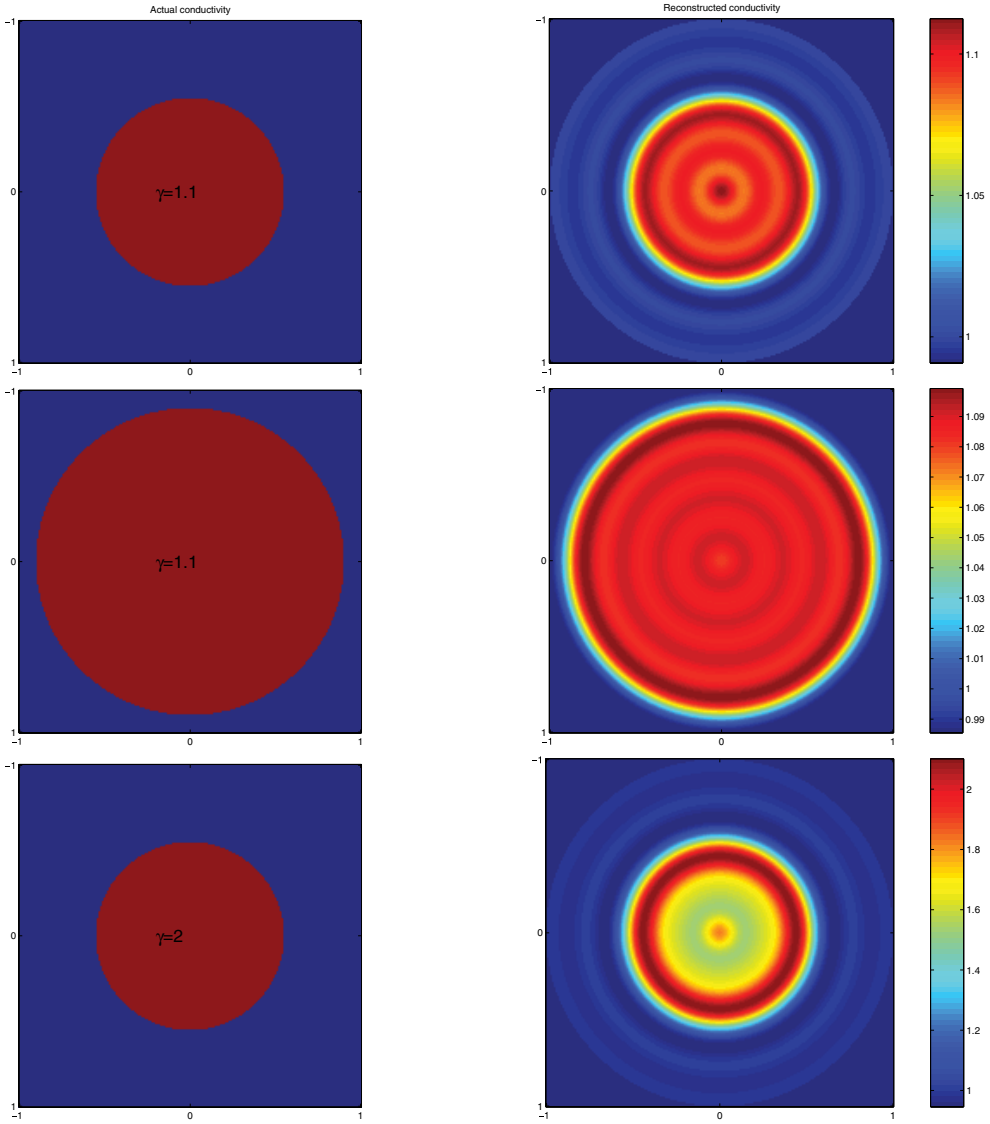


FIG. 8.5. *Three discontinuous conductivity distributions (left) reconstructed (right) from* $\mathbf{t}_R^{\exp}$ *with truncation radius* $R = 15$.

1 widens. This corresponds to the fact that the approximation $\psi|_{\partial\Omega} = e^{izk}|_{\partial\Omega}$ is better the smaller the contrast and the smaller the support of $\gamma - 1$. We see that the reconstructions tend to underestimate the actual amplitude of the conductivity more markedly as the support of $\gamma - 1$ widens and as the magnitude of γ increases. Also note that the reconstructions of the discontinuous conductivities are smooth, as predicted by Proposition 3.5. In Figure 8.3, the nature of the dependence of the smooth approximations on R can be observed. A Gibbs-type phenomenon is indeed present, as suggested by formula (69). Also, the support of $\gamma - 1$ is reconstructed with reasonable accuracy even for very small truncation radii, while the general shape and amplitude of γ is reconstructed with increasing accuracy as R increases. Thus, in

practice one should choose R as large as possible before blow-up of $\mathbf{t}^{\exp}$ is evident. In [29, 24, 15, 16], R was chosen by inspection.

The reconstructions from Calderón's linearization method are found in Figure 8.4. It is interesting to note that the linearized reconstruction from the DN map (67) achieves a more accurate approximation to the amplitude of the conductivity than the linearized reconstruction from the ND map (68). This result also holds for conductivities whose jump is negative ($0 < \sigma < 1$) in $|x| < r$. Note that the linearization formula (67) actually achieves the amplitude of the actual conductivity (albeit only at a single point in some cases), while the D-bar reconstruction $\gamma^{\exp}$ does not. This is presumably due to the damping effect of the convolution with the Fourier transform of $\overline{\mu^{\exp}}$ in (69).

Finally, in Figure 8.5 we display three reconstructions in the typical 2-D display mode for reconstructions from experimental data. This figure further illustrates the ringing effect in the reconstructions of the discontinuous conductivities.

Acknowledgments. The authors thank David Isaacson for helpful discussions and the anonymous reviewers for their suggestions.

REFERENCES

[1] M. B. P. Amato, C. S. Barbas, D. M. Medeiros, R. B. Magaldi, G. P. Schettino, G. Lorenzi-Filho, R. A. Kairalla, D. Deheinzelin, C. Morais, E. O. Fernandes, T. Y. Takagaki, and C. R. R. Carvalho, *Effect of a protective-ventilation strategy on mortality in the acute respiratory distress syndrome*, New England J. Med., 338 (1998), pp. 347–354.

[2] K. Astala and L. Päivärinta, *Calderón's inverse conductivity problem in the plane*, Ann. of Math. (2), 163 (2006), pp. 265–299.

[3] K. Astala, M. Lassas, and L. Päivärinta, *Calderón's inverse problem for anisotropic conductivity in the plane*, Comm. Partial Differential Equations, 30 (2005), pp. 207–224.

[4] J. A. Barceló, T. Barceló, and A. Ruiz, *Stability of the inverse conductivity problem in the plane for less regular conductivities*, J. Differential Equations, 173 (2001), pp. 231–270.

[5] L. Borcea, *Electrical impedance tomography*, Inverse Problems, 18 (2002), pp. 99–136.

[6] R. M. Brown and G. Uhlmann, *Uniqueness in the inverse conductivity problem for nonsmooth conductivities in two dimensions*, Comm. Partial Differential Equations, 22 (1997), pp. 1009–1027.

[7] A. P. Calderón, *On an inverse boundary value problem*, in Seminar on Numerical Analysis and its Applications to Continuum Physics, Sociedade Brasileira de Matemática, Rio de Janeiro, Brazil, 1980, pp. 65–73.

[8] M. Cheney, D. Isaacson, and J. C. Newell, *Electrical impedance tomography*, SIAM Rev., 41 (1999), pp. 85–101.

[9] W. Daily and A. L. Ramirez, *Electrical imaging of engineered hydraulic barriers*, Geophysics, 65 (2000), pp. 83–94.

[10] W. Daily, A. Ramirez, and R. Johnson, *Electrical impedance tomography of a perchloroethylene release*, J. Environ. Eng. Geophys., 2 (1998), pp. 189–201.

[11] M. R. Eggleston, R. J. Schwabe, D. Isaacson, and L. F. Coffin, *The application of electric current computed tomography to defect imaging in metals*, in Review of Progress in Quantitative NDE, Vol. 9A, Plenum Press, New York, 1990, pp. 455–462.

[12] D. G. Gisser, D. Isaacson, and J. C. Newell, *Electric current computed tomography and eigenvalues*, SIAM J. Appl. Math., 50 (1990), pp. 1623–1634.

[13] D. Isaacson and M. Cheney, *Effects of measurement precision and finite numbers of electrodes on linear impedance imaging algorithms*, SIAM J. Appl. Math., 51 (1989), pp. 1705–1731.

[14] D. Isaacson and E. L. Isaacson, *Comment on Calderón's paper: "On an inverse boundary value problem,"* Math. Comp., 52 (1989), pp. 553–559.

[15] D. Isaacson, J. L. Mueller, J. C. Newell, and S. Siltanen, *Reconstructions of chest phantoms by the D-bar method for electrical impedance tomography*, IEEE Trans. Med. Imaging, 23 (2004) pp. 821–828.

[16] D. Isaacson, J. L. Mueller, J. C. Newell, and S. Siltanen, *Imaging cardiac activity by the D-bar method for electrical impedance tomography*, Physiol. Meas., 27 (2006), pp. S43–S50.

[17] A. Kemna, A. Binley, A. Ramirez, and W. Daily, *Complex resistivity tomography for environmental applications*, Chem. Eng. J., 77 (2000), pp. 11–18.
[18] K. Knudsen, *On the Inverse Conductivity Problem,* Ph.D. thesis, Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark, 2002.
[19] K. Knudsen, *A new direct method for reconstructing isotropic conductivities in the plane,* Physiol. Meas., 24 (2003), pp. 391–401.
[20] K. Knudsen and A. Tamasan, *Reconstruction of less regular conductivities in the plane*, Comm. Partial Differential Equations, 29 (2004), pp. 361–381.
[21] K. Knudsen, J. L. Mueller, and S. Siltanen, *Numerical solution method for the D-bar equation in the plane*, J. Comput. Phys., 198 (2004), pp. 500–517.
[22] J. L. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications,* Vol. I, Springer-Verlag, New York, Heidelberg, 1972.
[23] L. Liu, *Stability Estimates for the Two-Dimensional Inverse Conductivity Problem*, Ph.D. thesis, University of Rochester, Rochester, NY, 1997.
[24] J. L. Mueller and S. Siltanen, *Direct reconstructions of conductivities from boundary measurements*, SIAM J. Sci. Comput., 24 (2003), pp. 1232–1266.
[25] A. I. Nachman, *Global uniqueness for a two-dimensional inverse boundary value problem*, Ann. of Math. (2), 143 (1996), pp. 71–96.
[26] A. I. Nachman, *Global Uniqueness for a Two-Dimensional Inverse Boundary Value Problem*, Preprint Series 19, Department of Mathematics, University of Rochester, Rochester, NY, 1993.
[27] A. L. Ramirez and W. Daily, *Electrical imaging at the large block test—Yucca Mountain, Nevada*, J. Appl. Geophys., 46 (2001), pp. 85–100.
[28] M. Reed and B. Simon, *Methods of Modern Mathematical Physics.* I: *Functional Analysis*, 2nd ed., Academic Press, New York, 1980.
[29] S. Siltanen, J. Mueller, and D. Isaacson, *An implementation of the reconstruction algorithm of A. Nachman for the* 2*-D inverse conductivity problem,* Inverse Problems, 16 (2000), pp. 681–699.
[30] L. Slater, A. M. Binley, W. Daily, and R. Johnson, *Cross-hole electrical imaging of a controlled saline tracer injection*, J. Appl. Geophys., 44 (2000), pp. 85–102.
[31] S. Stefanescu, C. Schlumberger, and M. Schlumberger, *Sur la distribution électrique potentielle autour d'une prise de terre ponctuelle dans un terrain à couches horizontales, homogènes et isotropes*, J. Physics and Radium Ser., 7 (1930), pp. 132–140.
[32] J. Sylvester, *A convergent layer stripping algorithm for the radially symmetric impedance tomography problem*, Comm. Partial Differential Equations, 17 (1992), pp. 1955–1994.
[33] M. E. Taylor, *Partial Differential Equations: Basic Theory*, Texts Appl. Math., 23, Springer-Verlag, New York, 1996.
[34] F. C. Trigo, R. Gonzalez-Lima, and M. B. P. Amato, *Electrical impedance tomography using the extended Kalman filter*, IEEE Trans. Biomed. Eng., 51 (2004), pp. 72–81.
[35] I. N. Vekua, *Generalized Analytic Functions*, Pergamon Press, London, Paris, Frankfurt, 1962.
[36] R. A. Williams and M. S. Beck, eds., *Process Tomography-Principles, Techniques, and Applications*, Butterworth-Heinemann, Oxford, UK, 1995.
[37] C. G. Xie, S. M. Huang, B. S. Hoyle, and M. S. Beck, *Tomographic imaging of industrial process equipment-development of system model and image reconstruction algorithm for capacitive tomography*, in Proceedings of the 5th Conference on Sensors and Their Applications, Edinburgh, Scotland, 1991, pp. 203–208.
[38] G. A. Uhlmann, *Developments in inverse problems since Calderón's foundational paper*, in Harmonic Analysis and Partial Differential Equations (Chicago, IL, 1996), University of Chicago Press, Chicago, IL, 1999.
[39] C. G. Xie, A. Plaskowski, and M. S. Beck, 8*-electrode capacitance system for two-component flow identification*, IEEE Proc. A, 136 (1989), pp. 173–190.

SIAM J. APPL. MATH.
Vol. 67, No. 4, pp. 915–937

MODELING CYCLIC WAVES OF CIRCULATING T CELLS IN AUTOIMMUNE DIABETES*

JOSEPH M. MAHAFFY† AND LEAH EDELSTEIN-KESHET‡

Abstract. Type 1 diabetes (T1D) is an autoimmune disease in which immune cells, notably T lymphocytes, target and kill the insulin-secreting pancreatic beta cells. Elevated blood-sugar levels and full-blown diabetes result once a large enough fraction of these beta cells has been destroyed. Recent investigation of T1D in animals, namely nonobese diabetic (NOD) mice, has revealed large cyclic fluctuations in the levels of T cells circulating in the blood, weeks before the onset of diabetes [J. D. Trudeau, C. Kelly-Smith, C. B. Verchere, J. F. Elliott, J. P. Dutz, D. T. Finegood, P. Santamaria, and R. Tan, *J. Clin. Invest.*, 111 (2003), pp. 217–223], but the mechanism for these oscillations is unclear. We here describe a mathematical model for the immune response that suggests a possible explanation for the cyclic pattern of behavior. We show that cycles similar to those observed experimentally can occur when activation of T cells is an increasing function of self-antigen level, whereas the production of memory cells declines with that level. Our model extends previous theoretical work on T-cell dynamics in T1D [A. F. M. Marée, P. Santamaria, and L. Edelstein-Keshet, *Int. Immunol.*, 18 (2006), pp. 1067–1077], and leads to interesting nonlinear dynamics, including Hopf and homoclinic bifurcations in biologically reasonable regimes of parameters. The model leads to the following explanation for cycles: High rates of beta-cell death, and corresponding elevation of self-antigen, shut off memory-cell production, leading to a gap in the population of activated T cells. Once peptide has been cleared by nonspecific mechanisms, the memory pool is renewed, and the cyclic behavior results.

Key words. autoimmune diabetes, type 1 diabetes, $CD8^+$ T cells, cycles, homoclinic bifurcation, mathematical model

AMS subject classifications. 92C30, 92C50, 70K05, 70K50, 70K44

DOI. 10.1137/060661144

1. Introduction.

Type 1 diabetes (T1D) is an autoimmune disease in which pancreatic beta cells are killed by the immune system, shutting off insulin secretion, and resulting in elevated blood glucose. The disease affects young people, severely impacting their health, and requiring perpetual insulin injection. Finding cures and/or treatment to replace the beta cells (e.g., by transplanting islets from organ donors) remains problematic, mainly because the damage is caused by the body's own immune system, which also attacks the transplant.

Studying autoimmune diabetes in humans presents ethical and clinical challenges. Therefore, animals with diabetic tendency, including *nonobese diabetic* (NOD) mice, are used to gain a basic scientific understanding of the disease. In NOD mice, T1D arises when populations of immune cells called T cells become primed to specifically target and kill beta cells. Such cytotoxic T cells belong to a class of lymphocytes displaying a surface marker called CD8. (Hence, they are denoted $CD8^+$ T cells.) We first briefly describe the background immunology and then present the detailed aspects specific to diabetes, the data on circulating T cells, and our model.

*Received by the editors May 28, 2006; accepted for publication (in revised form) November 17, 2006; published electronically April 24, 2007.

http://www.siam.org/journals/siap/67-4/66114.html

†Department of Mathematical Sciences, Nonlinear Dynamical Systems Group, Computational Sciences Research Center, San Diego State University, San Diego, CA 92182 (mahaffy@math.sdsu.edu).

‡Department of Mathematics and Institute of Applied Mathematics, University of British Columbia, 1984 Mathematics Road, V6T 1Z2, Vancouver, BC, Canada (keshet@math.ubc.ca). This author's research was funded by the Juvenile Diabetes Research Foundation and by the Mathematics of Information Technology and Complex Systems (MITACS), Canada.

1.1. Immunology primer. For an excellent survey of immunology, see [9]. T cells mature in the thymus, where those that cross-react with self-proteins are normally eliminated to prevent autoimmunity. After this period of development, they are released, circulate, and migrate to lymph nodes. In the lymph nodes, T cells interact with antigen-presenting cells (APCs) that display stimuli, consisting of a small fragment of antigen protein (i.e., a peptide of about nine amino acids in length) held inside a cleft of a larger protein (named major histocompatibility complex, or MHC, for historical reasons) [4]. The peptide-MHC complex (p-MHC for short) interacts with specific receptors on the surface of the T cells ("T-cell receptors," abbreviated TCRs). The strength, duration, and number of such interactions experienced by a given T cell determines its subsequent fate [24, 26, 15, 27, 21]. Within the right range of affinity to and quantity of p-MHC encountered, T cells with the appropriate specificity undergo activation, and the immune response is initiated.

Under normal conditions, APCs display antigens that are derived from foreign proteins such as viral or bacterial coat proteins. Then appropriately specific T cells are primed to form a large battalion of effector cells to combat the infection. Activated T cells proliferate, undergoing about six cell divisions. Their daughters are mostly effector cells (also called cytotoxic T lymphocytes, or CTLs), efficient and specific killers that seek out and destroy targeted cells. These effector cells, though deadly, are relatively short-lived [5]. A few daughters of activated T cells are memory cells that retain the same specificity but have no immediate effect [8, 25]. However, when the stimulus (e.g., the same foreign antigen) is encountered for a second time, memory cells can be activated rapidly to mount a faster immune response.

In autoimmune diseases such as T1D, the antigen peptide derives from normal proteins in the host. Infection or other injury can expose such proteins and initiate the disease, but once in progress, successive killing of targeted cells, and consequent release and exposure of more self-antigen, can sustain the inappropriate immune response. As the immune system is a complex web of nonlinear interactions between cells, chemicals, and tissues, rich dynamical behavior can be expected and indeed does occur. Our first goal in this paper is to point out interesting immunological dynamics to an audience of applied mathematicians. Our second goal is to present a plausible explanation of the cycles in autoimmune diabetes observed by [23], based on an established set of known and hypothesized interactions.

1.2. Autoimmunity in T1D. It has been shown that normal development of NOD mice includes a wave of programmed cell death (apoptosis) of pancreatic beta cells shortly after birth [18, 19]. In these same experiments, it was also determined that clearance of the apoptotic cells (by macrophages, nonspecific cells of the innate immune system) is reduced in NOD mice, leading to the conjecture that material from these dead beta cells forms self-antigen that triggers the autoimmune response. Previous modeling efforts have focused on such early initiation events [12, 13], but here we are mainly concerned with later stages in which the adaptive immune system is involved.

A number of proteins, including insulin, have been implicated as self-antigens in T1D. Most recently, experimental collaborators in Calgary (in the laboratory of P. Santamaria) have identified a new dominant self-antigen, IGRP (glucose-6-phosphatase catalytic subunit-related protein), a protein of beta cells whose normal function is yet to be determined. A fragment of this protein (consisting of amino acids 206–214) is the "peptide" to which most $CD8^+$ T cells in T1D react [10]. The discovery of this specific self-antigen in NOD mice followed years of experiments in which

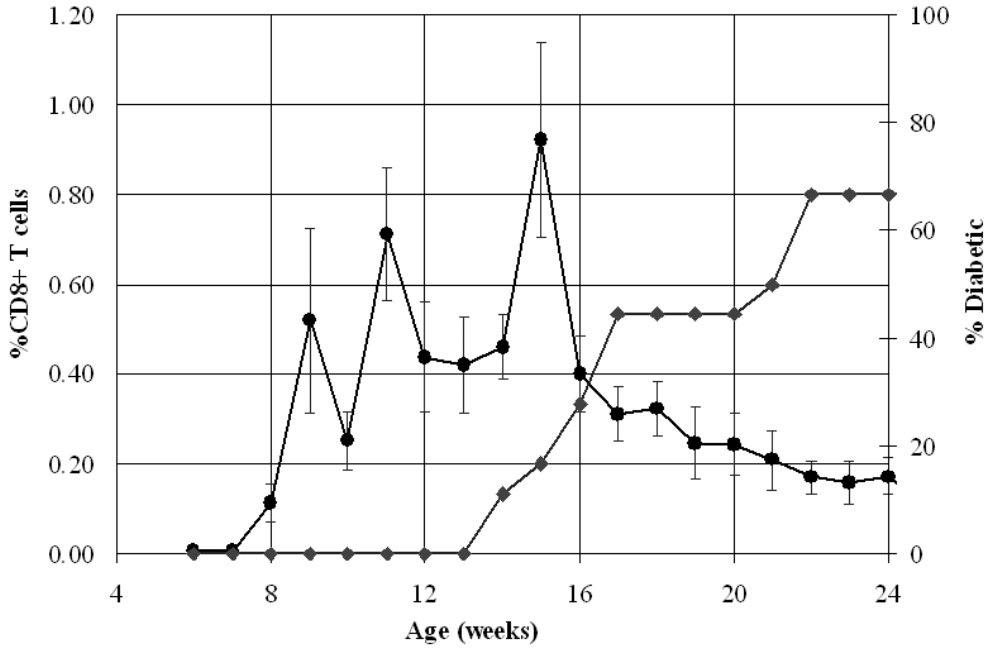


FIG. 1.1. *Periodic waves of circulating T cells occur in mice prone to diabetes (NOD mice) in the weeks before the onset of the disease. Data courtesy of the authors of* [23]. *Dark line, circles: T-cell level. Grey line, squares: percentage of the animals that became diabetic. Our model accounts for the cyclic waves but not for the period of initialization in weeks* 0–5, *the time when other processes prime the adaptive immune system.*

libraries of artificially synthesized peptides were used to identify and label T cells [2, 1, 6]. Use of tetramer probes (constructed of four copies of peptide-MHC with a fluorescent tag) allowed careful investigation of the levels and dynamics of these cells by enhancing the ability to label cells that were previously undetectable.

Using such tetramer staining experiments, it was shown by Trudeau et al. [23] that the level of autoreactive $CD8^+$ T cells is detectable in the pancreatic islets in 4–5-week-old NOD mice and at elevated levels by weeks 11–14. Correlated with this rise, populations of T cells circulating in the blood are also noticeably elevated over weeks 4–16 of age, before the high blood-sugar symptoms of diabetes occur. Surprisingly, the levels of these cells do not simply rise monotonically as the disease progresses but, rather, undergo dramatic fluctuations over this time frame, as shown in Figure 1.1.

Not all NOD mice develop diabetes, but the presence of these cyclic T-cell waves in a given animal predicts that it will become diabetic. Data for each one of the mice were aligned at the time of onset of high blood-sugar symptoms, so that the time axis could be "normalized" before combining and averaging. These pooled data show three peaks in the level of T cells starting at about 8 weeks of age and declining from about 16 weeks. The amplitude of the cycles increases over this time, and a slight increase in the period is also visible. The fact that Figure 1.1 was produced experimentally as an average of data for many mice suggests that there is some robustness in the cycling (as well as in its period) in NOD mice. These mice are all genetically identical, which means that parameters typical of their physiological and immunological processes are likely very similar (with some possible exceptions due to environmental effects).

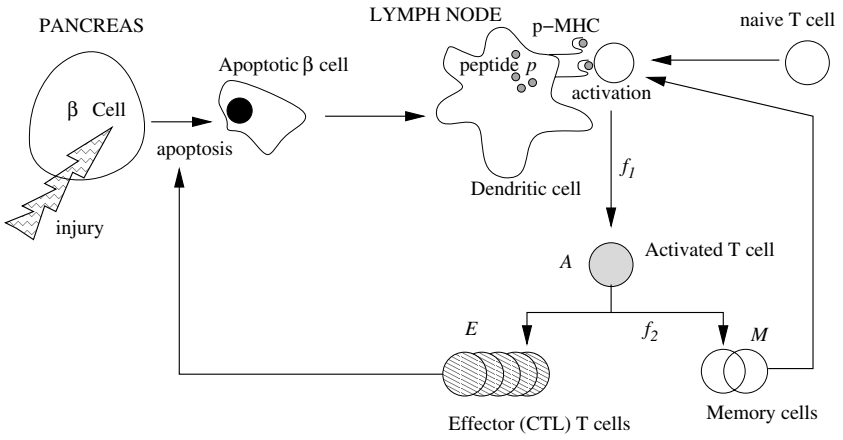


FIG. 2.1. *Scheme of the model. Programmed cell death (apoptosis) of pancreatic (insulin-producing) beta cells generates self-antigen peptide (p). In the pancreatic lymph nodes, this peptide is presented as part of cell-surface complexes (peptide-MHC, or p-MHC) on APCs called dendritic cells. The amount of p-MHC presented affects the activation and the differentiation of naive T cells into memory cells (for self-renewal) and into effector cells (CTLs) that seek and kill beta cells. This leads to more peptide exposure and results in positive feedback that eventually culminates in autoimmunity and T1D.*

Trudeau et al. speculated that each of these cycles represents "a round of proliferation of autoreactive T cells undergoing avidity maturation" [23], but the details of the underlying mechanism were not explored. This exploration is the subject of our paper.

2. Background for the model. Our main hypotheses stem from a recent model by Marée, Santamaria, and Edelstein-Keshet [14] that addressed the dynamics of T cells and peptide. In the latter paper, the focus was on artificial peptide used to treat the disease in a therapy similar to vaccination. It was shown that the competition of T-cell clones during peptide treatment could explain some of the puzzling dose-response behavior of the treatment and predict its success or failure. In their discussion, Marée, Santamaria, and Edelstein-Keshet [14] speculated that the increase in level of peptide antigen that results from beta-cell killing could be a feedback that explains the periodic waves of T cells observed by [23]. However, this idea has not yet been tested rigorously in a mathematical setting. We use some of the formalism and lessons learned in that model to investigate cyclic dynamics seen in [23]. We will show that an explanation for such dynamics is already inherent in the framework of the model of [14], or slight variations thereof.

Figure 2.1 summarizes the essential ingredients of our model. As shown, the process might be initiated by some injury or infection of beta cells, or by the normal wave of programmed cell death (apoptosis), not shown. Fragments of apoptotic cells are processed and presented as p-MHC on dendritic cells in the lymph nodes, and naive T cells interact with these complexes. It is known that the level of peptide presentation (i.e., amount of p-MHC) and the affinity of the T-cell receptors for the peptide determine whether a T cell encountering the APC will become activated to proliferate [4, 6, 16]. When naive T cells are activated, they proliferate to produce about 60 effector cells and about 1–4 memory cells [8]. Memory cells have a low turnover rate. They are able to undergo reactivation in response to antigen and to proliferate again, replenishing the pool of T cells. By killing beta cells, the effector T cells lead to a positive feedback on the amount of peptide produced and hence

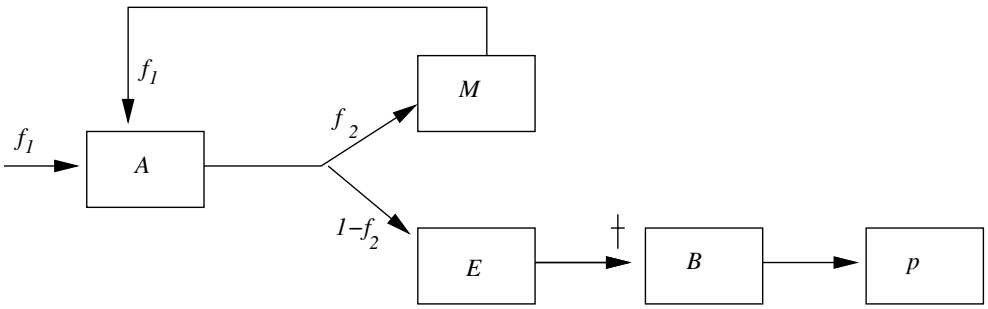


FIG. 2.2. *Simplified model scheme showing the main variables considered: A, E, and M are the number of activated, effector, and memory T cells. B denotes beta cells, and p is peptide. The two peptide-dependent functions are the fraction of T cells activated, f_1, and the fraction of memory cells produced, f_2. (The feedback from peptide to these has been omitted in the diagram for clarity.) The † represents the killing of beta cells by effector T cells.*

on further activation of T cells. The lifetime of the effector T cells is about 3 days [5, 7, 22] versus about 100 days for memory cells.

The level of peptide influences two important aspects of the process described above. First, the rate of activation of T cells depends on peptide level. Second, the fraction of daughter cells that are memory cells versus those that are effector cells is also peptide-dependent. Experimental evidence [11, 17] points to the fact that, at high-peptide doses, too few memory cells are produced. (This is termed "clonal exhaustion.") Following [14], we assume that the fraction of naive and memory T cells activated is given by a sigmoidal increasing function, $f_1(p)$, whereas the fraction, f_2, of daughter cells of activated T cells that become memory cells decreases sigmoidally as peptide increases. We also chose f_1 and f_2 to be Hill functions, i.e., rational functions with powers of degree > 1 (the degree is called the Hill coefficient; see section 3.2.)

In Figure 2.2, we show a simplified scheme, outlining our basic assumptions for the model: A fraction f_1 of incoming naive T cells becomes activated, (A); a fraction, f_2, of their offspring is memory cells, (M), and the rest, $1 - f_2$, are effector cells, (E). Memory cells can be reactivated (same peptide-dependent fraction, f_1, as incoming naive T cells). The effector cells cause the death of beta cells, (B), which, in turn, creates the peptide, (p). The peptide level affects both f_1 and f_2.

3. The model.

3.1. Assumptions. The following assumptions enter the model:

1. We do not consider the distinct compartments of blood, pancreas, and lymph nodes at this stage. Since the dynamics of interest take place over many weeks, whereas the trafficking between these compartments takes place on the time scale of hours, we approximate all variables as densities or concentrations in a single, well-mixed compartment.
2. We do not model the pathogenesis of the disease over the first 4–5 weeks. At this early stage, it is likely that the innate immune system (e.g., macrophages) may set up conditions that eventually give rise to the priming of T cells. See [13] for an analysis of that stage.
3. We assume that effector cells are terminal. (Some controversy exists about whether they give rise to some memory cells.) We also investigated a model in which memory cells are a progeny of effector cells and found essentially similar results.

4. We do not discuss the competition of many distinct "clones" of T cells for sites on ACPs or for p-MHC [14]. We model only the development of one dominant clone.
5. We assume that material from dead beta cells produces self-antigen peptide at a linear rate and that this peptide is presented proportionally as p-MHC on the dendritic cells. In [14], this p-MHC level was denoted m_t and modeled as a quantity in a quasi-steady state (QSS) with peptide and MHC molecules. Here we simplify such details.
6. We assume that once beta cells are gone, the production of the autoantigen ceases, and the immune response stops, since T-cell activation does not occur in the absence of peptide.

3.2. Model equations. Our full model consists of the following set of ordinary differential equations (ODEs):

$$\frac{dA}{dt} = (\sigma + \alpha M) f_1(p) - (\beta + \delta_A) A - \epsilon A^2, \tag{3.1}$$

$$\frac{dM}{dt} = \beta 2^{m_1} f_2(p) A - f_1(p) \alpha M - \delta_M M, \tag{3.2}$$

$$\frac{dE}{dt} = \beta 2^{m_2} (1 - f_2(p)) A - \delta_E E, \tag{3.3}$$

$$\frac{dp}{dt} = REB - \delta_p p, \tag{3.4}$$

$$\frac{dB}{dt} = -\kappa EB, \tag{3.5}$$

where $A(t), M(t)$, and $E(t)$ are the population levels of activated, memory, and effector T cells at time t, $p(t)$ is the peptide level, and $B(t)$ is the population of remaining beta cells. For the peptide-dependent functions, we take Hill functions,

$$f_1(p) = \frac{p^n}{k_1^n + p^n}, \tag{3.6}$$

$$f_2(p) = \frac{a {k_2}^m}{k_2^m + p^m}, \tag{3.7}$$

with $m, n > 1$. The parameters $k_1 > 0$ and $k_2 > 0$ in (3.6) and (3.7) denote typical levels of peptide at which the response of these functions is half-maximal, and $0 < a < 1$ is the maximal value of $f_2(p)$. Note that $f_1(p)$ is monotonic increasing whereas $f_2(p)$ is monotonic decreasing with p. In (3.1)–(3.3), all T cells represent members of clones whose specificity to beta-cell peptide is high. In (3.1), σ is the rate that naive T cells enter the circulation from the thymus. The fraction of incoming naive and memory cells that become activated is governed by the peptide-dependent sigmoidal function, $f_1(p)$ (α is a factor that represents the higher rate of activation of memory cells relative to naive cells). The rate of decay of A, δ_A, is augmented by a term for competition, ϵA^2, as discussed in [14]. Activated cells progress to a differentiated stage at rate β. They then proliferate by a series of cell doublings to produce $2^{m_2} \approx 60$ effector cells and $2^{m_1} \approx 3$–4 memory cells. The commitment to development into these two types of daughter cells depends on peptide according to the decreasing sigmoidal function $f_2(p)$. Effector cells are terminal and have a shorter half-life than memory cells ($\delta_M < \delta_E$).

Equation (3.4) depicts our simple assumption about production and clearance of peptide: the level of "peptide," p, is produced with mass-action kinetics when effector cells kill beta cells (at rate R per effector per beta cell) and cleared with linear kinetics at rate δ_p. Recall that clearance of dead beta cells and their fragments by macrophages is defective in NOD mice [12, 18, 19], and this defect can theoretically lead to the early chronic inflammation that initiates the priming of T cells [13]. Therefore, it is of interest to ask whether this same defect can also account partly for the dynamics of T cells at this later stage of the disease. We investigate this later.

We use the simplest possible model for decay of beta cells due to killing by effector T cells in (3.5). The parameter κ denotes the rate of killing per effector cell. We ignore the (limited) ability of beta cells to regenerate and the very slow aging and turnover rate of beta cells in the healthy individual. Currently, the extent to which beta cells can self-renew after immune attack is still under investigation, and this process is likely to occur on a slow time scale. For this reason, we did not explicitly include this in the model at this stage.

3.3. Model equations for a reduced QSS system. Our analysis begins with a reduction of the full system of equations (3.1)–(3.5) to a simpler model using separation of time scales. First, we argue that the time scale of peptide dynamics—hours—is faster than any of the time scales of cell dynamics—days and weeks—justifying a QSS assumption on the peptide. Hence, we set $dp/dt = 0$ in the model, so that $p = (RB/\delta_p)E$.

The model then consists of (3.1), (3.2), (3.3), and (3.5). The functions f_1, f_2 now depend on E and B via the QSS peptide expression. We refer to this as the *reduced QSS model*. Our first step was to explore this model computationally. To do so, we had to estimate parameters and consider appropriate scaling. Our steps and results are described below.

3.4. Parameter estimates, scaling arguments, and computations. Based on nonlinearities (in the functions f_1, f_2), the model consisting of (3.1)–(3.5) can have a range of interesting behaviors. As we are interested in the possible biological and medical applications of this model, it is essential to study its behavior within a biologically reasonable range of parameter values. Almost all parameters in the model were based on experimental information previously compiled by Marée, Santamaria, and Edelstein-Keshet [14]. Some exceptions include parameters associated with beta-cell killing and peptide production, as these were not considered in the previous treatment. To avoid lengthy diversion into the details, we concentrate all details of the parameter estimates in the appendices. The meanings, units, and values of the parameters are presented in Table B.1. The level of cells of type A, E, M varies on a range of several orders of magnitude. As we wanted to present these all on the same plot, we scaled these population densities by the appropriate powers of ten. Scaling arguments are also given in the appendices. We left the time variable in units of days to emphasize the period and timing of the cycles that we obtained.

Simulations of the dynamics were carried out in MATLAB. Initial conditions were chosen to depict some (preexisting) stimulus to the immune system stemming from earlier stages of the disease (e.g., as speculated in [13]). Bifurcation diagrams were composed with the AUTO feature of XPP, freely available software written by G. Bard Ermentrout.[1] Unless otherwise indicated, all simulations use the basic core set of parameter values, as shown in Table B.1.

[1] www.math.pitt.edu/∼bard/xpp/xpp.html.

4. Results. Starting any simulation with the healthy state as initial condition, i.e., $A = M = E = 0$, $B = 1$ (and thus also $p = 0$), clearly results in continued health, since this point is a steady state of the system. Moreover, the stability of this equilibrium implies that even some (sufficiently small) perturbation rapidly returns to this state. Hence to get any immune dynamics of interest in our model, the system should be initiated with some T cells already "primed." Typically, we start simulations with $A = 0.5, M = 0$, and $E = 1$. This state ensures that effector cells are present to lead to peptide production and that activated T cells are available to renew that pool of effectors. Other initiation values are possible, depending on parameter settings (discussed later). This prototypical set of values represents the outcome of earlier events that our model is not describing (but see, e.g., [13] for a possible description.)

Not all NOD mice develop diabetes. Therefore, any model for this disease also has to account for the fact that some initial stimuli will be resolved without full-blown autoimmunity. We first discuss this baseline control for the model. Running the reduced QSS model from an initially "primed" state with default parameter values gives rise to the behavior shown in Figure 4.1; that is, an initial elevated level of effector and memory T cells is resolved, after some time, and the immune response ceases. This corresponds to resolution of the immune attack with no autoimmunity even though the immune system has been provoked to respond. The beta-cell population decreases by 40% during the immune attack. Since our model does not address replenishment of the beta cells by reproduction or stem-cell differentiation, the beta-cell mass remains constant after this isolated immune response.

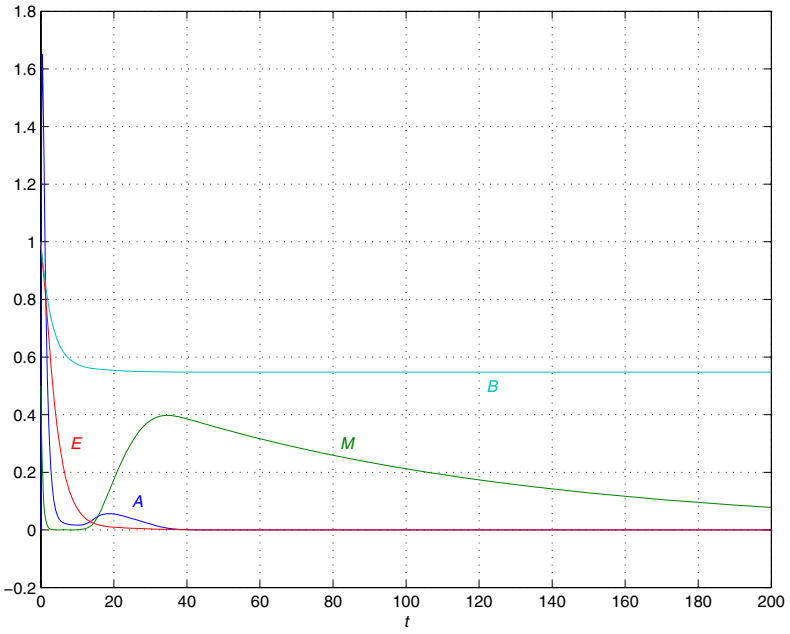


FIG. 4.1. *(color online). Simulation of the model for NOD mice that* do not *become diabetic. Number of circulating cells (scaled) versus time (days). Dark blue: A* ($\times 10^3$ *cells), green: M* ($\times 10^4$ *cells), red: E* ($\times 10^6$ *cells), light blue: B (fraction of beta-cell mass remaining). Simulation uses default ("NOD") parameter values given in Tables B.2 and B.1. For the initial conditions* $A = 0, M = 0.5, E = 1, B = 1$*, the immune response is resolved without chronic disease or cyclic waves.*

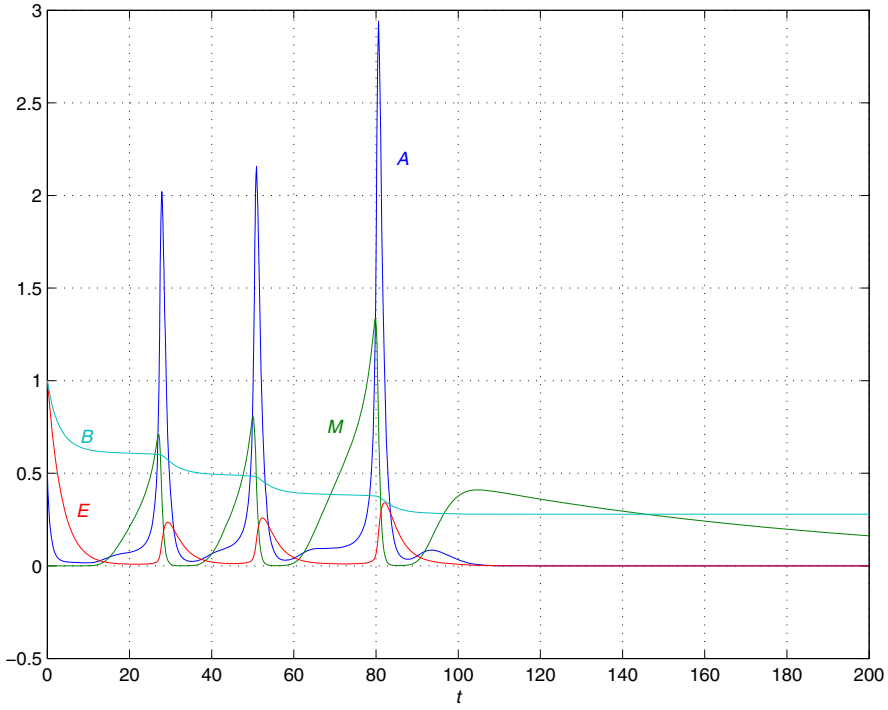


FIG. 4.2. *(color online). Simulation of the model for NOD mice that* do *become diabetic (by* 80–90 *days of age). Default ("NOD") parameter values and scaling are as in Figure* 4.1 *but with initial conditions* $A = 0.5, M = 0, E = 1, B = 1$ *that evoke the elevated periodic immune response. Dark blue: A, green: M, red: E, light blue: B. The disease progresses with cycles of T cells that cause waves of beta-cell killing, as predicted by the model.*

When the initial conditions include more elevated levels of activated T cells (with all other parameters left as is), oscillations can appear, as shown in Figure 4.2. As in Trudeau et al. [23], three peaks with increasing amplitude of effector T cells occur over days 30–80 at the period of approximately 3–4 weeks, as in the experimental data. This run is in close agreement with the data for mice that develop full-blown diabetes, as shown in Figure 1.1.

We can understand intuitively how such cycles occur by reasoning as follows: In our model, (3.5) leads to the decay of beta cells whenever effector cells are present. Due to the assault on beta cells by the T cells (specified by our choice of initial conditions), peptide level increases, T cells are activated, and effector cells are formed. However, once peptide rises to a high level, memory-cell production is turned off (as f_2 decreases with p). Thus, replenishment of activated T cells drops, and subsequently E also declines and is not renewed. Once the effector cells decline, new peptide is hardly produced. It is gradually cleared and eventually reaches a low level that is then consistent with memory-cell production. This then stimulates production of new activated T cells, and the cycle repeats. Periodic peaks and troughs continue until beta cells are depleted, and then no more peptide is formed, and T cell activation stops altogether. At this stage, since beta cells are gone, full-blown diabetes sets in, and the immune response decays to its trivial equilibrium. This reasoning is plausible but relies on an appropriate combination of parameters governing rates of depletion and renewal of the various cell types.

It is noteworthy that merely by increasing the rate of clearance of the peptide, δ_p, we end the tendency of the system to cycle. We ran simulations with elevated values of δ_p and found behavior similar to that of Figure 4.1 for much broader ranges of initial conditions (results not shown). These results can be taken as indications that in "control" mice whose peptide clearance rate is normal, immune response is less likely to lead to prolonged cycling attack. These results are discussed in more detail later.

5. Analysis of a reduced model with B as a parameter. To gain a clearer understanding of the behavior described above, we reduce the four-dimensional model simulated above further yet by considering the level of beta cells, B, to be a parameter. The onset of diabetes in NOD mice requires about 16 weeks, at which time there are very few remaining beta cells in the pancreas. This indicates that the variable B in the full model acts more like a slowly varying parameter compared to the other variables in the model. We therefore consider a reduction to three variables (A, M, E) and analyze the model behavior. We then discuss how the gradual decrease of B influences the dynamics of the whole system. The model to be analyzed now consists of the three equations

$$\frac{dA}{dt} = (\sigma + \alpha M) f_1(p) - (\beta + \delta_A) A - \epsilon A^2, \tag{5.1}$$

$$\frac{dM}{dt} = \beta 2^{m_1} f_2(p) A - f_1(p) \alpha M - \delta_M M, \tag{5.2}$$

$$\frac{dE}{dt} = \beta 2^{m_2} (1 - f_2(p)) A - \delta_E E, \tag{5.3}$$

together with (3.6) and (3.7) and the QSS peptide expression

$$p \approx (RB/\delta_p) E. \tag{5.4}$$

This three-dimensional system of differential equations permits a more complete analysis.

5.1. Steady states and stability properties. The three-dimensional system of differential equations given by (5.1)–(5.3) has several types of feedback. Peptide level (and therefore effector cell level) leads to positive feedback on T-cell activation via f_1. Simultaneously, these levels produce negative feedback on the memory-cell production via f_2. When combined, these nonlinear feedbacks lead to the possibility of multiple steady states, depending on the parameters. Numerical experiments suggest that this mixed feedback system can have from one up to five equilibria.

In the biologically relevant regime of parameters (discussed in the appendices), we find that there are three equilibria. One of these is clearly the trivial equilibrium $A = M = E = 0$. This follows immediately from the fact that $f_1(0) = 0$. This equilibrium corresponds to a disease-free state and is easily shown to be a stable node. The fact that the origin is an attractor means that a small disturbance that provokes the immune system should be resolved, provided it is sufficiently weak.

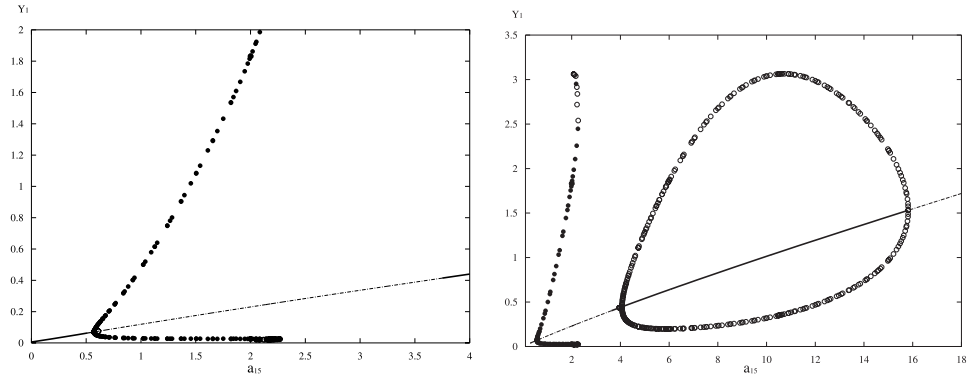
There also exists a positive equilibrium that corresponds to a state of elevated immune cell levels. In that state, effector T cells are continuously killing beta cells, and this corresponds to an autoimmune attack that eventually leads to diabetes. This equilibrium has various stability properties that depend on the parameters. We discuss this in more detail below. A third equilibrium is a saddle with a two-dimensional

stable manifold, which for some parameters separates the "healthy" and diseased equilibria. For these parameters, stimuli that fall on the wrong side of this separatrix will be attracted to the diseased equilibrium. For other parameter values, the unstable manifold of the diseased state connects to the stable manifold of the saddle point. In this case, almost all positive initial conditions asymptotically approach the "healthy" state.

As a specific example of the local analysis, we considered the system of equations (5.1)–(5.3) with the parameters given in Table B.1 and $B = 1$. Due to the nonlinearities in the functions f_1 and f_2, it is not possible to solve explicitly for equilibria. Therefore, we determined steady states, eigenvalues, and eigenvectors numerically using the software program Maple. We found the following results: The disease-free equilibrium, $(\bar{A}_0, \bar{M}_0, \bar{E}_0) = (0, 0, 0)$, is a stable node with the three eigenvalues $\lambda = -1, -0.3, -0.01$. A saddle node at $(\bar{A}_s, \bar{M}_s, \bar{E}_s) = (0.0116, 0.696, 0.00116)$ has a two-dimensional stable manifold (eigenvalues $\lambda_1 = -1.52, \lambda_2 = -0.0188$ and associated eigenvectors $v_1 = [1, 0.495, 0.0245]$, $v_2 = [1, -68.5, 0.107]$) and an unstable manifold (eigenvalue $\lambda_3 = 0.210$ with eigenvector $v_3 = [1, 2.62, 0.0589]$). Finally, the diseased equilibrium, $(\bar{A}_d, \bar{M}_d, \bar{E}_d) = (0.119, 0.0141, 0.0356)$, has a stable manifold (with eigenvalue $\lambda_1 = -2.37$ and associated eigenvector $v_1 = [1, -0.108, -0.0414]$). It also has a two-dimensional unstable manifold (eigenvalues $\lambda = 0.0129 \pm 0.553i$) that spirals outward toward a limit cycle. From this local analysis, we could see that at each equilibrium, one eigenvalue is significantly more negative than the others. This suggests that there is a globally attracting two-dimensional manifold containing the three equilibria, where the interesting dynamic behavior occurs.

5.2. Bifurcations. We first discuss bifurcations with respect to a relevant parameter and later assemble the sequence of dynamical behaviors in Figure 6.2. In the model given by (5.1)–(5.3), we have assumed that the destruction of beta cells occurs on a slow time scale. Thus, the level of beta cells, B, makes a natural bifurcation parameter to consider. At the beginning of our simulations, we normalize $B = 1$ and set $\delta_p = 1$. By the QSS assumption for peptide, a gradual loss of beta cells in this model variant is dynamically equivalent to a gradual increase in the peptide clearance rate δ_p. (Both parameter variations essentially describe the decreasing QSS value, $p = (RB/\delta_p)E$.) We explored this parameter variation using the AUTO option of the software XPP. Figure 5.1 shows the result obtained thereby.

The diagram given in Figure 5.1(a) shows the basic bifurcation behavior of the model (and uses the default parameter values given in Tables B.2 and B.1). Moving across this diagram from left to right along the horizontal axis represents increasing values of the peptide decay rate δ_p or, equivalently, a decreasing level of beta cells, B. Close to the leftmost edge (high B, or low peptide clearance rate), we find a stable diseased state (solid line with shallow slope). The "healthy" state, also stable, and the saddle node are not indicated in the diagram. Moving towards the right leads to a supercritical Hopf bifurcation at $a_{15} = \delta_p = 0.571$, spawning a stable limit cycle. Here we enter the regime of cyclic behavior evidenced in Figure 4.2. The diseased equilibrium is then an unstable spiral, as predicted by the local analysis described above. The limit cycle persists, and its amplitude increases as the parameter increases (respectively, as the beta-cell level decreases) up to a homoclinic bifurcation at $\delta_p = 2.268$ (equivalently at $B = 0.441$, i.e., when only about 44% of beta-cell mass remains). As seen in our runs and in the upper branch of this bifurcation line on the zoomed out diagram of Figure 5.1(b), AUTO has difficulty resolving this global bifurcation. We discuss the nature of this dynamical shift later.



(a) (b)

FIG. 5.1. *Bifurcation diagram for the peptide decay rate, $a_{15} = \delta_p$, with all other parameters set at their default values, as in Tables B.2 and B.1. The vertical axis is A in units of 10^3 cells.* (a) *A portion of the diagram, enlarged, showing the typical bifurcation: a Hopf bifurcation occurs at $a_{15} = 0.5707$, spawning a stable limit cycle. A homoclinic bifurcation occurs at $a_{15} = 2.268$.* (b) *Further bifurcations on an expanded scale: another Hopf bifurcation (to an unstable limit cycle) occurs at $a_{15} = 4.063$. This limit cycle vanishes at $a_{15} = 20.28$.*

Following the homoclinic bifurcation, the diseased state remains unstable, and the origin is the only global attractor for some range of the bifurcation parameter. Interpreting this bifurcation diagram in terms of normal and reduced levels of (peptide) clearance rates (by control versus NOD macrophages) suggests why the clearance defect itself could make the difference between healthy (control) mice versus diabetes-prone (NOD) mice: for example, as seen in Figure 5.1(a), a "control" peptide clearance rate of $\delta_p = 3$ per day leads to dynamics that always resolve any initial stimulus (returning to the baseline where no immune cells persist, since the limit cycle does not occur, and the disease state is unstable), whereas a factor of two decrease to $\delta_p = 1.5$ per day (representing reduced clearance in NOD mice) puts the same system into the regime of cyclic T-cell waves and autoimmunity.

Reinterpreting this diagram in terms of the gradual decrease of beta-cell mass (from left to right starting from $B = 1$) explains the following features shared by the data of Figure 1.1 and the simulation of Figure 4.2: (1) the increase in the amplitude of the cycles, (2) the fact that the cyclic behavior stops abruptly (e.g., around days 80–90 in the simulation of Figure 4.2) when the homoclinic bifurcation occurs, and (3) the slight lengthening of the period just before this transition. It also explains why (4) the immune cells then decay to the baseline state $A = M = E = 0$. Thus, the bifurcation diagram can help to provide a plausible scenario for a mechanism underlying these dynamics.

As previously noted, immunological systems present a menagerie of curious dynamical behaviors that can be an enticing invitation to the applied mathematician. As our model is nonlinear, other interesting behavior is to be anticipated. In Figure 5.1(b), we show an expanded scale, with much higher values of the peptide turnover parameter. As seen here, at $\delta_p = 4.063$ per day, a second subcritical Hopf bifurcation takes place. Thus, for a range of values of $4.063 < \delta_p < 20.28$ per day, the diseased state becomes (locally) stable once more, with a domain of attraction bounded by an unstable limit cycle. All solutions inside this domain will evolve towards the diseased state, whereas outside this domain of attraction, solutions eventually lead to the origin. Aside from purely mathematical interest, this diagram suggests

that there are as yet other unexplored behaviors in this and other immunological models. On one hand, biologically, this result could be interpreted to mean that increased removal of peptide is not always advantageous (since it can reinstate the stability of the diseased state). On the other hand, the dynamics shown in this expanded parameter regime might be more of a mathematical curiosity than a result that is directly relevant to diabetes in NOD mice.

We investigated a number of other parameter variations and bifurcations (diagrams omitted), starting from the default parameter set. For example, we varied the parameter $a = a_4 < 1$ of the function f_2. This parameter specifies the maximal fraction of memory cells produced (when $p = 0$). We found that decreasing a from 1 leads to the homoclinic bifurcation at $a = 0.45$. Similarly, for the T-cell competition parameter, the range $0 \leq \epsilon \leq 2.17$ lies within the stable limit cycle regime. A Hopf bifurcation occurs at $\epsilon = 2.173$, leading to stability of the diseased state. No homoclinic bifurcation was obtained by varying this parameter. Finally, changing k_2, the peptide level that corresponds to the half-maximal value of f_2, gave a stable diseased state when $k_2 = 2$, a Hopf bifurcation at $k_2 = 1.112$, and a homoclinic bifurcation when $k_2 = 0.825$. Due to space constraints, these bifurcation diagrams are not shown.

6. Geometry of the solutions. Figure 6.1 shows two stereograms of the three-dimensional AME system in the regime of parameters consistent with the stability

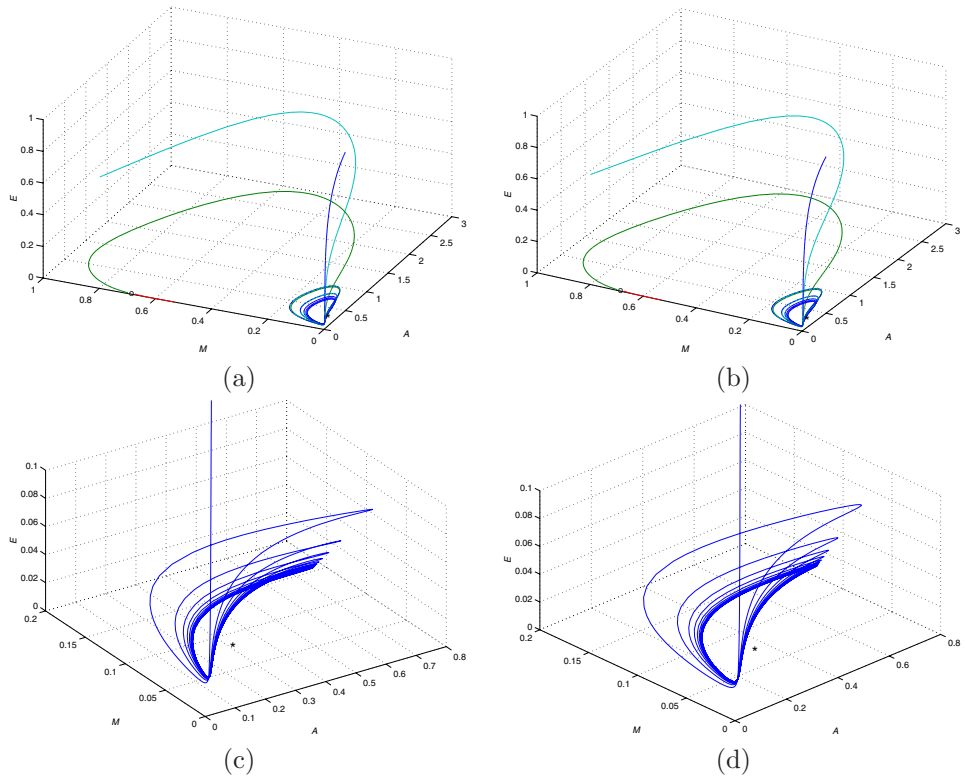


FIG. 6.1. *(color online). Two stereograms showing in* (a) *and* (b) *the limit cycle and location of the saddle point (marked ∘) and in* (c) *and* (d) *the limit cycle and its unstable diseased equilibrium (denoted by ∗). Both diagrams were made for the basic model with default parameter values but with beta-cell mass treated as a (constant) parameter.*

of the limit cycle oscillations. In (a) and (b), we show the position of the saddle node (close to the M-axis) with unstable manifolds in green and red. One branch of the unstable manifold (in red) flows towards the stable disease-free state, while the other branch (in green) spirals towards the limit cycle about the disease state. (The two-dimensional stable manifold is not indicated in this figure.) The limit cycle and two trajectories attracted to it are also shown. In (c) and (d), a zoomed-in view of the limit cycle is shown. The location of the (unstable spiral equilibrium) diseased state is indicated by a small star.

Figure 6.2(a)–(d) shows a sequence of diagrams that illustrates the bifurcations and dynamics described in the previous section. We show two-dimensional "cartoons" that give the overall picture (although our AME system is three-dimensional), since it is difficult to numerically simulate the precise parameter set that leads to the homoclinic connection and equally challenging to represent all stable and unstable manifolds in a three-dimensional plot. As shown in this figure, the origin (heavy dot labeled H for "healthy") retains its stability and is a local attractor in all cases, but its basin of attraction can vary greatly. In (a) and (b), a separatrix (one branch of the stable manifold of the saddle node S) defines the boundary between those states attracted to H and others that remain in the positive orthant. In (a), these other

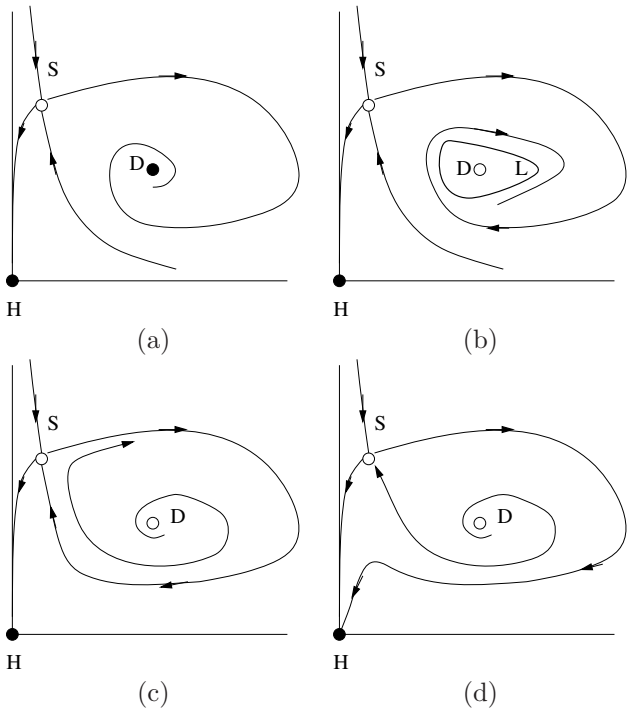


FIG. 6.2. *This sequence of four sketches illustrates the essential geometry of the dynamics and bifurcations. H: "healthy" state in which there are no circulating immune cells, D: diseased equilibrium; S: saddle node, L: stable limit cycle. Heavy dots indicate stable equilibria, and open dots indicate unstable ones. In* (a)*, an unstable manifold of S winds into the stable spiral at D. In* (b)*, just past a Hopf bifurcation, there is a stable limit cycle to which this manifold is attracted.* (c) *represents the homoclinic bifurcation. In* (d)*, the unstable manifold of S makes a detour past the unstable D, ending at H. The state H is always stable. However, the boundary of its basin of attraction is formed by the stable manifold of S. In* (d)*, every initial condition will eventually evolve towards the origin.*

points are attracted to the stable diseased state (heavy dot at D), whereas in (b), past a Hopf bifurcation, the limit cycle is attracting. A homoclinic connection (which exists for one specific set of values of the parameters) is illustrated in (c). In (d), states close to the unstable point D may take an "excursion" towards S but eventually arrive at H. In this case, all solutions of (5.1)–(5.3), except for a set of measure zero (on the stable manifold of the saddle node), would eventually converge to the disease-free state $(\bar{A}, \bar{M}, \bar{E}) = (0, 0, 0)$.

7. Parameter sensitivity. The hallmark of autoimmune diabetes in NOD mice is that many small perturbations and treatments can "cure" the disease, delay its onset, or prevent it from occurring. Thus, the actual (biological) system is sensitive to relatively small changes in essential parameters of the system. In order to explore the sensitivity of the model, we tested how increases and decreases in each of the parameters in (3.1)–(3.5) affect the dynamics. We used the values of parameters that generated Figure 4.2 as a basic set and varied each in turn by 10% up and down. The results are shown in Table B.2.

Recall that the original parameter set is consistent with a stable limit cycle for $B = 1$. In Table B.2, we note whether the dynamics obtained by a given parameter change has moved the system in the direction of the homoclinic ($\rightarrow$) or the Hopf bifurcation ($\leftarrow$) or, in some cases, beyond those bifurcations. (Arrows are indications of shifts along the type of bifurcation shown in Figure 5.1(a).) We also indicate the number of peaks observed between $t = 0$ and the time at which the homoclinic bifurcation occurs. It can be seen that changing some parameters (e.g., n, m, k_2, k_1 of the peptide-dependent response functions f_1, f_2) has a large effect on the number of cycles that occur, increasing the number of peaks up to 9–10. These parameters control the location of the "activation switch" and the switch in commitment to memory versus effector cells with respect to peptide level. The parameter β and the number of memory cells produced, 2^{m_1}, also has a dramatic effect on the behavior. Other variations, e.g., δ_M, a, ϵ, have a very minor effect.

It is interesting to note that certain slight parameter shifts place the system beyond the homoclinic bifurcation, leading to global stability of the origin (as in Figure 6.2(d)). This includes a 10% decrease in the rate of memory-cell reactivation, α, or memory-cell production, a, or a 10% increase in the peptide clearance rate, δ_p, or the effector T cell death rate, δ_E (entries in Table B.2 marked with S, $\rightarrow$). Making these adjustments takes the system out of the cyclic regime and restores global stability of the "healthy" state at the origin.

Here we venture to speculate on implications to the disease itself and possible treatments. One can envision medical interventions that are designed to affect one or another of the parameters mentioned above in patients with a known genetic tendency to autoimmunity. If any of these parameter change(s) could be made before beta-cell mass is destroyed, the immune attack could be resolved or prevented. Alternately, if cycles of circulating T cells are observed, treatments could be applied to knock the system out of its destructive cyclic regime, back to the baseline state. The most effective treatment would be one that targets any of the more sensitive parameters in our model. Because our model is fairly simplistic, it is premature to draw firm conclusions about optimal therapeutic strategies. However, studying parameter sensitivity and bifurcations of more detailed and more realistic models for this disease (or other autoimmune disorders) could possibly lead to new therapeutic strategies. Clearly, in the context of a mathematical model, one can also identify and possibly avoid unforeseen complications (e.g., the unstable limit cycle regime in Figure 5.1(b),

where the disease state regains stability in another range of the parameter(s)).

8. Other variants of the model. We considered several variants of the model that incorporated other features or relaxed certain assumptions. First, we considered a model in which memory cells are offspring (rather than sisters) of effector cells. (In that model, a function like f_2 represented the probability that an effector cell differentiates into a memory cell.) Similar behavior was obtained in a narrower range of parameters. As this scheme of differentiation is less widely accepted, here we omit the details.

The immune response has several inherent delays. After beta cells die, it takes around 8 hours to 1 day for their fragments to be collected, transported to the pancreatic lymph node, processed, and presented by APCs. Once T cells are activated, it take a further 2–3 days for proliferation and production of effector cells. This means that an immune response can take 4–6 days from the time of stimulus. We explored some of the effects of delay in the system by investigating variants of the model that had one or two delays. We found similar dynamics within a slightly shifted set of parameter regimes. Results were similar to figures previously displayed and here are omitted.

We briefly explored competition of various T-cell clones to determine how competition between different peptide-dependent cells could affect the dynamics. We found that similar clones tend to cycle together and that competition was not a major force in the cyclic dynamics. The details are omitted.

9. Experimental tests of the model. This model has been informed by previous theory [13, 14], supplemented by experimental observations. In turn, it suggests new experiments that can be used to verify or refute its conclusions.

First, the model predicts a specific sequence of events, with peaks in memory cells preceding peaks in activated T cells and preceding peaks in effector cells (as shown in Figure 4.2). Further, the model predicts that during these cycles, one should be able to observe cycles of apoptotic beta cells in the pancreas (since killing by effector cells occurs via apoptosis). If the presence or sequence of cell types follows some other trend, our model would have to be revised.

Second, the model predicts outcomes of specific interventions. For example, once NOD T-cell cycles are observed, poisoning some fraction of their macrophages by administering silica, a known poison for such cells (i.e., reducing the innate ability to clear dead beta-cell material and hence reducing δ_p) should decrease the amplitude of the cycles as well as the period of the cycles (see parameter sensitivity, Table B.2). A dose response of this "macrophage poison" versus dynamical behavior would show successively decreased cycle amplitude (see also Figure 5.1(a) for the dependence of cycle amplitude on $a_{15} = \delta_p$). Alternately, treatments that enhance macrophage clearance of apoptotic material (if possible) could, at sufficient dose, stop the cycles and retard the development of the disease. A number of similar interventions are predicted by parameter sensitivity. While we cannot expect that we have captured all NOD parameters accurately in this preliminary model, general trends "towards" or "away from" the Hopf or homoclinic bifurcation predicted by the various changes in basic parameters should be indicative of the accuracy or fallibility of the assumptions on which the model is based. Some (but not all of these) are experimentally feasible. Future work with experimental colleagues will address such issues in an experimental setting.

10. Discussion. Our main conclusion in this paper is that cyclic dynamics can arise spontaneously in the immune response leading up to T1D, at least under conditions typical of the susceptible (NOD) mice. This fact was conjectured in [14] as a possible outcome of the interplay between the effector T cells killing the insulin-producing beta cells and the feedback from self-antigen produced when those cells are killed. We confirmed this conclusion by extending the model in [14] to include the death of beta cells and the accumulation of the antigen that results. Our cyclic dynamics (Figure 4.2) are similar to the experimentally observed cycles (Figure 1.1) in three important ways: (1) it shows cycles of increasing amplitudes; (2) the interpeak time length increases slightly; and (3) the cycles stop, and the levels of T cells drop around 16–18 weeks. (At this point, the mouse becomes diabetic in the experimental system.) This behavior was obtained in a regime of parameters that is based mainly on values assembled from the experimental literature in [14].

We showed that one explanation for these oscillations, illustrated by our model, is as follows: beta-cell killing produces large quantities of self-antigen peptide, expanding the population of effector cells at the expense of memory cells. This creates a gap in self-renewal of the T cells that leads to a pause in their reproduction and reduced effector levels for killing. After a suitable interval, when peptide is cleared, the memory-cell production is reinstated, and the cycle begins once more. The gradual loss of beta-cell mass limits the number of cycles that can occur (to three in the case of NOD mice). The cyclic dynamics are found for a wide range of parameter values, provided the peptide-dependent functions that control T-cell activation, f_1, and memory-cell production, f_2, ramp up (respectively, down) as peptide level increases. Since the immune system is highly complex, with many feedbacks between cells, chemicals, and tissues, it is possible that other explanations for cycles can be equally compelling. For example, recent work by an experimental collaborator (P. Santamaria, U. Calgary) has focused on the role of regulatory T cells and their cytokine IL-2. Positive and negative feedback that is emerging in these experimental investigations will provide future opportunities for modeling and analysis using the tools of nonlinear dynamics.

The main contribution of our study is to explain the mechanism underlying the observed cycles by studying the nonlinear dynamics of the extended model and uncovering its bifurcations. This aspect of our work is particularly apt for readers of SIAM, some of whom may not yet be aware of rich dynamics in immunology. We showed that all three of the observations listed above can be explained as the gradual shift of a parameter (the mass of beta cells remaining) during the course of the immune attack by effector T cells. This gradual shift moves the system from a regime in which there is a stable limit cycle towards a homoclinic bifurcation. The amplitude of the limit cycle expands very quickly just before this bifurcation, and its period increases (theoretically up to infinity) at the homoclinic connection itself. Beyond that, all points are attracted towards the origin; i.e., the levels of T cells drop.

We found that the number of peaks that occur in the model shifts when certain parameters are changed (by 10%). The most sensitive parameters are those appearing in the functions f_1 and f_2, but it is unlikely that these are easy to manipulate in an experimental system. As our model is the simplest possible variant of [14] that produces cycles, this sensitivity to parameters may be a price paid for omitting other regulatory features of the immune system. On the other hand, the sensitivity to parameters also suggests numerous experimental tests of our predictions that could, in principle, validate or falsify the model. Such tests will be under consideration in

future work on this problem. In the original model of [14], competition between various clones of T cells for sites on APCs was considered as an important determinant of dynamics. Here, in preliminary investigations of competition of two clones, we found that clones tended to cycle synchronously. Future work will also address the effect of competition in greater detail.

Our model did not address any of the spatial or compartmental aspects of the immune response. For example, we also did not consider the details of trafficking of T cells between blood, lymph nodes, and tissue. Some of the detailed movement and interactions of T cells with dendritic cells in lymph nodes is currently being modeled in conjunction with experimental observations by the group of R. J. de Boer and A. F. M. Marée (Utrecht, The Netherlands). These insights will inform future models in immunology, including extended models of autoimmune diabetes.

Finally, our model suggests that there are two distinct outcomes in an autoimmune attack typical of T1D: (1) The immune attack clearly subsides once the beta cells have been depleted, but here the outcome is full-blown diabetes. This explains observations in NOD mice, but, therapeutically, it is an outcome to be prevented. (2) More intriguing, any parameter change that shifts the system beyond its homoclinic bifurcation would also end the immune attack. This can happen through the process of "clonal exhaustion"; i.e, so much peptide is presented that memory-cell production is turned off completely. It could also happen through arrest of activation, where so little peptide is presented that T cells no longer become activated. In either case, if this happens before a significant fraction of the beta cells have been killed, it could provide a "cure" that resolves the autoimmunity without diabetes. Here we have hinted at several parameters that could have precisely this type of effect. This suggests that studying more detailed and hence more realistic variants of this model could indicate possible therapeutic strategies by highlighting which parameters give promising leads for medical targets.

Appendix A. Estimation of parameter values. We explain our procedure for estimating parameters below and summarize the values we used in Table B.1.

A.1. Cell turnover rates. The death rate of memory cells is estimated as $\delta_M = 0.01$ day^{-1} versus $\delta_E = 0.3$ day^{-1} for effector cells. We here assumed that activated cells have a relatively low death rate, as most are converted into differentiated cells. Consequently, we assumed that $\delta_A \approx 0.02$ day^{-1}.

A.2. Cell-division rates and numbers. We approximated an 8-hour cell cycle for the immune cells and thereby obtained $\beta \approx 1$–6 day^{-1}. The number of memory and effector cells produced per activated T cell is 0–8 versus 60, respectively, according to [14], leading to values for the factors 2^{m_i}.

A.3. Circulating cell levels. According to [14], around 1–10 naive T cells produced by the thymus per day will have the correct specificity. Consequently, $\sigma \approx$ 1–10 cells day^{-1}. To then determine the competition parameter, ϵ, we first considered the possibility of a QSS for activated T cells of the form $\sigma - \delta_A A - \epsilon A^2 = 0$. We found that this cannot be a correct approximation, because the reactivation of memory cells plays a much greater role in sustaining the level of A than the (limited) entry of new naive T cells from the thymus.

In our subsequent approach, we approximated $M \approx (\beta 2^{m_1} f_2/\delta_M)A \approx 10^4$ circulating memory cells and $E \approx (\beta 2^{m_2}(1-f_2)/\delta_E)A \approx 10^6$ circulating effector cells. The first of these implies that $\beta f_2 A \approx 10$, whereas the second implies that $\beta(1-f_2)A \approx 5 \times 10^3$. These approximations lead to $f_2 \approx 0.002$, and $A \approx 1 - 3 \times 10^3$.

Now considering the situation at a high-peptide level, near the peak of activated T-cell levels, we have $dA/dt \approx 0$, i.e., $\sigma + \alpha M f_1 - (\beta + \delta_A)A - \epsilon A^2 \approx 0$. The relative magnitudes of terms in this equation are as follows: σ is very small and can be neglected in the high-peptide scenario. If $\alpha \approx 1 - 5$ day^{-1} (which means that on average, a memory cell takes a few hours to be reactivated), and $f_1 \approx 1$ at high peptide, then $\alpha M f_1 \approx 1{-}5{\times}10^4$. From the above estimates, $(\beta{+}\delta_A)A \approx \beta A \approx 5{\times}10^3$ is of lower order, and $A^2 \approx 1 - 4 \times 10^6$. The balance is mainly between the terms $\alpha M f_1$ and ϵA^2. We can use these figures to estimate the size of the competition parameter, ϵ, from

$$\epsilon \approx \frac{\alpha M f_1}{A^2} \approx \frac{1 - 5 \times 10^4}{1 - 4 \times 10^6} \approx 1 - 5 \times 10^{-2}.$$

The units of ϵ are day^{-1}cell^{-1}.

A.4. Peptide and beta-cell levels. Because peptide level is not directly observed experimentally, its level in the model is on a relative, rather than absolute, scale. The important relation is the relative magnitude of k_1, k_2, the parameters that represent the level of peptide at which memory-cell production falls off and T-cell activation turns on, respectively. We arbitrarily chose $k_2 = 1$ and $k_1 = 2$. This means that a reasonable scale of peptide level is 0–10 "peptide units." Since peptide time scale is fast, and the peptide variable is assumed to be on QSS, only the ratio of the turnover rate, δ_p, and the production rate, R, of the peptide influence the dynamics. Based on the estimated levels of circulating effector T cells, we used $R \approx 10^{-5}$ per cell per day and $\delta_p = 1$ per day to give the QSS value of peptide in the range of 0–10. We also use a relative scale for the level of remaining beta cells; i.e., B represents the fraction of beta cells still remaining, and so $0 \le B \le 1$.

The removal of peptide by macrophages, by diffusion, and by other influences is assumed to be in the range of $\delta_p \approx 1$day^{-1}. When effector cell levels are high, $E \approx 10^6$ cells, this leads to the approximation $p \approx 10 \approx RBE/\delta_4 \approx R \times 10^6$. This leads to an estimate $R \approx 10^{-5}$ peptide units day^{-1}cell^{-1}.

A.5. Typical values of variables. The results of above ballpark estimates lead to the following ranges of the variables concerned:

$$A \approx 1 - 2 \times 10^3, \quad M \approx 1 - 5 \times 10^4, \quad E \approx 1 - 6 \times 10^6, \quad p \approx 1 - 10, \quad B \approx 1.$$

The populations of the three types of T cells differ by many orders of magnitude. We therefore scaled each variable in terms of some power of 10 for convenient graphics. The scaling considerations are discussed in the next section.

Appendix B. Scaling the equations. Let $A = A^*\bar{A}$, $M = M^*\bar{M}$, $E = E^*\bar{E}$, etc., where stars denote numerical values and overbars denote quantities carrying units. We keep time in units of days; i.e., time is not scaled. Equations (3.1)–(3.7) can be rewritten as follows:

$$\frac{dA^*}{dt} = \left(\frac{\sigma}{\bar{A}} + \left(\frac{\alpha\bar{M}}{\bar{A}}\right) M^*\right) f_1(p) - (\beta + \delta_A)A^* - (\epsilon\bar{A})(A^*)^2, \tag{B.1}$$

$$\frac{dM^*}{dt} = \left(\beta 2^{m_1}\frac{\bar{A}}{\bar{M}}\right) f_2(p)A^* - f_1(p)\alpha M^* - \delta_M M^*, \tag{B.2}$$

$$\frac{dE^*}{dt} = \left(\beta 2^{m_2}\frac{\bar{A}}{\bar{E}}\right)(1 - f_2(p))A^* - \delta_E E^*, \tag{B.3}$$

$$\frac{dB^*}{dt} = -(\kappa \bar{E}) E^* B^*, \tag{B.4}$$

$$\text{QSS}: \qquad p = \left(\frac{R \bar{E} \bar{B}}{\delta_p} \right) E^* B^*. \tag{B.5}$$

Since peptide is already in arbitrary units, we did not rescale the peptide or the functions f_1, f_2. Dropping the *'s, we thus obtained a new system of (scaled) equations,

$$\frac{dA}{dt} = (a_6 + a_7 M)\, f_1(p) - a_8 A - a_9 A^2, \tag{B.6}$$

$$\frac{dM}{dt} = a_{10} f_2(p) A - f_1(p) a_7 a_{16} M - a_{11} M, \tag{B.7}$$

$$\frac{dE}{dt} = a_{12}(1 - f_2(p)) A - a_{13} E, \tag{B.8}$$

$$\frac{dB}{dt} = -a_{17} E B, \tag{B.9}$$

$$\text{QSS}: \qquad p = (a_{14}/a_{15}) E B, \tag{B.10}$$

where the new parameters so defined are as follows:

TABLE B.1

Default "NOD mouse" parameter values used to simulate the model. See Appendix A *for a description of how these values were estimated.*

Par.	Meaning	Default value	Units	Ref.
σ	influx naive T cells from thymus	1–10	cell day^{-1}	[3, 14]
α	rate of production of A per M	1–5	day^{-1}	estimated
β	rate of cell division	1–6	day^{-1}	typical
δ_A	death rate, activated T cells	≈0.01	day^{-1}	[22, 7]
δ_M	death rate, memory T cells	≈0.01	day^{-1}	[22, 7, 14]
δ_E	death rate, effector T cells	0.3	day^{-1}	[5, 14]
δ_p	peptide turnover rate	0-1	day^{-1}	estimated
ϵ	T-cell competition parameter	$1 - 5 \times 10^{-2}$	(cell day)$^{-1}$	estimated
k_1	peptide level for $\frac{1}{2}$ max activation	2	peptide units	arbitrary
k_2	peptide level for $\frac{1}{2}$ max memory cells	1	peptide units	arbitrary
m	Hill coeff. for memory-cell production	2	-	[14]
n	Hill coeff. for T cell activation	3	-	[14]
2^{m_1}	maximum number of memory cells produced per proliferating T cell	8	-	[27, 20, 25]
2^{m_2}	number of effector cells produced per proliferating T cell	60	-	[20, 25]
a	maximal fraction of memory cells produced	< 1	-	fitted
R	peptide accumulation rate	10^{-5}	day^{-1}cell^{-1}	estimated
κ	beta-cell killing per effector T cell	0.14×10^{-6}	day^{-1}cell^{-1}	fitted

TABLE B.2

Parameter sensitivity. The default value of each (scaled) parameter is shown, and the effect of a 10% *increase and decrease is recorded. (See Appendix* C *for the definition of scaled parameters.) Original parameter values produce a stable limit cycle (i.e., dynamics between the Hopf and the homoclinic bifurcations).* $\rightarrow$ *denotes a shift towards the homoclinic bifurcation,* $\leftarrow$ *denotes a shift towards the Hopf bifurcation, or even beyond it, i.e., toward the stable steady state.* S *denotes the return to healthy state, and* P *signifies how many peaks (cycles) are seen before the homoclinic bifurcation occurs.* NC *means little or no change.*

Scaled parameter	Original parameter	Default value	Increase +10%	Decrease −10%
a_1	n	2	$S, \rightarrow$	10P, $\leftarrow$
a_2	k_1	2	S, $\rightarrow$	10P, $\leftarrow$
a_3	m	3	S, $\rightarrow$	9P, $\leftarrow$
a_4	a	0.7	NC, 3P	S, $\rightarrow$
a_5	k_2	1	9P$\leftarrow$	S, $\rightarrow$
a_6	σ	0.02	4P$\leftarrow$	S, $\rightarrow$
a_7	α	20	7P$\leftarrow$	S, $\rightarrow$
a_8	$\beta + \delta_A$	1	S$\rightarrow$	9P, $\leftarrow$
a_9	ϵ	1	NC, 3P	NC, 3P
a_{10}	$\beta 2^{m_1}$	1	8P$\leftarrow$	S, $\rightarrow$
a_{11}	δ_M	0.01	NC, 3P	NC,3P
a_{12}	$\beta 2^{m_2}$	0.1	4P$\leftarrow$	S, $\rightarrow$
a_{13}	δ_E	0.3	S$\rightarrow$	7P, $\leftarrow$
a_{14}	R	50	4P$\leftarrow$	S, $\rightarrow$
a_{15}	δ_p	1	S, $\rightarrow$	4P, $\leftarrow$
a_{16}	scale	0.1	S, $\rightarrow$	5P, $\leftarrow$
a_{17}	κ	0.14	1P, $\rightarrow$	4P, $\leftarrow$

$$a_6 = \frac{\sigma}{\bar{A}}, \quad a_7 = \frac{\alpha \bar{M}}{\bar{A}}, \quad a_8 = \beta + \delta_A, \quad a_9 = \epsilon \bar{A},$$

$$a_{10} = \beta 2^{m_1} \frac{\bar{A}}{\bar{M}}, \quad a_{11} = \delta_M, \quad a_{16} = \frac{\bar{A}}{\bar{M}},$$

$$a_{12} = \beta 2^{m_2} \frac{\bar{A}}{\bar{E}}, \quad a_{13} = \delta_E, \quad a_{17} = \kappa \bar{E}, \quad a_{14} = R \bar{E} \bar{B}, \quad a_{15} = \delta_p.$$

The original variables, A, M, E, differ by six orders of magnitude. We therefore selected a scaling of the main variables in various powers of 10, so as to display all results on a common coordinate system within the range of 1–10 units. To do so, we scaled variables by selecting the following reference scales:

$$\bar{A} = 10^3, \quad \bar{M} = 10^4, \quad \bar{E} = 10^6, \quad \bar{p} = 10, \quad \bar{B} = 1.$$

Appendix C. XPP code. Below is a typical file used for figures in this paper.

```
#XPP file for simulating AME system
#y1=A, y2=M, y3=E, y4=p, y5=B

y4 = a14*y3*y5/a15
f1 = y4^a1/(a2^a1+y4^a1)
f2 = a4*a5^a3/(a5^a3+y4^a3)
y1' = f1*(a6+a7*y2)-a8*y1-a9*y1^2
y2'= a10*f2*y1-f1*a16*a7*y2-a11*y2
y3' = a12*(1-f2)*y1-a13*y3
```

```
# If beta cell mass is a variable:
# y5' = -a17*y3*y5
# init y1=0.5,y2=0,y3=1,y5=1

#otherwise, for constant beta cells we use this:
init y1=0.5,y2=0,y3=1
par y5=1

par a1=2,a2=2,a3=3,a4=0.7,a5=1,a6=0.02
par a7=20,a8=1.0,a9=1.0,a10=1,a11=0.01,a12=0.1
par a13=0.3,a14=50,a15=1,a16=0.1,a17=0.14

@ dt=0.05, total=200
@ xlo=0,xhi=200,ylo=0,yhi=4
@ NPLOT=4, XP=t, YP=y1, XP2=t, YP2=y2, XP3=t, YP3=y3

done
```

Appendix D. List of abbreviations. We used the following abbreviations.

APC: antigen-presenting cell
CTL: cytotoxic T-lymphocyte
IGRP: islet-specific glucose-6-phosphatase catalytic subunit-related protein
MHC: major histocompatibility complex
NOD: nonobese diabetic mouse
ODE: ordinary differential equation
T1D: type 1 diabetes
TCR: T-cell receptor
QSS: quasi-steady state

Acknowledgments. We thank A. F. M. Marée, Eric Cytrynbaum, and members of Beta-CAAN (P. Santamaria, D. Finegood, B. Verchere, J. Dutz, D. Coombs, A. Khadra) for helpful discussions and comments.

REFERENCES

[1] A. Amrani, P. Serra, J. Yamanouchi, J. D. Trudeau, R. Tan, J. F. Elliott, and P. Santamaria, *Expansion of the antigenic repertoire of a single T cell receptor upon T cell activation*, J. Immunol., 167 (2001), pp. 655–666.

[2] A. Amrani, J. Verdaguer, P. Serra, S. Tafuro, R. Tan, and P. Santamaria, *Progression of autoimmune diabetes driven by avidity maturation of a T-cell population*, Nature, 406 (2000), pp. 739–742.

[3] J. A. M. Borghans, A. J. Noest, and R. J. De Boer, *How specific should immunological memory be?* J. Immunol., 163 (1999), pp. 569–575.

[4] J. A. M. Borghans, L. S. Taams, M. H. M. Wauben, and R. J. De Boer, *Competition for antigenic sites during T cell proliferation: A mathematical interpretation of* in vitro *data*, Proc. Natl. Acad. Sci. USA, 96 (1999), pp. 10782–10787.

[5] R. J. De Boer, M. Oprea, R. Antia, K. Murali-Krishna, R. Ahmed, and A. S. Perelson, *Recruitment times, proliferation, and apoptosis rates during the CD*8^+ *T-cell response to lymphocytic choriomeningitis virus*, J. Virol., 75 (2001), pp. 10663–10669.

[6] B. Han, P. Serra, A. Amrani, J. Yamanouchi, A. F. M. Marée, L. Edelstein-Keshet, and P. Santamaria, *Prevention of diabetes by manipulation of anti-IGRP autoimmunity: High efficiency of a low-affinity peptide*, Nat. Med., 11 (2005), pp. 645–652.

[7] D. R. Jackola and H. M. Hallgren, *Dynamic phenotypic restructuring of the CD4 and CD8 T-cell subsets with age in healthy humans: A compartmental model analysis*, Mech. Ageing Dev., 105 (1998), pp. 241–264.

[8] J. Jacob and D. Baltimore, *Modelling T-cell memory by genetic marking of memory T cells* in vivo, Nature, 399 (1999), pp. 593–597.

[9] C. Janeway, *Immunobiology: The Immune System in Health and Disease*, Garland Science, New York, 2005.

[10] S. M. Lieberman, A. M. Evans, B. Han, T. Takaki, Y. Vinnitskaya, J. A. Caldwell, D. V. Serreze, J. Shabanowitz, D. F. Hunt, S. G. Nathenson, P. Santamaria, and T. P. DiLorenzo, *Identification of the β cell antigen targeted by a prevalent population of pathogenic CD8$^+$ T cells in autoimmune diabetes*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 8384–8388.

[11] R. Maile, B. Wang, W. Schooler, A. Meyer, E. J. Collins, and J. A. Frelinger, *Antigen-specific modulation of an immune response by* in vivo *administration of soluble MHC class* I *tetramers*, J. Immunol., 167 (2001), pp. 3708–3714.

[12] A. F. M. Marée, M. Komba, C. Dyck, M. Łabeçki, D. T. Finegood, and L. Edelstein-Keshet, *Quantifying macrophage defects in type* 1 *diabetes*, J. Theoret. Biol., 233 (2005), pp. 533–551.

[13] A. F. M. Marée, R. Kublik, D. T. Finegood, and L. Edelstein-Keshet, *Modelling the onset of type* 1 *diabetes: Can impaired macrophage phagocytosis make the difference between health and disease?* Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 364 (2006), pp. 1267–1282.

[14] A. F. M. Marée, P. Santamaria, and L. Edelstein-Keshet, *Modeling competition among autoreactive CD8(+) T-cells in autoimmune diabetes: Implications for antigen-specific therapy*, Int. Immunol., 18 (2006), pp. 1067–1077.

[15] D. H. Margulies, *Interactions of TCRs with MHC-peptide complexes: A quantitative basis for mechanistic models*, Curr. Opin. Immunol., 9 (1997), pp. 390–395.

[16] T. W. McKeithan, *Kinetic proofreading in T-cell receptor signal transduction*, Proc. Nat. Acad. Sci. U.S.A., 92 (1995), pp. 5042–5046.

[17] D. Moskophidis, F. Lechner, H. Pircher, and R. M. Zinkernagel, *Virus persistence in acutely infected immunocompetent mice by exhaustion of antiviral cytotoxic effector T cells*, Nature, 362 (1993), pp. 758–761.

[18] B. A. O'Brien, W. E. Fieldus, C. J. Field, and D. T. Finegood, *Clearance of apoptotic beta-cells is reduced in neonatal autoimmune diabetes-prone rats*, Cell Death. Differ., 9 (2002), pp. 457–464.

[19] B. A. O'Brien, Y. Huang, X. Geng, J. P. Dutz, and D. T. Finegood, *Phagocytosis of apoptotic cells by macrophages from NOD mice is reduced*, Diabetes, 51 (2002), pp. 2481–2488.

[20] J. T. Opferman, B. T. Ober, and P. G. Ashton-Rickardt, *Linear differentiation of cytotoxic effectors into memory T lymphocytes*, Science, 283 (1999), pp. 1745–1748.

[21] P. A. Savage, J. J. Boniface, and M. M. Davis, *A kinetic basis for T cell receptor repertoire selection during an immune response*, Immunity, 10 (1999), pp. 485–492.

[22] D. F. Tough and J. Sprent, *Turnover of naive- and memory-phenotype T cells*, J. Exp. Med., 179 (1994), pp. 1127–1135.

[23] J. D. Trudeau, C. Kelly-Smith, C. B. Verchere, J. F. Elliott, J. P. Dutz, D. T. Finegood, P. Santamaria, and R. Tan, *Prediction of spontaneous autoimmune diabetes in NOD mice by quantification of autoreactive T cells in peripheral blood*, J. Clin. Invest., 111 (2003), pp. 217–223.

[24] S. Valitutti, S. Müller, M. Cella, E. Padovan, and A. Lanzavecchia, *Serial triggering of many T-cell receptors by a few peptide-MHC complexes*, Nature, 375 (1995), pp. 148–151.

[25] H. Veiga-Fernandes, U. Walter, C. Bourgeois, A. McLean, and B. Rocha, *Response of naïve and memory CD8$^+$ T cells to antigen stimulation* in vivo, Nat. Immunol., 1 (2000), pp. 47–53.

[26] A. Viola and A. Lanzavecchia, *T cell activation determined by T cell receptor number and tunable thresholds*, Science, 273 (1996), pp. 104–106.

[27] A. D. Wells, H. Gudmundsdottir, and L. A. Turka, *Following the fate of individual T cells throughout activation and clonal expansion: Signals from T cell receptor and CD28 differentially regulate the induction and duration of a proliferative response*, J. Clin. Invest., 100 (1997), pp. 3173–3183.

SIAM J. APPL. MATH.
Vol. 67, No. 4, pp. 938–959

WAVE SCATTERING AT THE SEA-ICE/ICE-SHELF TRANSITION WITH OTHER APPLICATIONS*

TIMOTHY D. WILLIAMS† AND VERNON A. SQUIRE†

Abstract. We present a mathematical model that describes how ice-coupled (flexural-gravity) waves traveling beneath a uniform, floating sea-ice sheet, defined over $(-\infty, 0)$, propagate into a second ice sheet (l, ∞) of different thickness by way of an arbitrarily defined transition region of finite width $(0, l)$. Each ice sheet is represented as an Euler–Bernoulli thin plate with a prescribed thickness and material properties, either or both of which vary across the transition. The most familiar application of this geometry is to sea-ice abutting an ice-shelf—a common occurrence found in the waters around Antarctica and parts of the Arctic or to sea-ice skirting *sikussak*—the band of extremely thick coastal fast ice that can form when local ice is sheltered from destructive storms. Another application is to breakwaters, and this is also discussed. By using Green's theorem two coupled integral equations are derived: one defined over $(0, l)$ and the second of the Wiener–Hopf type, defined over (l, ∞). The latter is solved analytically, allowing the integral equations to be decoupled and the first equation to be solved numerically. Results are presented for the geophysical and engineering examples referred to above.

Key words. Wiener–Hopf method, scattering, sea-ice/ice-shelf transition, flexible breakwaters

AMS subject classifications. 76B15, 86A40, 45A05, 34B60

DOI. 10.1137/060659351

1. Introduction.

In a recent paper [19] the authors consider a related geophysical problem, namely, how flexural-gravity waves propagating at the plate/water interface move between three floating elastic plates of different uniform thickness. That work enables several phenomena commonly observed in the polar regions to be modeled, e.g., wave propagation in the marginal ice zone—exploiting a feature of the model that allows any of the plate thicknesses to be zero, and wave scattering by open or refrozen leads and the abrupt thickness changes encountered at the edges of floes. It does not permit the two exterior sheets, defined over $(-\infty, 0)$ and (l, ∞), to be joined smoothly (or otherwise) through a transition region $(0, l)$ of prescribed variable thickness, as the thickness in each outer region is required to be constant. In this paper we relax that requirement, thereby expanding the range of both geophysical and marine engineering problems that can be modeled.

Of particular interest is what occurs when waves impinge on an ice-shelf of significantly greater thickness than the adjoining sea-ice sheet via a transition ramp that may stretch for tens or hundreds of meters or, in some cases, several kilometers. While it might be reasonably assumed that a typical ice-shelf would resist waves, in the Southern Ocean waves can be very long because of the immense fetches involved, and the passage of these waves is effectively unhindered by the sea-ice encountered en route. Waves of a somewhat shorter period—but still of sizable wavelength can also reach the transition depending on the nature of the sea-ice they encounter during their journey, as a cover of sea-ice acts to "attenuate" such waves through

*Received by the editors May 9, 2006; accepted for publication (in revised form) January 8, 2007; published electronically April 24, 2007. This work was supported by the Marsden Fund Council from Government funding, administered by the Royal Society of New Zealand, and by the University of Otago.

http://www.siam.org/journals/siap/67-4/65935.html

†Department of Mathematics and Statistics, University of Otago, P.O. Box 56, Dunedin, New Zealand (vernon.squire@otago.ac.nz, twilliams@maths.otago.ac.nz).

hysteresis, scattering, mechanical impacts and fracture, and hydrodynamical turbulence in the water column. Given sufficient amplitude ice-coupled waves of this type have the capacity to destroy an ice-shelf; this is a motivating interest of the current study.

Taken in the context of global warming this is of immediate topical significance, as temporal adjustments in pack ice serve as a proxy of climate change [14, 16, 4] and, especially in the vicinity of the Antarctic peninsula, there has been an increased incidence of breakup of ice-shelves with a concomitant recession in their fronts [20]. This paper addresses the complex relationship between an ice-shelf and its sea-ice shield in the context of how incident waves are scattered at the transition and, accordingly, are or are not allowed to progress into the shelf with reduced amplitude. It is found that reflection coefficients are oscillatory but that, because conservation of energy acts to suppress their fine structure, the amplitude transmission coefficient curves are less complicated and with increasing period simply decrease slightly from their low period values to a minimum before increasing monotonically towards perfect transmission at long periods (see Figure 7). In principle, the magnitudes of the waves within the ice-shelf, while reduced from their sea-ice amplitude, are still sufficient to cause the ice-shelf to flex and possibly to fracture and create an iceberg, although in practice, because the wavelength in the ice-shelf is much longer than in the neighbouring sea-ice, the associated strains will be modest. It is also possible that the finite geometry of the ice-shelf will cause certain wave modes to be favored at the expense of others, i.e., resonance [7, 15], but this is not modeled here as we have assumed the ice-shelf is semi-infinite in extent (l, ∞).

2. Equations and boundary conditions. Figure 1 illustrates schematically the situation that we are modeling: A plane wave with unit amplitude and radial frequency ω travels from the left-hand region through a central transition—where it is partially reflected and partially transmitted, into the right-hand region. The plates to the left and right are, respectively, of constant thickness h_0 and h_2, while the thickness across the transition varies according to $h_1(x)$. Throughout the paper

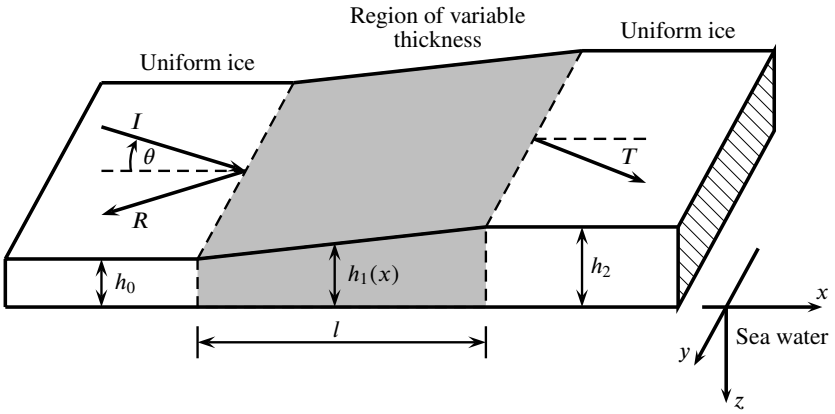


FIG. 1. *A plane flexural-gravity wave arrives from the left-hand region at an angle θ to normal incidence and is partially reflected and partially transmitted through a central transition of width l into the region beneath the right-hand plate. The thicknesses of the three plates are denoted by $h_0, h_1(x), h_2$, as shown, and each is modeled using a Euler–Bernoulli thin plate. Submergence is neglected, so the bottom of each plate is taken to be in the $z = 0$ plane. The left-hand edge of the central plate is located at $x = 0$ (the coordinate axes are displaced to the right to avoid clutter), and the sea water has a finite depth of H.*

subscripts of $j = 0$, 1, or 2 will be used to denote quantities referring to the different plates in the same way that they are in the h_j.

The amplitudes of the reflected and transmitted waves are denoted by R and T, respectively; hereafter referred to as the reflection and transmission coefficients. The determination of their values is the main purpose of our solution.

If we assume that the sea water beneath the ice is inviscid and of constant density and that the fluid flow is irrotational, then there exists a potential function $\Phi(x, y, z, t)$ such that the velocity of a fluid particle is given by $\boldsymbol{\nabla}\Phi$. Since the forcing from the incident wave is periodic in time and since the geometry of the problem is shift-invariant in the y direction, we assume that Φ has the following form:

$$\Phi(x, y, z, t) = \mathrm{Re}\big[\phi(x, z)e^{i(\alpha_y y - \omega t)}\big]. \tag{2.1}$$

This reduces the dimension of the problem from four to two. The wave number α_y will be related to the incoming wave's angle of incidence.

For each plate, define the flexural rigidity D_j, the mass per unit area m_j, and the characteristic length and time L_j and τ_j, as follows:

$$D_j = \frac{E_j h_j^3}{12(1-\nu_j)^2}, \quad m_j = \rho_j h_j, \quad L_j = \sqrt[4]{\frac{D_j}{\rho g}}, \quad \tau_j = \sqrt{\frac{L_j}{g}}.$$

E_j, ν_j, and ρ_j are the Young's modulus, Poisson's ratio, and density of the plate in the j^{th} region, respectively, while ρ is the water density and g is the acceleration due to gravity.

We now denote by the index m the region with the largest flexural rigidity and define the natural length that we will nondimensionalize lengths with respect to, L, as $L = (D_m/\rho\omega^2)^{1/5}$. If we also scale times by a factor of τ_m, then we have

$$(\bar{x}, \bar{y}, \bar{z}) = (x, y, z)/L, \quad \bar{t} = t/\tau_m, \quad \bar{\phi}(\bar{x}, \bar{z}) = \frac{\tau_m}{L^2}\phi(x, z), \quad \bar{\alpha}_y = \alpha_y L.$$

The other two significant lengths, l and H, are also scaled by L, so that $\bar{l} = l/L$ and $\overline{H} = H/L$. Further quantities that we will refer to are

$$\overline{D}_j = D_j/D_m, \quad \overline{m}_j = m_j/m_m, \quad \lambda = \frac{g}{L\omega^2} - i\varepsilon, \quad \mu = \frac{m_m}{\rho L},$$

where ε is an infinitesimal quantity introduced to force the reflected and transmitted waves to decay exponentially as they travel away from the central ice strip. The limit as it becomes zero will be taken once the solution has been completed.

Dropping the overbars to avoid clutter, $\phi(x, z)$ will satisfy the following system of equations:

$$(\nabla^2 - \alpha_y^2)\phi(x, z) = 0, \tag{2.2a}$$

$$\boldsymbol{\mathcal{L}}(x, \partial_x)\phi_z(x, 0) + \phi(x, 0) = 0, \tag{2.2b}$$

$$\phi_x(x^+, z) - \phi_x(x^-, z) = \phi(x^+, z) - \phi(x^-, z) = 0, \tag{2.2c}$$

$$\phi_z(x, H) = 0, \tag{2.2d}$$

where $\boldsymbol{\mathcal{L}}(x, \partial_x) = (\partial_x^2 - \alpha_y^2)D(x)(\partial_x^2 - \alpha_y^2) + (1-\nu)\alpha_y^2 D''(x) + \lambda - m(x)\mu$. The function $D(x)$ is defined piecewise, as follows:

$$D(x) = \begin{cases} D_0 & \text{for } x < 0, \\ D_1(x) & \text{for } 0 < x < l, \\ D_2 & \text{for } x > l, \end{cases}$$

and $m(x)$ is defined analogously in terms of the m_j.

As well as applying the above equations and the required radiation conditions (see section 3.2 below), the full solution must also satisfy some conditions at the two edges $x_e = 0$ and $x_e = l$. If two adjacent plates are joined or frozen together at a given edge, then we must apply what we shall subsequently call the fixed edge conditions:

$$\phi_z(x_e^+, 0) = \phi_z(x_e^-, 0), \tag{2.3a}$$

$$\phi_{zx}(x_e^+, 0) = \phi_{zx}(x_e^-, 0), \tag{2.3b}$$

$$\boldsymbol{\mathcal{M}}(x_e^+, \partial_x)\phi_z(x_e^+, 0) = \boldsymbol{\mathcal{M}}(x_e^-, \partial_x)\phi_z(x_e^-, 0), \tag{2.3c}$$

$$\boldsymbol{\mathcal{S}}(x_e^+, \partial_x)\phi_z(x_e^+, 0) = \boldsymbol{\mathcal{S}}(x_e^-, \partial_x)\phi_z(x_e^-, 0), \tag{2.3d}$$

where

$$\begin{aligned}\boldsymbol{\mathcal{M}}(x, \partial_x) &= D(x)\boldsymbol{\mathcal{L}}_-(\partial_x),\\ \boldsymbol{\mathcal{S}}(x, \partial_x) &= D(x)\boldsymbol{\mathcal{L}}_+(\partial_x) + D'(x)\boldsymbol{\mathcal{L}}_-(\partial_x),\\ \boldsymbol{\mathcal{L}}_\pm(\partial_x) &= (\partial_x^2 - \alpha_y^2) \mp (1-\nu)\alpha_y^2.\end{aligned}$$

If, on the other hand, the two plates are free to move independently, we must apply the free edge conditions:

$$\boldsymbol{\mathcal{M}}(x_e^\pm, \partial_x)\phi_z(x_e^\pm, 0) = 0, \tag{2.4a}$$

$$\boldsymbol{\mathcal{S}}(x_e^\pm, \partial_x)\phi_z(x_e^\pm, 0) = 0. \tag{2.4b}$$

Both sets of conditions imply that energy is conserved at each edge (i.e., no translational or rotational work is done by any of the edges). Note that even if the free edge conditions (2.4) are applied, (2.3c) and (2.3d) are still satisfied.

3. Solution method. To begin we use Green's theorem to derive a pair of coupled IEs (integral equations). The first depends on $\phi_z(x, 0)$ over the finite interval $(0, l)$, and the second is an integral of the Wiener–Hopf type [13] over the semi-infinite interval (l, ∞). The latter IE may be solved analytically using the Wiener–Hopf technique, allowing the two IEs to be decoupled. The IE over $(0, l)$ may then be solved numerically and, once the appropriate edge conditions have been applied, R and T can be calculated.

3.1. Green's function. We use a Green's function that satisfies the following set of equations:

$$\left(\partial_\xi^2 + \partial_\zeta^2 - \alpha_y^2\right)G(x-\xi, z, \zeta) = \delta(x-\xi, z-\zeta), \tag{3.1a}$$

$$\boldsymbol{\mathcal{L}}_0(\partial_\xi)G_\zeta(x-\xi, z, 0) + G(x-\xi, z, 0) = 0, \tag{3.1b}$$

$$G_\zeta(x-\xi, z, H) = 0, \tag{3.1c}$$

where $\boldsymbol{\mathcal{L}}_j(\partial_x) = D_j(\partial_x^2 - \alpha_y^2)^2 + \lambda - m_j\mu$ $(j = 0, 2)$.

This Green's function depends on the dispersion function for the left-hand region $f_0(\gamma)$, where $f_j(\gamma) = \coth(\gamma H)/\gamma - \Lambda_j(\gamma)$, $\Lambda_j(\gamma) = \boldsymbol{\mathcal{L}}_j(i\alpha) = D_j\gamma^4 + \lambda - m_j\mu$ $(j = 0, 2)$, and $\gamma(\alpha) = (\alpha^2 + \alpha_y^2)^{1/2}$. G is presented in [5] in terms of its Fourier transform with respect to $x - \xi$, $\hat{G}(\alpha, z, \zeta)$, which is given by

$$\hat{G}(\alpha, z, \zeta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} G(x-\xi, z, \zeta)e^{i\alpha(x-\xi)}d(x-\xi) = \chi(z_-, \gamma)\frac{\varphi(z_+, \gamma)}{f_0(\gamma)}, \tag{3.2}$$

where $z_+ = \max\{z, \zeta\}$, $z_- = \min\{z, \zeta\}$, and

$$\chi(z,\gamma) = \frac{\Lambda_1(\gamma)\gamma\cosh(\gamma z) - \sinh(\gamma z)}{\gamma^2 \tanh(\gamma H)}, \quad \varphi(z,\gamma) = \frac{\cosh\gamma(z-H)}{\cosh(\gamma H)}.$$

The derivative of G that is most relevant to this problem is $G_{z\zeta}(x-\xi,0,0)$, which we write as $g(x-\xi)$. It is given by

$$g(x-\xi) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{f_0(\gamma)} e^{-i\alpha(x-\xi)} d\alpha = i \sum_{\alpha\in S_0} A_0(\alpha) e^{i\alpha|x-\xi|}, \tag{3.3}$$

where, for $j = 0, 2$, $S_j = \{\alpha \,|\, f_j(\gamma) = 0 \ \&\ \mathrm{Im}(\alpha) > 0\}$ and

$$A_j(\alpha) = \gamma/\alpha f'(\gamma) = -\big(\gamma^2/\alpha\big)\big/\big(H(\Lambda_j^2(\gamma)\gamma^2 - 1) + 5D_j\gamma^4 + \lambda - m_j\mu\big).$$

In the limit as ε becomes zero in the definition of λ, the roots γ of each dispersion relation $f_j(\gamma) = 0$ are distributed throughout the complex plane as illustrated by Fox and Squire [6]. There is one positive real root that we shall label γ_j, a complex conjugate pair in the right-hand half-plane, an infinity of pure imaginary roots in the upper half-plane, and the negatives of the previously mentioned roots. The effect of the ε on the location of the roots is to produce an infinitesimal counterclockwise rotation. In particular, each γ_j is moved slightly off the real line into the upper half-plane.

The elements of each set S_j are given by $\alpha = \sqrt{\gamma^2 - \alpha_y^2}$, taking the square root from the upper half-plane. If $\varepsilon = 0$ and $\alpha_y < \gamma_j$, S_j also contains a real root α_j that moves into the upper half-plane as ε is increased. This forces the sum in (3.3) to decay exponentially as $|x-\xi| \to \infty$. The other members of S_j are two complex roots with the same imaginary parts and an infinity of pure imaginary roots.

If $\alpha_y < \gamma_0$, then $\alpha_y = \gamma_0 \sin\theta$, where θ is the angle of incidence. If $\alpha_y \geq \gamma_0$, then no propagating waves can exist in the left-hand region—in that case waves may travel parallel to the central transition but decay exponentially with distance in the perpendicular direction [5].

One further property of g that we will need to be aware of later is that it has a delta function type singularity in its fourth derivative. This can be seen by noting that

$$\frac{\Lambda_0(\gamma)}{f_0(\gamma)} = -1 + \frac{\coth\gamma H}{\gamma f_0(\gamma)}, \tag{3.4}$$

which is the Fourier transform of the equation

$$\boldsymbol{\mathcal{L}}_0(\partial_x) g(x) = -\delta(x) - G_\zeta(x,0,0). \tag{3.5}$$

3.2. Green's theorem. The radiation conditions alluded to above require that in the limit as $\varepsilon \to 0$ the potential corresponds to an incident wave arriving from the left and being either reflected or transmitted, i.e.,

$$\phi(x,z) \sim \begin{cases} \big(e^{i\alpha_0 x} + Re^{-i\alpha_0 x}\big)\varphi_0(z) & \text{as } x \to -\infty, \\ Te^{i\alpha_2 x}\varphi_2(z) & \text{as } x \to \infty, \end{cases} \tag{3.6}$$

where $\varphi_j(z) = \varphi(z, \gamma_j)$ for $j = 0, 1, 2$.

For $\varepsilon > 0$, $|\phi|$ will become infinite as $x \to -\infty$ due to the incident wave potential $\phi_0(x,z) = e^{i\alpha_0 x}\varphi_0(z)$. To circumvent this we solve for the function $\psi(x,z) = \phi(x,z) - \phi_0(x,z)$ instead, which decays exponentially as $|x| \to \infty$.

We begin the solution by using Green's theorem to derive a pair of coupled IEs. Using (2.2) and (3.1), ψ can be written

$$\psi(x,z) = \iint_\Omega \big(\nabla^2_{\xi\zeta} G\psi - G\nabla^2_{\xi\zeta}\psi\big)\,d\xi\,d\zeta = \oint_{\partial\Omega} \big(\partial_n G - G\partial_n\psi\big)\,ds \tag{3.7a}$$

$$= -\int_{-\infty}^{\infty} \big(G_\zeta(x-\xi,z,0)\psi(\xi,0) - G(x-\xi,z,0)\psi_z(\xi,0)\big)\,d\xi, \tag{3.7b}$$

where Ω is the fluid region, $\{\,(\xi,\zeta)\,|\,-\infty < \xi < \infty \;\&\; 0 < \zeta < H\}$, $\partial\Omega$ is the positively oriented boundary of Ω, s is the arc length as we travel around $\partial\Omega$, and ∂_n is the derivative with respect to the outward normal to $\partial\Omega$. (3.7b) follows from (3.7a) because the exponential decay of G and ψ as $|\xi| \to \infty$, along with the sea floor conditions (2.2d) and (3.1c), forces the other line integrals to vanish. If we eliminate G and ψ from (3.7b) using (2.2b) and (3.1b), integrate by parts, and simplify the resulting expression, we can write ψ entirely in terms of $\phi_z(x,0)$ for $x > 0$:

$$\psi(x,0) = \sum_{x_e \in X_e} \boldsymbol{P}^T_{x_e} \boldsymbol{\mathcal{L}}_{\text{edge}}(\partial_x) G_\zeta(x - x_e, z, 0) + \int_0^\infty \big(\boldsymbol{\mathcal{L}}(\xi,\partial_\xi) - \boldsymbol{\mathcal{L}}_0(\partial_\xi)\big) G_\zeta(x-\xi,z,0)\,\phi_{0,z}(\xi,0)\,d\xi, \tag{3.8}$$

where $\boldsymbol{P}_{x_e} = \boldsymbol{P}^+_{x_e} - \boldsymbol{P}^-_{x_e}$ are vectors containing four unknown constants that must be determined from the edge conditions. If $\boldsymbol{\mathcal{P}}(x,\partial_x) = D(x)\partial_x - D'(x)$, then the $\boldsymbol{P}^\pm_{x_e}$ and the operator $\boldsymbol{\mathcal{L}}_{\text{edge}}$ are

$$\boldsymbol{\mathcal{L}}_{\text{edge}}(\partial_x) = -\begin{pmatrix} \boldsymbol{\mathcal{L}}_+(\partial_x)\,\partial_x \\ \boldsymbol{\mathcal{L}}_-(\partial_x) \\ \partial_x \\ 1 \end{pmatrix}, \quad \boldsymbol{P}^\pm_{x_e} = \begin{pmatrix} D(x_e^\pm) \\ \boldsymbol{\mathcal{P}}(x_e^\pm,\partial_x) \\ \boldsymbol{\mathcal{M}}(x_e^\pm,\partial_x) \\ \boldsymbol{\mathcal{S}}(x_e^\pm,\partial_x) \end{pmatrix} \phi_z(x_e^\pm,0).$$

3.3. Integral equations. We now differentiate (3.8) with respect to z and let $z \to 0$ to give an IE in $\phi_z(x,0)$. We will solve it by splitting it into two different equations, corresponding to regions 0 and 2. The first equation, defined for $0 < x < l$, is

$$\phi_z(x,0) = e^{i\alpha_0 x}\varphi_0'(0) + \sum_{x_e \in X_e} \boldsymbol{P}^T_{x_e}\boldsymbol{\psi}(x - x_e) + \sum_{\alpha\in S_0} \beta_+(\alpha)e^{i\alpha x} + \int_0^l K(x,\xi)\phi_z(\xi,0)\,d\xi, \tag{3.9}$$

where $\boldsymbol{\psi}(x) = \boldsymbol{\mathcal{L}}_{\text{edge}} g(x)$ and

$$\beta_+(\alpha) = -iA_0(\alpha) f_2(\gamma)\Phi^+(\alpha), \tag{3.10a}$$

$$\Phi^+(\alpha) = \int_0^l \phi_z(\xi,0)e^{i\alpha\xi}\,d\xi, \tag{3.10b}$$

arise from differentiating the integral from l to ∞ in (3.8), letting $z \to 0$, and substituting (3.3) into the result.

The kernel $K(x,\xi)$ in (3.9) is given by

$$K(x,\xi) = \big(\boldsymbol{\mathcal{L}}(\xi,\partial_\xi) - \boldsymbol{\mathcal{L}}(\partial_\xi)\big)g(x-\xi) = \sum_{j=1}^{4} d_j(\xi)\boldsymbol{\mathcal{L}}_{1j}(\partial_\xi)g(x-\xi), \tag{3.11}$$

where if $d_0(x) = D_1(x)/D_0$, $d_1(x) = D_0\big(d_0(x)-1\big)$, $d_2(x) = 2D_1'(x)$, $d_3(x) = D_1''(x)$, and $d_4(x) = m_1(x) - m_0$. The four $\boldsymbol{\mathcal{L}}_{1j}$ operators are $\boldsymbol{\mathcal{L}}_{11}(\partial_x) = (\partial_x^2 - \alpha_y^2)^2$, $\boldsymbol{\mathcal{L}}_{12}(\partial_x) = (\partial_x^2 - \alpha_y^2)\partial_x$, $\boldsymbol{\mathcal{L}}_{13}(\partial_x) = \boldsymbol{\mathcal{L}}_{-}(\partial_x)$, and $\boldsymbol{\mathcal{L}}_{14}(\partial_x) = -\mu$.

Now, we know from (3.5) that $\boldsymbol{\mathcal{L}}_{11}(\partial_\xi)g(x-\xi)$ has a delta function type singularity. (This also implies that $\boldsymbol{\mathcal{L}}_{12}\,g$ will have a jump discontinuity.) When this is integrated out, it produces a $\big(1-d_0(\xi)\big)\phi_z(x,0)$ term on the right-hand side of (3.9). Hence, that equation becomes

$$\begin{aligned} d_0(x)\phi_z(x,0) = e^{i\alpha_0 x}\varphi_0'(0) &+ \sum_{x_e\in X_e} \boldsymbol{\psi}^T(x-x_e)\boldsymbol{P}_{x_e} \\ &+ \sum_{\alpha\in S_0} \beta_+(\alpha)e^{i\alpha x} + \sum_{j=1}^{4} \fint_0^l d_j(\xi)\boldsymbol{\mathcal{L}}_{1j}(\partial_\xi)g(x-\xi)\phi_z(\xi,0)d\xi, \end{aligned} \tag{3.12}$$

which can now be solved using numerical quadrature in terms of the unknown $\beta_+(\alpha)$ ($\alpha\in S_0$) coefficients. The Cauchy principal value symbol

$$\fint_0^l = \lim_{\varepsilon'\to 0}\left(\int_0^{x-\varepsilon'} + \int_{x+\varepsilon'}^{l}\right)$$

is used to show that any delta function singularities have been integrated out (as they are identically zero outside the interval $x-\varepsilon' < \xi < x+\varepsilon'$) and none of the integrands are actually Cauchy singular. The above limit can be taken analytically. The next step in the solution is to set up an IE that will facilitate the elimination of β_+ from (3.12) and allow us to solve for $\phi_z(x,0)$ within the ramp independently of its value in region 2.

The equation over $x\in(l,\infty)$ is obtained in the same way that (3.9) was found from (3.8), only this time we assume that $x > l$. Doing this gives us

$$\phi_z(x,0) - \sum_{\alpha\in S_0}\beta_-(\alpha)e^{i\alpha(x-l)} = \int_l^\infty \big(\boldsymbol{\mathcal{L}}_2(\partial_\xi) - \boldsymbol{\mathcal{L}}_0(\partial_\xi)\big)g(x-\xi,z,0)\,\phi_z(\xi,0)d\xi, \tag{3.13}$$

where if $\boldsymbol{p}^T(\alpha) = \boldsymbol{\mathcal{L}}_{\text{edge}}(-i\alpha)$, then

$$\begin{aligned} \beta_-(\alpha) = e^{-i\alpha l}\varphi_0'(0)\delta_{\alpha,\alpha_0} &+ iA_0(\alpha)\sum_{x_e\in X_e}\boldsymbol{p}^T(-\alpha)\boldsymbol{P}_{x_e}e^{i\alpha(l-x_e)} \\ &+ iA_0(\alpha)\boldsymbol{\mathcal{L}}_{1j}(-i\alpha)\sum_{j=1}^{4}\int_0^l d_j(\xi)\phi_z(\xi,0)e^{i\alpha(l-\xi)}d\xi. \end{aligned} \tag{3.14}$$

Equation (3.13) is an IE of the Wiener–Hopf type and thus may be solved analytically in terms of the β_- coefficients. We now proceed to show how this is done.

3.4. Solution of Wiener–Hopf integral equation. Let

$$\Phi^-(\alpha) = -\int_{-\infty}^{l}\int_{l}^{\infty} \big(\boldsymbol{\mathcal{L}}_2(\partial_\xi) - \boldsymbol{\mathcal{L}}_0(\partial_\xi)\big) g(x-\xi,z,0)\,\phi_z(\xi,0) e^{i\alpha(x-l)} d\xi\, d(x-l),$$

and let $\mathbb{C}^+$ and $\mathbb{C}^-$ be the upper and lower complex half-planes, respectively. Φ^- is analytic in $\mathbb{C}^-$ and, in the following, all functions with a "$-$" superscript will be analytic in $\mathbb{C}^-$ and will be termed "minus functions," and all functions with a "$+$" superscript will be analytic in $\mathbb{C}^+$ and will be termed "plus functions."

Assuming that the left-hand side of (3.13) is zero for $x < l$, taking its Fourier transform with respect to $x - l$ gives

$$\Phi^+(\alpha) - i\sum_{k\in S_0} \frac{\beta_-(k)}{\alpha+k} = \Phi^-(\alpha) + \left(1 - \frac{f_2(\gamma)}{f_0(\gamma)}\right)\Phi^+(\alpha), \tag{3.15}$$

so that

$$\frac{f_2(\gamma)}{f_0(\gamma)}\Phi^+(\alpha) - i\sum_{k\in S_0} \frac{\beta_-(k)}{\alpha+k} = \Phi^-(\alpha). \tag{3.16}$$

From [3], the quotient f_2/f_0 can be written as the product of a plus function $K^+(\alpha)$ and a minus function $K^-(\alpha) = K^+(-\alpha)$, where

$$K^+(\alpha) = \prod_{k\in S_2} \frac{\alpha+k}{\gamma(k)} \Big/ \prod_{k'\in S_0} \frac{\alpha+k'}{\gamma(k')}. \tag{3.17}$$

Consequently, we can rearrange (3.16) to give

$$\begin{aligned} K^+(\alpha)\Phi^+(\alpha) - i\sum_{k\in S_0} \frac{\beta_-(k)/K^+(k)}{\alpha+k} \\ = \frac{\Phi^-(\alpha)}{K^-(\alpha)} + i\sum_{k\in S_0} \frac{\beta_-(k)}{\alpha+k}\left(\frac{1}{K^-(\alpha)} - \frac{1}{K^+(k)}\right). \end{aligned} \tag{3.18}$$

Now the above equation states that its left-hand side, which is a plus function, agrees with its right-hand side, a minus function, on the real line. By the Riemann principle both sides must therefore be equal to a single entire function $J(\alpha)$. If $h_2 > 0$, then $K^\pm(\alpha) \sim O(1)$ as $\alpha \to \infty$, so it is clear that both sides of (3.18) are $O(\alpha^{-1})$. Hence $J(\alpha) = 0$ by Liouville's theorem, and we can write

$$\Phi^+(\alpha) = i\sum_{k\in S_0} \frac{\beta_-(k)/K^+(k)}{K^+(\alpha)(\alpha+k)}. \tag{3.19}$$

If $h_2 = 0$, $K^\pm(\alpha) \sim O(\alpha^{-2})$, so the right-hand side of (3.18) is potentially linear as $a \to \infty$. However, the left-hand side is again $O(\alpha^{-1})$, so $J = 0$ again and Φ^+ is still given by (3.19).

Substituting this formula for Φ^+ into (3.10a), we can now eliminate the β_+ coefficients from (3.9), putting them in terms of integrals involving $\phi_z(x,0)$ over $(0,l)$. These integrals may be approximated by quadrature, and (3.9) can be solved as a standard Fredholm IE over a finite interval. (We use the same method of doing this as [18].)

3.5. Scattering coefficients and application of the edge conditions. $\phi_z(x,0)$ can be calculated for $x<0$ from (3.8) and for $x>l$ from (3.19), giving the following eigenfunction expansions:

$$\phi_z(x,0)=\begin{cases} e^{i\alpha_0 x}+\sum_{\alpha\in S_0} a(\alpha)e^{-i\alpha x} & \text{for } x<0,\\ \sum_{\alpha\in S_2} b(\alpha)e^{i\alpha(x-l)} & \text{for } x>l,\end{cases} \tag{3.20}$$

where

$$a(\alpha)=iA_0(\alpha)\sum_{x_e\in X_e}\boldsymbol{p}^T(\alpha)\boldsymbol{P}_{x_e}e^{i\alpha x_e}+\beta_+(\alpha)e^{i\alpha l} + iA_0(\alpha)\sum_{j=1}^{4}\boldsymbol{\mathcal{L}}_{1j}(i\alpha)\int_0^l d_j(\xi)\phi_z(\xi,0)e^{i\alpha\xi}d\xi, \tag{3.21a}$$

$$b(\alpha)=A_2(\alpha)f_0(\gamma)\sum_{k\in S_0}\frac{K^+(\alpha)\beta_-(k)}{K^+(k)(\alpha-k)}. \tag{3.21b}$$

Once the unknowns in the $\boldsymbol{P}_{x_e}$ vectors have been found (by applying the edge conditions), R and T are given by $R=a(\alpha_0)/\varphi_0'(0)$ and $T=b(\alpha_2)e^{-i\alpha_2 l}/\varphi_2'(0)$, respectively.

The edge conditions (2.3c) and (2.3d) are applied by simply setting $\big[\boldsymbol{P}_{x_e}\big]_3$ and $\big[\boldsymbol{P}_{x_e}\big]_4$ to zero. If $X_e=\{0,l\}$, then the remaining two frozen edge conditions can be applied by substituting (3.20) into (2.3a) and (2.3b). However, if X_e has other elements, this is not possible. In that case (2.3a) is most easily applied by observing that, when it holds,

$$\big[\boldsymbol{P}_{x_e}\big]_1=\big(D(x_e^+)-D(x_e^-)\big)\phi_z(x_e^\pm,0)$$

(note that this is zero if D is continuous at x_e), while (2.3b) is best applied by requiring that

$$D(x_e^\pm)\big[\boldsymbol{P}_{x_e}\big]_2=\big(D(x_e^+)-D(x_e^-)\big)\boldsymbol{\mathcal{P}}(x_e^\pm,\partial_x)\phi_z(x_e^\pm,0) +\big(D(x_e^+)D'(x_e^-)-D(x_e^-)D'(x_e^+)\big)\phi_z(x_e^\pm,0). \tag{3.22}$$

The quantity $\boldsymbol{\mathcal{P}}(x_e^\pm,\partial_x)\phi_z(x_e^\pm,0)$ can be calculated without the necessity of numerical differentiation from the formula

$$\frac{1}{D_0}\boldsymbol{\mathcal{P}}(x_e^\pm,\partial_x)\phi_z(x_e^\pm,0)=i\alpha_0e^{i\alpha_0x}\varphi_0'(0)+\sum_{x_f\in X_e}\boldsymbol{P}_{x_f}^T\boldsymbol{\psi}'(x_e^\pm-x_f) +\sum_{j=1}^{4}\int_0^l d_j(\xi)\boldsymbol{\mathcal{L}}_{1j}(\partial_\xi)g'(x_e^\pm-\xi)\phi_z(\xi,0)d\xi, \tag{3.23}$$

which is obtained by differentiating (3.9) with respect to x and integrating out the delta function singularities produced.

When D is continuous, the continuous-slope condition simplifies to

$$\big[\boldsymbol{P}_{x_e}\big]_2=\big(D'(x_e^+)-D'(x_e^-)\big)\phi_z(x_e^\pm,0),$$

the right-hand side of which vanishes when D' is also continuous.

The free edge conditions can usually be applied by substituting (3.20) into (2.4). However, when $h_2 = 0$, or when there are free edges inside $(0, l)$, we must adjust our procedure. Applying $\boldsymbol{\mathcal{L}}_-(\partial_x)$ and $\boldsymbol{\mathcal{L}}_+(\partial_x)\partial_x$ to (3.9) and again allowing for delta function singularities gives the following formulae for the bending moment and the transverse edge force:

(3.24a)
$$\frac{1}{D_0}\boldsymbol{\mathcal{M}}(x_e^{\pm}, \partial_x)\phi_z(x,0) = \boldsymbol{\mathcal{L}}_-(\alpha_0)e^{i\alpha_0 x}\varphi_0'(0) + \sum_{x_f \in X_e} \boldsymbol{P}_{x_f}^T \boldsymbol{\mathcal{L}}_-(\partial_x)\boldsymbol{\psi}(x_e^{\pm} - x_f) + \sum_{j=1}^{4} \int_0^l \boldsymbol{\mathcal{L}}_{1j}(\partial_\xi)\boldsymbol{\mathcal{L}}_-(\partial_x) g(x_e^{\pm} - \xi) d_j(\xi)\phi_z(\xi, 0) d\xi,$$

(3.24b)
$$\frac{1}{D_0}\boldsymbol{\mathcal{S}}(x_e^{\pm}, \partial_x)\phi_z(x,0) = i\alpha_0\boldsymbol{\mathcal{L}}_+(\alpha_0)e^{i\alpha_0 x}\varphi_0'(0) + \sum_{x_f \in X_e} \boldsymbol{P}_{x_f}^T \boldsymbol{\mathcal{L}}_+(\partial_x)\boldsymbol{\psi}'(x_e^{\pm} - x_f) + \sum_{j=1}^{4} \int_0^l \boldsymbol{\mathcal{L}}_{1j}(\partial_\xi)\boldsymbol{\mathcal{L}}_+(\partial_x) g'(x_e^{\pm} - \xi) d_j(\xi)\phi_z(\xi, 0) d\xi.$$

Setting these to zero, choosing whether to apply them at x_e^+ or x_e^- since we have already applied (2.3c) and (2.3d), allows us to find $\left[\boldsymbol{P}_{x_e}\right]_1$ and $\left[\boldsymbol{P}_{x_e}\right]_2$.

4. Results. We begin this section by documenting the different types of transition profiles that can be used, classified in terms of the continuity properties of $D(x)$ and $D'(x)$ at the ends of the variable region (section 4.1). This is followed in section 4.2 by some results that aim to validate the theory and its numerical implementation.

After those two sections we describe two applications. Section 4.3 models a sea-ice/ice-shelf transition, such as occurs where the Ross Sea sea-ice meets the Ross ice-shelf, while section 4.4 presents a discussion on the effectiveness of breakwaters with different widths and thickness profiles. In principle the latter results could be used to protect a very large floating structure (VLFS) of pontoon-type such as a floating airport from incoming ocean waves.

4.1. Types of transition profiles. Transition profiles are classed as either type 0, type 1, or type 2. Type 0 transitions have the property that $D(x)$ and $D'(x)$ are continuous at their edges, and thus no edge conditions are required there (assuming frozen edge conditions are applied, one must always apply two conditions at each free edge); for type 1 transitions $D(x)$ is continuous but $D'(x)$ is not, and so one edge condition must be applied; for type 2 transitions $D(x)$ and/or $D'(x)$ are discontinuous, so that two edge conditions must be applied. Examples of each type are plotted in Figures 2(a), 2(d), and 2(g), and the scattering behavior of each is plotted in the two graphs that appear to their right. The middle column of plots (2(b), 2(e), and 2(h)) shows the scattering of normally incident waves, while the right-hand column (2(c), 2(f), and 2(i)) shows the scattering of obliquely incident waves for a number of periods.

All three profiles have the same average thickness and consequently produce similar amounts of reflection at normal incidence. The type 2 scattering curve, however, has the most structure at low periods, although we observe that the linear (type 1) transition also has a zero in $|R|$ at about 0.3 s. We attribute structure to the greater coherence of the shorter, uniform wavelengths in the intermediate 1.5-m-thick ice

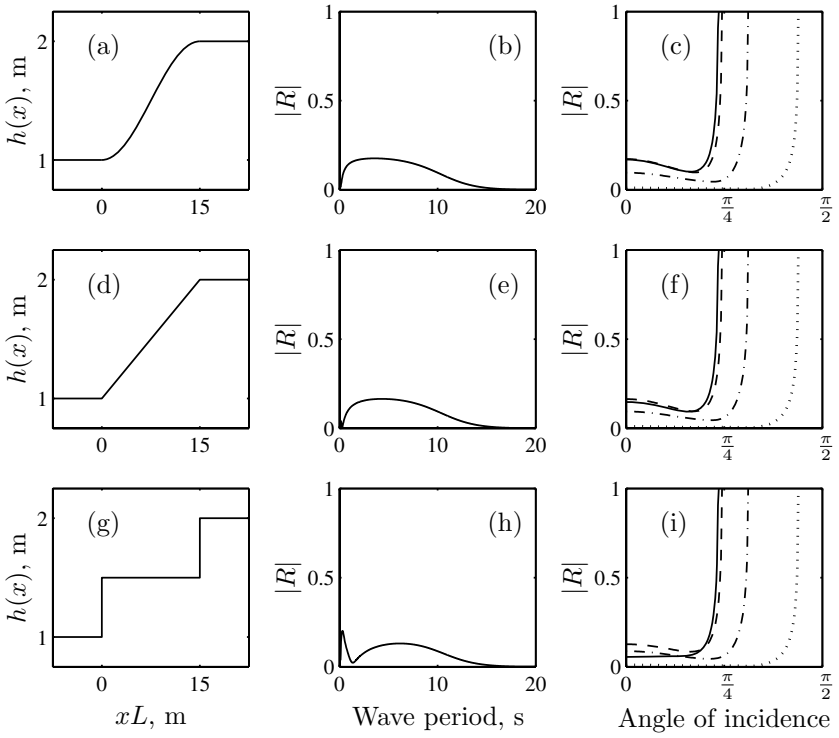


FIG. 2. *The scattering of waves incident on three different types of transition—a smooth type* 0 *transition with profile shown in* (a)*, a linear type 1 transition* (d)*, and a double step type* 2 *transition* (g)*. Scattering results for each different type of transition are shown in the two plots to the right of the corresponding thickness profile—figures in the second column* (b)*,* (e)*, and* (h) *illustrate the variation in the amount of reflection of normally incident waves with wave period, while figures in the right-hand column* (c)*,* (f)*, and* (i) *show the behavior of* $|R|$ *with the angle of incidence at a number of different wave periods. The periods used are* 2 *s (solid curves),* 5 *s (dashed curves),* 10 *s (dashed-dotted curves), and* 15 *s (dotted curves). The water depth is infinite.*

compared to the varying dispersion of the type 0 and type 1 cases. The more complicated the thickness profile is in a functional sense, the less the fine structure in the scattering curve and the more the reflection. This is also found for smooth ridges, which seem to produce more reflection than linear ones [18].

There are two main differences between the scattering of obliquely incident waves by transition zones and by ridges—both arising because $h_0 \neq h_2$. The first is that zeros in $|R|$ occur only when the period exceeds a certain value, e.g., in the three 15 s curves (dotted) of Figures 2(c), 2(f), and 2(i). Curves corresponding to lower periods either have minima or increase monotonically (see the 2 s curve in 2(i)). The second is that each period has a critical angle, less than $\pi/2$, above which any incident waves will be completely reflected. This is due to the wave number in the thinner sea-ice to the left (γ_0) being less than that in the thicker shelf ice on the right (γ_2). If α_y, the component of the wave number in the y direction, exceeds γ_2, no wave is able to propagate into the right-hand region in the x direction. (This corresponds to α_2 either vanishing or becoming imaginary.)

Figure 3 plots this critical angle as a function of period for a series of values of the ratio h_2/h_0. The two smallest values of h_2/h_0 show that as the ratio becomes smaller and smaller there is a "critical period" at which the critical angle reaches $\pi/2$ that

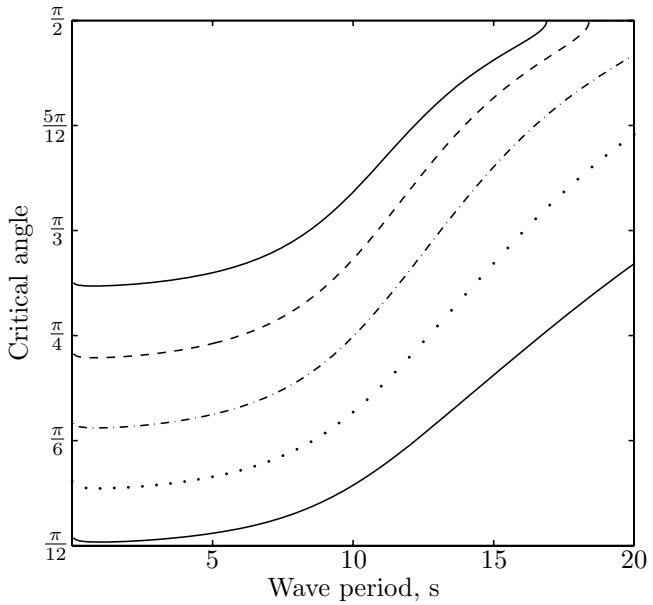


FIG. 3. *The variation of the critical angle of reflection with the period for different values of* h_2/h_0*:* 1.5 *(upper solid curve),* 2 *(dashed curve),* 3 *(dashed-dotted curve),* 5 *(dotted curve), and* 10 *(lower solid curve). The water depth is infinite.*

moves closer to the vertical axis. In the limit, when h_2 reaches h_0, the critical angle is constantly $\pi/2$ since the wave numbers on both sides of the transition are equal (cf. Figure 2). This critical period moves to larger periods as h_2 increases relative to h_0 and has moved out of the plotted range for the three curves corresponding to $h_2/h_0 = 3$, 5, and 10. A result of this is that the critical angle curves move downward as h_2/h_0 increases so that very little transmission of obliquely incident waves can occur when an ice-shelf is very thick.

4.2. Validation of results. The figures in this section are mostly intended to establish that the theory described in sections 2 and 3 produces results that converge to ones that can be independently verified by alternative solutions. Before this is done, however, in Figure 4 we attempt to reproduce Figure 5(a) of the paper by Porter and Porter [12]. These authors develop a variational solution for the spatially inhomogeneous floating plate and then invoke a mild slope approximation to enable results to be computed; zero submergence is not required.

Figure 4(a) plots the thickness profiles of the linear ramps that Porter and Porter use, and 4(b) plots the corresponding scattering calculated by our method as a function of incident wave number. The two methods agree well, although they diverge in the long wave limit. As $\alpha_0 \to 0$, Figure 4(b) predicts that $|R|$ will vanish for all three profiles, while the figure in [12] shows it tending towards a small finite value that decreases as h_2 decreases. This value is predicted in [9] and is the result of narrowing of the right-hand fluid region due to the increased submergence of the ice on that side. This example shows that the no-submergence assumption is unsuited to small finite depths.

Figure 5 illustrates how the scattering for a series of increasingly steep type 1 linear ramps converges to the scattering by a step. Results for a step can be checked

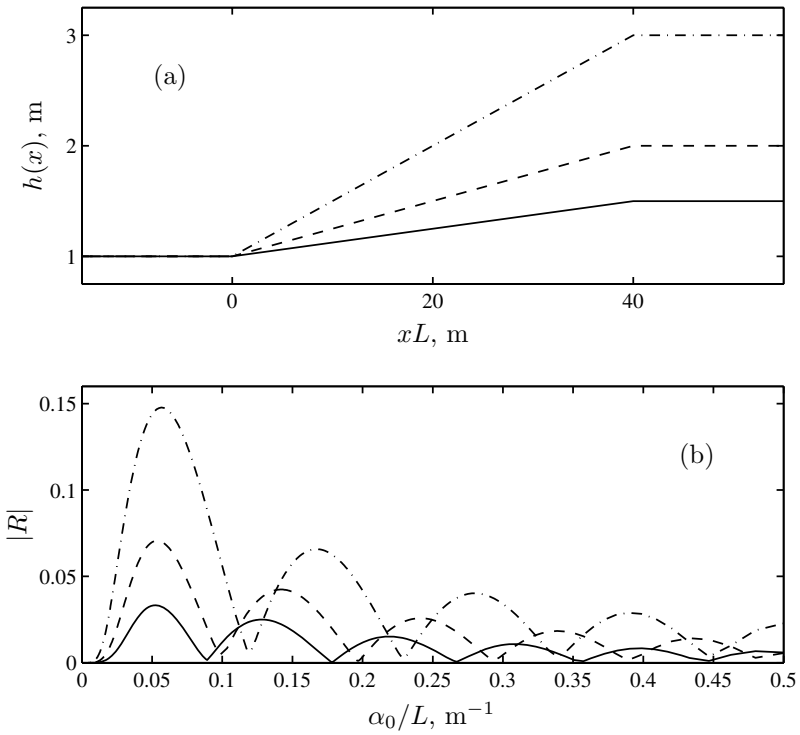


FIG. 4. *Verification of linear ramp results. The plot shown is a reproduction of Figure* 5(a) *of* [12]. *The upper plot* (a) *shows three different thickness profiles that ramp up linearly from* 1 *m to either* 1.5, 2, *or* 3 *m. The scattering by each ramp is plotted in* (b) *using the same line style used for its profile's line style in* (a). *In that figure* $|R|$ *is plotted against its dimensional wave number for the left-hand region,* α_0/L. *The incident waves are normally incident, and the water depth is* 20 *m.*

by comparing them to Figure 8 in [1], where $|R|$ displays a monotonic decrease from about 0.22 as the period is increased. As the steepness of the ramps is increased the progression of their reflection coefficient curves towards the step's scattering behavior can be seen, especially at large periods. It can also be seen in the period below which each ramp curve starts to significantly depart from the step curve. For the 100-m-wide ramp this happens at about 14 s, while the 25-m-wide ramp's reflection is still quite close to that for the step for periods above about 10.5 s. After each curve drops away, it has a maximum followed by either a minimum or a zero. The height of this maximum increases towards the step curve as steepness increases, while the period at which the zero occurs marches to the left. It has almost moved out of the plotted range by the time the ramp width has dropped to 25 m. These results are consistent with those of [18], which shows that the larger a feature's width in relation to the incident wavelength, the more opportunity there is for resonance and the more fine structure that is observed in the corresponding scattering curves.

An analogous computation for type 0 transition can also be done. The smooth transition curves display less fine structure than their linear equivalents, although their reflection is similar to that of the step for larger period ranges. That is, $|R|$ for each smooth transition begins to depart from $|R|$ for the step at a period about

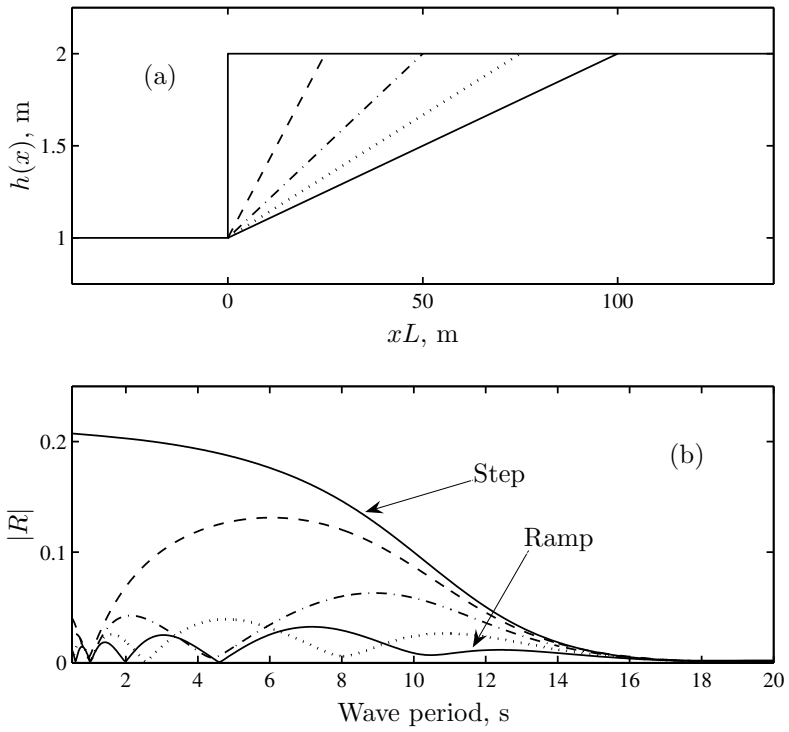


FIG. 5. *Convergence of the scattering by a series of increasingly steep type* 1 *linear ramps to the scattering by a step, which may be considered as an infinitely steep ramp. The curve corresponding to the step is pointed out in* (b)*, as is the curve corresponding to the ramp profile plotted in* (a) *with a solid line; the scattering curves for the other ramps are also plotted using the line style used for their profile in* (a)*. The incident waves are normally incident, and the water depth is infinite.*

1 s lower than the equivalent period for the linear ramp of the same width. Likewise the scattering by a sequence of smooth transitions as their profiles become more and more similar to a double step can be modeled. These examples and more are provided in [17].

We finish this section by reminding the reader that all of the results provided in [17] and in this paper assume that submergence is negligible. Comparing Figure 4 with the equivalent figure of [12] showed that this makes a slight difference for long waves. While this should be borne in mind at small depths, it will not cause problems for deeper water and especially not in the infinite depth limit.

4.3. The sea-ice/ice-shelf transition. The configurations modeled in sections 4.1 and 4.2 of one ice thickness ramping up to a different thickness over a given distance lend themselves well to modeling a ramp connecting relatively thin sea-ice to a much thicker, fresh water ice-shelf. Such a situation is found in the Ross Sea, where the transition from the sea-ice up to the Ross ice-shelf is gentle enough that one can easily walk up it.

Figure 6 plots the scattering by such a transition when the sea-ice has thickness $h_0 = 1\,\text{m}$ and the thickness of the ice-shelf, h_2, is (a) 5 m, (b) 10 m, (c) 15 m, and (d) 20 m. Results are plotted for transition widths of 250, 500, and 750 m.

A typical sea-ice/ice-shelf transition would ramp up from 1 to 15 m over about

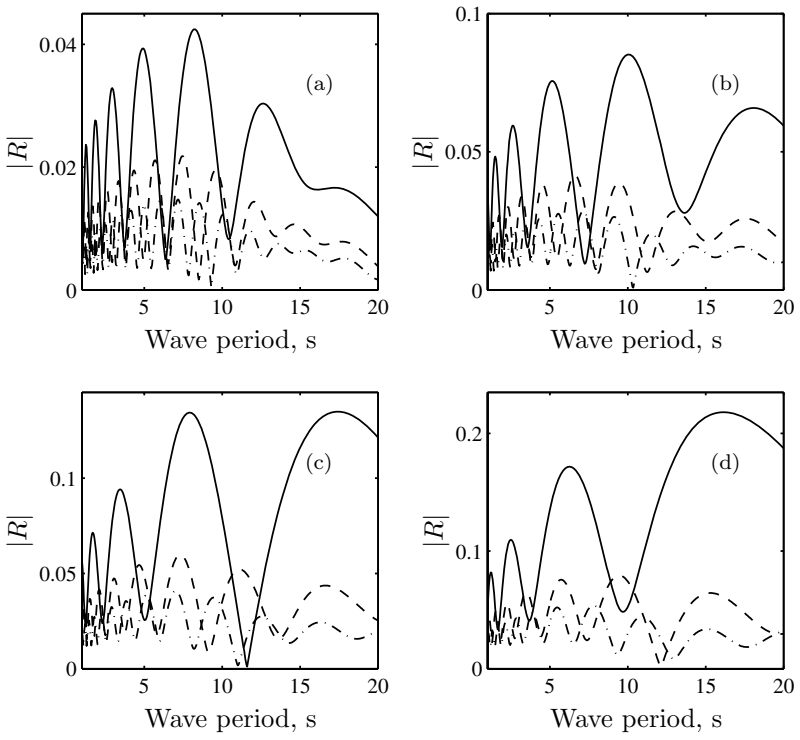


FIG. 6. *The reflection of normally incident waves produced by a sea-ice/ice-shelf transition. The sea-ice has thickness* $h_0 = 1\,m$*, and the thickness then increases linearly to values for* h_2 *of* (a) $5\,m$, (b) $10\,m$, (c) $15\,m$, *and* (d) $20\,m$. *The widths of the ramps are* $250\,m$ *(solid curves),* $500\,m$ *(dashed curves), and* $750\,m$ *(dashed-dotted curves), and the water depth is infinite.*

500 m (cf. the dashed curve in Figure 6(c)), but the other plots are included to give the reader more of an idea about the effect of changing the different dimensions.

Clearly, increasing the width of the ramp produces more complicated reflection patterns due to the occurrence of more resonances when several wavelengths repeat within the transition. In addition, increased width without a concomitant change in thickness lowers the slope of the ramp, making the overall change in the thickness less abrupt and in turn reducing the amount of reflection. In contrast, increasing the thickness h_2 while keeping the width constant increases the slope and thus causes more reflection. The next most noticeable effect of an increase in h_2 is that the scale over which the $|R|$ curves change increases: The spacings between successive maxima and minima become wider, and $|R|$ takes a lot longer to begin dropping to zero. The increase in spacings is attributed to the fact that as the ice becomes thicker the waves in the transition become much longer in comparison to its width. The slow decay in reflection is due to the overall size of the features taking longer to become insignificant in relation to the incident wavelength.

In general it appears that sea-ice/ice-shelf transitions of this type have the most marked effect on the longer waves (in comparison to other features in sea-ice). A value of $|R| = 0.19$ is the largest reflection of a 20 s wave produced by any feature modeled in the previous figures, and none of the features to be discussed later will produce such a large reflection for so long a wave. On the other hand, waves at smaller periods are not unaffected by the considerable increase in thickness as they travel into the region

beneath the ice-shelf. We have already mentioned how the wavelength of a wave of a given period increases with thickness. This increase is actually quite dramatic. For example, when the wave period is 5 s, the wavelength increases from 85 to 205 m as thickness increases from 1 to 5 m. When $h_2 = 20$ m, the wavelength becomes 450 m.

Before leaving Figure 6 we return again to the matter of submergence, as we wish to reassure the reader that the effects we are reporting would not be dwarfed by its assimilation into the model. To do this we invoke results from new work that allows both smooth and abrupt changes of property with submergence correctly incorporated [2]. Reassuringly the results are very similar to those reported herein [21] for the most challenging configuration, namely, the solid curve of Figure 6(b) where the thickness change is largest over the shortest horizontal distance. Differences between the current paper and [2] for no submergence are small (and explainable), giving us further confidence that the present theory is correct because the methods employed are mathematically independent, and the change when submergence is introduced is no more than 10% at worst. Of interest, the separation of the no-submergence and submergence results is negligible at short periods and for very long waves, with an intermediate range of periods where the curves deviate most (up to the 10% figure mentioned previously). It is argued that this occurs because at short periods the slope is so mild that the submergence is not noticed. As the period is gradually increased, the wavelength becomes longer with the result that the slope of the feature above and below the water line appears more abrupt. However, at very long periods the submergence is inconsequential compared to the wavelength. The validation of our results using [2] reassures us that the inclusion of submergence will not significantly alter our conclusions.

The wave amplitude is less under thicker ice due to its greater rigidity and, to a lesser extent, its greater mass. Accordingly, if an incident wave of amplitude A was perfectly transmitted into the region under the ice-shelf, its amplitude would drop to σA, where $\sigma = \varphi_2'(0)/\varphi_0'(0) < 1$. Figure 7 plots σ as a function of period for the four different values of h_2 used in Figure 6. It can be seen that the drop in amplitude is greatest at lower periods but that, as the period increases, the difference diminishes. As might have been expected, σ takes longer to climb to 1 as the ice-shelf becomes thicker. Figure 7 also plots the modulus of $\mathcal{T} = \sigma T$, the amplitude transmission coefficient, for the twelve ramps featured in Figure 6. Allowing for imperfect transmission, therefore, the final amplitude of a wave that initially has amplitude A is $A\mathcal{T}$.

We can see immediately that the width of the sea-ice/ice-shelf transition has only a negligible effect on the amplitude transmission coefficient. The most divergence is in Figure 7(d) when $h_2 = 20$ m, where the three curves separate very slightly around 19 s. In addition, the longer scale over which the curves for the thicker ice-shelves take to reach perfect transmission (the dotted curve) is also apparent.

A final point to make on comparison of Figure 7 with Figure 6 is that the $|\mathcal{T}|$ curves show no evidence at all of the oscillatory maxima and minima that were evident in the $|R|$ curves. This is expected and is a consequence of conservation of energy, i.e., $s|T|^2 = 1 - |R|^2$, where $s = \sigma^2 A_0(\alpha_0)/A_2(\alpha_2)$ is the intrinsic admittance [17]: If two values of $|R|$ are both quite small, their squares will be even smaller and the corresponding values of $|T|$ will both be very close to $1/\sqrt{s}$.

4.4. A flexible breakwater. Another physical application of the method described in this paper is to a breakwater shielding a VLFS from waves arriving from the open ocean. While it would be unusual for such structures to be made from ice,

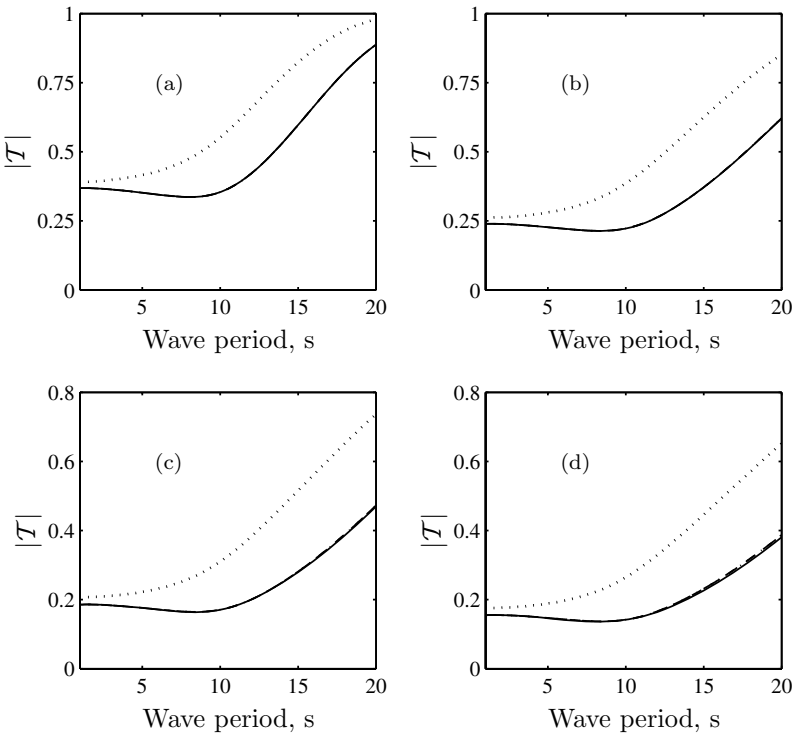


FIG. 7. *The relative amplitudes of normally incident waves transmitted by a sea-ice/ice-shelf transition. If a wave has amplitude A in the sea-ice region, its amplitude when it reaches the ice-shelf will be AT, where* $\mathcal{T} = \sigma T$, *and* $\sigma = \varphi_2'(0)/\varphi_0'(0)$ *is the relative amplitude of a perfectly transmitted wave.* σ *is plotted for reference as a dotted line when* $h_0 = 1\,m$ *and* h_2 *takes values of* (a) $5\,m$, (b) $10\,m$, (c) $15\,m$, *and* (d) $20\,m$. *In each figure, i.e., for each value of* h_2, $|\mathcal{T}|$ *is actually plotted when the widths of the transition regions are* $250\,m$ *(solid curve),* $500\,m$ *(dashed curve), and* $750\,m$ *(dashed-dotted curve) but the latter two curves are almost indistinguishable from the solid ones. The water depth is infinite.*

for convenience we shall take the building material to have the same density, Young's modulus and Poisson's ratio as sea-ice. As an aside the reader may be interested to know that there was actually a plan by the British during World War II to construct an aircraft carrier out of ice for use against German U-boats in the mid-Atlantic. Known as Project Habbakuk (an Admiralty clerk's misspelling of the biblical name Habakkuk), it was proposed that a 4000×600 ft VLFS with 40-ft-thick walls and a displacement of 2 million tons or more would be constructed in Canada from 280,000 blocks of ice. The building material was changed later to a mixture of ice and wood pulp known as pykrete but, because of the immense cost, the project was eventually scrapped.

Figure 8 shows the amounts of reflection produced by three different profiles for three different widths: 5 m (b), 15 m (c), and 30 m (d). The solid, dashed, and dashed-dotted curves correspond to the profiles in Figure 8(a) plotted in the same line style. The outer thicknesses are $h_0 = 1$ m and $h_2 = 0$, and the incident wave arrives normally from the right.

The differences in widths are not great enough for huge variations in the scattering patterns, but we can see that the maxima in $|R|$ at about 4 s in Figure 8(b) moves

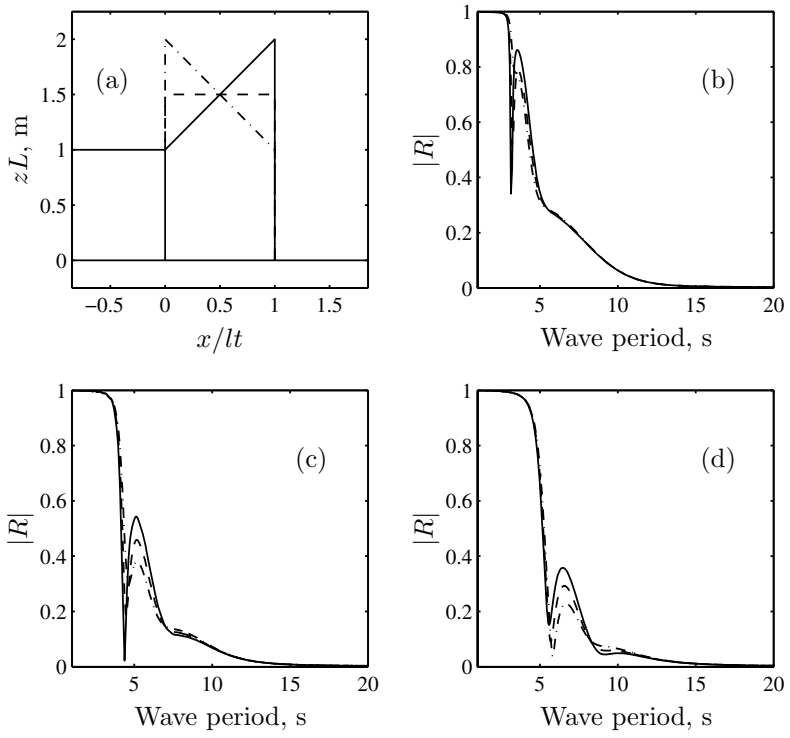


FIG. 8. *Wave scattering by a breakwater.* (a) *shows the three different shapes used. The incident wave arrives at a normal angle from the open water region to the right, and the breakwater is located between the* $x = 0$ *and* $x = l$ *planes and is separated by a free edge from the floating structure to the left that it shields. The reflection that each different shape produces is plotted in the same line style used for its profile in* (a) *when the width of the breakwater is either* (b) 5 *m,* (c) 15 *m, or* (d) 30 *m wide. The water depth is infinite.*

to the right and drops in height as l increases. The region of very high reflection at low periods also moves to the right. The breakwater that is sloping away from the open water (plotted as a solid curve) gives noticeably more reflection than the other breakwaters in the intermediate period range (about 5 to 10 s), although the size of this interval depends on the breakwater's width. When l is larger, it takes longer to become small in comparison to the incident wavelength, at which point the reflection by the three shapes converges. However, when we consider transmission into the region under the VLFS, the differences between the three different widths and shapes become less significant. This is demonstrated by Figure 9, which investigates the effect of a breakwater on a common open water wave spectrum.

The spectral density function (SDF) we use corresponds to a Pierson–Moskowitz wave spectrum [11], given by

$$\vec{f}_{\mathrm{sd}}(\tau) = \beta(\tau_0\tau)^3 \times \exp\big(-\eta(\tau_0\tau)^4\big), \tag{4.1}$$

where $\tau_0\tau$ is the dimensional wave period (τ is nondimensional), $\beta = 7.4\times10^{-4}\,\mathrm{m^2\,s^{-4}}$, and $\eta = 3.0 \times 10^{-4}\,\mathrm{s^{-4}}$. The arrow over the f is intended to indicate that this is the incident wave spectrum. In open water, this density function is very similar in shape to the dotted curves, climbing from zero in the short wave limit to a peak period at

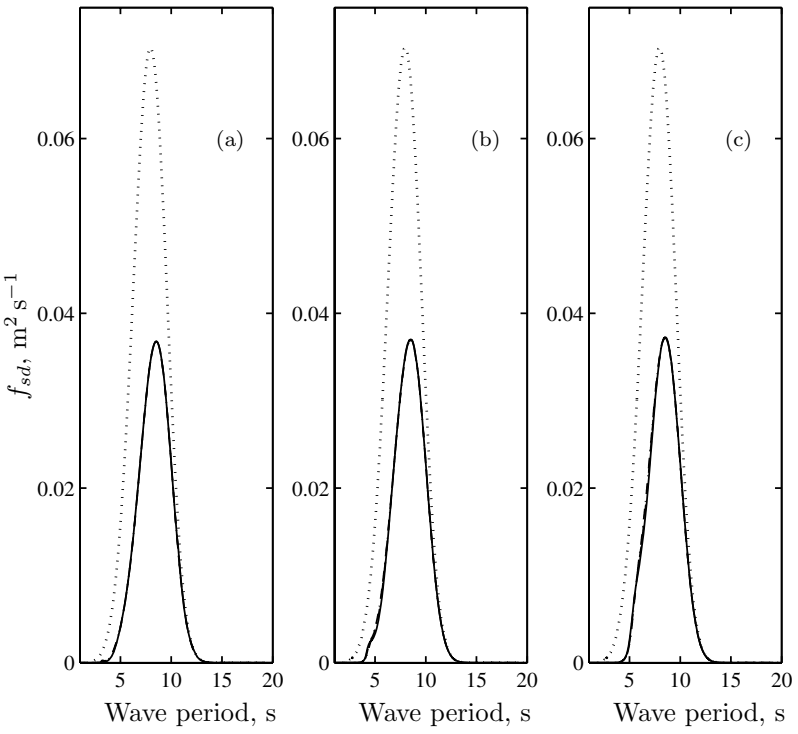


FIG. 9. *The effect of a breakwater on a Pierson–Moskowitz incident wave spectrum. The dotted lines in each figure are reference lines plotting the spectrum that would result if the Pierson–Moskowitz spectrum was transmitted perfectly from the open water region into the VLFS. If* $\vec{f}_{\rm sd}$ *is the SDF for the incident wave spectrum, this reference SDF is given by* $f_{\rm sd} = \sigma^2 \vec{f}_{\rm sd}$. *The dotted, dashed, and dashed-dotted curves show the SDFs resulting from the breakwater profiles plotted in Figure* 8 *in the same line styles, and their widths are either* (a) 5 *m,* (b) 15 *m, or* (c) 30 *m. (Note that the dashed and dashed-dotted curves are very hard to distinguish from the solid curves.) Their SDFs are given by* $f_{\rm sd} = \vec{f}_{\rm sd} \times |\mathcal{T}|^2$. *The incident waves are all taken to arrive normally, and the water depth is infinite.*

8 s, before dropping to zero again as the period increases further.

The dotted lines in Figure 9 actually plot $\sigma^2 \vec{f}_{\rm sd}(\tau)$, which is intended as a reference to show what the wave spectrum beneath the VLFS would look like if the breakwater did not produce any reflection. The solid, dashed, and dashed-dotted curves plot the spectra resulting from the different shaped breakwaters plotted in Figure 8(a), while the different subplots, Figures 9(a), 9(b), and 9(c), correspond to the three different widths used in Figure 8(b), 8(c), and 8(d), respectively.

The spectra are calculated by $f_{\rm sd}(\tau) = \vec{f}_{\rm sd}(\tau) \times |\mathcal{T}|^2$, and it can be seen that the different breakwater shapes produce a negligible difference in the wave spectra underneath the VLFS. The width does not seem to make a large amount of difference either, although the 5-m-wide breakwaters seem to filter out a little more wave amplitude.

Similarly, Figure 10 shows that using a breakwater with a larger average thickness does not affect the transmitted wave spectrum markedly, although in general it does increase $|R|$. Figures 10(a) and 10(c) show $|R|$ and $f_{\rm sd}$ for three 5-m-wide breakwaters: 1 m thick (solid curve), 2 m thick (dashed curve), and 3 m thick (dashed-dotted curve).

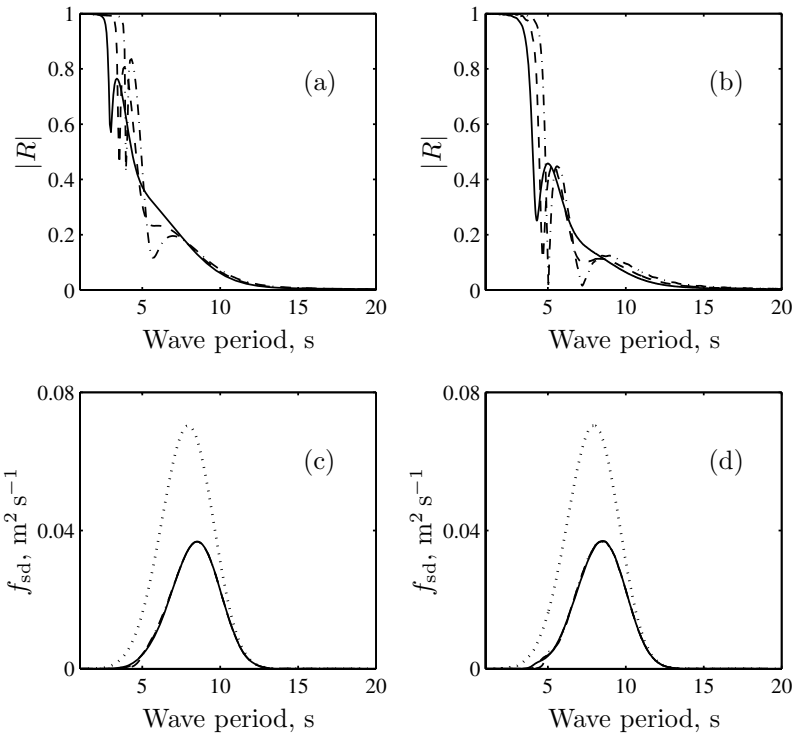


FIG. 10. *Further investigations of the reflection and transmission produced by a breakwater.* (a) *and* (c) *show* $|R|$ *and* f_{sd} *for three* 5*-m-wide breakwaters of* 1 *m thickness (solid curve),* 2 *m thickness (dashed curve), and* 3 *m thickness (dashed-dotted curve), and* (b) *and* (d) *show the analogous results for* 15*-m-wide breakwaters. The waves are normally incident, and the water depth is infinite.*

Figures 10(b) and 10(d) show the analogous results for 15-m-wide breakwaters.

The results of this section lead one to conclude that, should one be required to construct a flexible breakwater to protect a VLFS, the extra expense and effort needed to make it wider and/or thicker would not filter out a significant proportion of the incoming wave amplitude. Of course, if the thickness was reduced too much, reflection would start to decrease, but keeping it at 1 m would seem to be sufficient.

The most economical approach to improving the amount of protection that is afforded the VLFS would probably be to have several smaller breakwaters side by side—possibly separating them by small expanses of open water. Increasing the number of abrupt changes in surface properties that the incident waves must cross would help to produce greater reflection.

5. Conclusions. We have considered wave scattering arising from a change of physical properties—referred to as the transition, across two floating, Euler–Bernoulli, elastic plates. A new model is reported that utilizes Green's theorem to construct two coupled IEs, defined over $(0, l)$ and (l, ∞), respectively, and invokes the Wiener–Hopf method to decouple the system by first solving the IE in (l, ∞). After validating the model two applications are discussed.

1. Wave scattering at the sea-ice/ice-shelf transition, whereby ocean waves incident from the open ocean enter an extended region of sea-ice that may be shore-fast

or separated from the coast by a lead, i.e., a "river" in the ice, before impinging on the ice-shelf itself. This geophysical application of the theory has topical relevance because climate warming is known to be inducing ice-shelf melting in specific regions of Antarctic, e.g., the Antarctic Peninsula alongside the Weddell Sea, and is contributing to a reduction in pack ice and shore-fast ice that would formerly have protected ice-shelves from severe storms. The model described herein replicates this important marine system mathematically, potentially allowing anthropogenically assisted changes in the coastal ice masses to be utilized as a proxy of global change. By way of example it is proposed that in concert with surface-meltwater-enhanced fracture involved in the breakup of the 200–300 m-thick Larson B ice-shelf [10, 8], the removal of the sea-ice barrier and temperature-induced weakening of the ice were precursors that preconditioned the ice-shelf for breakup. An opportunity to test this hypothesis using the current model to compute the flexural waves that energize the ice-shelf with and without the sea-ice barrier in place has arisen and will be the focus of a separate paper directed at the geophysics as opposed to the mathematical development. While nothing in polar geophysics is ever straightforward, the mechanisms we have suggested are worthy of consideration, especially when the possibility of standing waves created by waves reflected at the hingeline and geometric resonance are also incorporated as contributing factors [7, 15].

2. Floating, flexible breakwaters are also considered as an application of the current model. Here, for mathematical convenience the configuration is reversed so that waves travel towards the transition from the right. Concomitantly, the *ice-shelf* is allowed to become vanishingly thin, i.e., open water. Waves then propagate through the transition region into a floating plate of uniform prescribed thickness. This is an effective model of a VLFS, e.g., a floating airport, shielded by a skirt of specified geometry, and the question becomes "how should the skirt be designed to minimize the amount of wave energy reaching the VLFS?" This is not outrageously futuristic. A six-year plan launched in 1995 to research and develop Megafloat, a floating airport in Yokosuka Bay, Japan, involved the construction of a 1,000 m by 60–121 m model that successfully passed takeoff and landing tests. A detailed evaluation of the tests on the 1000 m Megafloat and the 4000 m-class test design concluded that a Megafloat airport with a scale of up to 4000 m was more than feasible. Breakwaters of some type will be necessary to avoid significant wave-induced flexure of the VLFS, so it is fitting that we investigate their design using the tools we have developed. Of significance, we find that the effect of thickness and width of the transition is secondary once a certain thickness is reached, so a simple cost-benefit analysis needs to be done to decide the optimum configuration for a particular VLFS.

Acknowledgment. The authors are grateful to Luke Bennetts, who helped us to quantify the effect of assuming zero submergence.

REFERENCES

[1] M. Barrett and V. A. Squire, *Ice-coupled wave propagation across an abrupt change in ice rigidity, density or thickness*, J. Geophys. Res., 101 (1996), pp. 20825–20832.

[2] L. G. Bennetts, N. R. T. Biggs, and D. Porter, *A multi-mode approximation to wave scattering by ice sheets of varying thickness*, J. Fluid Mech., 579 (2007), pp. 413–443.

[3] H. Chung and C. Fox, *Calculation of wave-ice interaction using the Wiener-Hopf technique*, NZ J. Math, 31 (2002), pp. 1–18.

[4] J. Comiso, *A rapidly decliing perenial sea-ice cover in the Arctic*, Geophys. Res. Lett., 29 (2002), p. 1956.

[5] D. V. Evans and R. Porter, *Wave scattering by narrow cracks in ice sheets floating on water*

of finite depth, J. Fluid Mech., 484 (2003), pp. 143–165.
[6] C. Fox and V. A. Squire, *Reflection and transmission characteristics at the edge of shore fast sea ice*, J. Geophys. Res., 95 (1990), pp. 11629–11639.
[7] E. H. Gui and V. A. Squire, *Random vibration of floating ice tongues*, Ant. Sci., 1 (1989), pp. 157–163.
[8] C. L. Hulbe, D. R. MacAyeal, G. H. Denton, J. Kleman, and T. V. Lowell, *Catastrophic ice shelf break as the source of Heinrich event icebergs*, Paleoceanography, 19 (2004), p. PA1004.
[9] H. Lamb, *Hydrodynamics*, Cambridge University Press, New York, 1932, 6th ed.
[10] D. R. MacAyeal, T. A. Scambos, C. L. Hulbe, and M. A. Fahnestock, *Catastrophic ice-shelf break-up by an ice-shelf fragment capsize mechanism*, J. Glaciol., 49 (2003), pp. 22–36.
[11] O. M. Phillips, *The Dynamics of the Upper Ocean*, Cambridge University Press, New York, 1977, 2nd ed.
[12] D. Porter and R. Porter, *Approximations to wave scattering by an ice sheet of variable thickness over undulating bed topography*, J. Fluid Mech., 509 (2004), pp. 145–179.
[13] B. W. Roos, *Analytical Functions and Distributions in Physics and Engineering*, Wiley, New York, 1969.
[14] D. A. Rothrock, Y. Yu, and G. A. Maykut, *Thinning of the Arctic sea-ice cover*, Geophys. Res. Lett., 26 (1999), pp. 3469–3472.
[15] V. A. Squire, W. H. Robinson, M. H. Meylan, and T. G. Haskell, *Observations of flexural waves in the Erebus Glacier Tongue, McMurdo Sound, Antarctica, and nearby sea ice*, J. Glaciol., 40 (1994), pp. 377–385.
[16] P. Wadhams and N. R. Davis, *Further evidence of ice thinning in the Arctic ocean*, Geophys. Res. Lett., 27 (2000), pp. 3973–3976.
[17] T. D. Williams, *Reflections on Ice: The Scattering of Flexural-Gravity Waves by Irregularities in Arctic and Antarctic Ice Sheets*, Ph.D. thesis, University of Otago, 2005.
[18] T. D. Williams and V. A. Squire, *Oblique scattering of plane flexural-gravity waves by heterogeneities in sea ice*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 3469–3497.
[19] T. D. Williams and V. A. Squire, *Scattering of flexural–gravity waves at the boundaries between three floating sheets with applications*, J. Fluid Mech., 569 (2006), pp. 113–140.
[20] D. Vaughn, *private commumication*, 2006.
[21] L. Bennetts, *private commumication*, 2006.

SIAM J. APPL. MATH.
Vol. 67, No. 4, pp. 960–989

INVERSE PROBLEMS RELATED TO ION CHANNEL SELECTIVITY*

MARTIN BURGER†, ROBERT S. EISENBERG‡, AND HEINZ W. ENGL§

Abstract. Ion channels control many biological processes in cells, and, consequently, a large amount of research is devoted to this topic. Great progress in the understanding of channel function has been made recently using advanced mathematical modeling and simulation. This paper investigates another interesting mathematical topic, namely inverse problems, in connection with ion channels. We concentrate on problems that arise when we try to determine ("identify") one of the structural features of a channel—its permanent charge—from measurements of its function, namely current-voltage curves in many solutions. We also try to design channels with desirable properties—for example with particular selectivity properties—using the methods of inverse problems. The use of mathematical methods of identification will help in the design of efficient experiments to determine the properties of ion channels. Closely related mathematical methods will allow the rational design of ion channels useful in many applications, technological and medical. We also discuss certain mathematical issues arising in these inverse problems, such as their ill-posedness and the choice of regularization techniques, as well as challenges in their numerical solution. The L-type Ca channel is studied with the methods of inverse problems to see how mathematics can aid in the analysis of existing ion channels and the design of new ones.

Key words. ion channels, Poisson–Nernst–Planck equations, identification, optimal design, permanent charge, current-voltage relations

AMS subject classifications. 35R30, 92C05, 92C40, 65N21, 65R32

DOI. 10.1137/060664689

1. Introduction.

Ion channels are proteins with a hole down their middle that allow ions to move through otherwise impermeable cell membranes, thereby controlling many biological processes of great importance in health and disease. Interest in channels has grown rapidly because of their general role as controllers of biological function in health and disease. A quick glance at the literature through a search on the Internet will find hundreds of papers on channelopathies, diseases of channels (cf. [As99, LHJR00]). Specifically, channels are proteins akin to enzymes (cf. [Ei90]) that control the flow of ions through membranes and thus control a wide range of biological functions (cf. [Aletal94, Hi01]).

Channels generate the action potential which conducts all information in the nervous system and coordinates contraction, including the contraction which allows the heart to function as a pump. Channels are involved in nearly all sensory function, in the secretion of hormones, and in the function of the kidney and intestine. There is hardly a biological function that is not controlled by channels or transporters in an important way.

*Received by the editors July 10, 2006; accepted for publication (in revised form) January 22, 2007; published electronically April 24, 2007.
http://www.siam.org/journals/siap/67-4/66468.html

†Institut für Industriemathematik, Johannes Kepler Universität, Altenbergerstr. 69, A-4040 Linz, Austria. Current address: Institut für Numerische und Angewandte Mathematik, Westfälische Wilhelms-Universität Münster, Einsteinstr. 62, D-48149 Münster, Germany (martin.burger@uni-muenster.de).

‡Department of Molecular Biophysics and Physiology, Rush University Medical Center, 1750 W. Harrison St., Chicago, IL 60612 (beisenbe@rush.edu). This author's work was supported in part by NIH grant GM076013.

§Institut für Industriemathematik, Johannes Kepler Universität, Altenbergerstr. 69, A-4040 Linz, Austria (heinz.engl@jku.at).

The enormous importance of channels has generated enormous amounts of experimental work. Literally hundreds of laboratories and thousands of scientists measure channel properties every day with remarkable resolution, often studying the properties of just one protein molecule. Molecular genetics and molecular biology allow routine (although tedious) engineering of channel proteins nearly one atom at a time (cf. [Mietal06]). Few areas of biology are so well explored at such resolution.

Channels also are much simpler than enzymes. Channel function does not involve changes in covalent bonds or chemistry in that sense. Channels perform many of their functions without changing structure (on the biological time scale of msec). Ions move through channels driven by concentration gradients and electrical potential at room temperature. Channels form an unusual nearly unique system because they are both physically simple and biologically very important. The daunting complexity of the structure of many biological systems is not found in single molecules of channel proteins [TBSS01, Ei98, Maetal03].

One of the defining characteristics of proteins is their selectivity. Most proteins bind specific organic chemicals with great specificity even at very low concentrations, 10^{-5} times smaller than concentrations of ions always associated with proteins, e.g., K^+, Na^+, and Cl^-, which are typically found at 0.2 M concentration. These organic molecules often control the biological function of the protein with great specificity even at these very low concentrations. Ion channels (for example) conduct ions of one type much better than ions of another type, and this selectivity among ions is essential for their role in signaling in the nervous system, and coordination of muscle contraction, particularly in the heart. If the selectivity of ion channels is understood, and a physical theory is available showing how channel structure produces channel function, channel proteins can be designed to specification and built using the well-developed techniques of molecular engineering, e.g., by site-directed mutagenesis.

The design of ion channels to specification can also be seen as an application of the mathematical theory of inverse problems ("reverse engineering"). Design requires specialized mathematics because of the complexity and sensitivity (with respect to perturbations) of the system and the mutual dependence of various design goals: improving some properties can make others worse, and so mathematics is needed to find a good compromise. In this paper, we show how iterative and variational regularization methods developed for inverse problems can be applied to design or identify the function—in particular the selectivity—of ion channels using the physical chemistry of crowded charge, which is modeled through the Poisson–Nernst–Planck equations, a system of nonlinear partial differential equations combined with a density functional theory of excess chemical potential. The main idea of this approach is to formulate the design or identification of permanent charge as an abstract operator equation or optimization problem involving Poisson–Nernst–Planck (or related) models for the flow of electrical charge through the channel and to regularize it either by using an iterative method with appropriate stopping criterion and/or additional penalization of the objective functional. This regularization is necessary to compute numerical solutions in a stable and robust way, since the inverse problem is ill-posed in the sense that small differences in the electrical current can correspond to arbitrarily large differences in the permanent charge. In the context of identification, regularization methods allow computation of a stable approximation to the permanent charge in the channel. In the context of design, they also allow us to introduce a priori ideas of suitable designs.

Inverse problems arise whenever one searches for causes of desired or observed effects. Two problems are called inverse to each other if the formulation of one problem involves the solution of the other one. At first sight, it might seem arbitrary which of

these problems is called direct and which is called inverse. Usually, the direct problem is the more classical one. But there is an intrinsic mathematical reason to call one problem "inverse," namely the fact that it is usually ill-posed (cf. [EHN96]). When dealing with partial differential equations, the direct problem usually predicts the evolution of the described system from knowledge of its present state and the governing physical laws including information on all physically relevant parameters. A possible inverse problem would be to compute (some of) the parameters from observations of the evolution of the system; this particular inverse problem is called "parameter identification" and is usually ill-posed (cf. [CER90, EHN96, ER95, IS05, Na06]). We shall highlight the ill-posedness of the inverse problem in a simplified setup, which we nonetheless expect to capture the essential features of the problem, and we also discuss *identifiability*, i.e., the question of whether the unknowns in the inverse problem are determined uniquely from the data.

Regularization methods are needed to overcome these instabilities and to design solution techniques that are robust (i.e., that are stable with respect to data and numerical errors). In general terms, regularization methods replace an ill-posed problem by a family of neighboring well-posed problems. We perform this task for design and identification problems in ion channels. In addition to the stable approximation, the regularization methods are also used to introduce a priori knowledge about the ion channel structure. In a case study of an L-type Ca channel, we present various numerical results, which demonstrate the feasibility of our approach and highlight some particular issues that are likely to appear in channel and protein problems.

2. Modeling ion channels. In the following, we give a brief overview of continuum models of ion transport through channels. Such models need to incorporate the electrostatic interaction between the charged particles, the change of charge density by the mobile ions and a consequent change of the electric field, the generation of ion flux by the electric field, and the direct electrochemical interactions between the ions. Here we shall detail and use Poisson–Nernst–Planck (PNP) models, where the unknowns are the electric potential V and the densities ρ_k of the various (ionic) species present in the channel. Continuum models of this sort have received much attention in the literature (cf. [CE93, GNE02, GNE03, IR02a, IR02b, Maetal03, NCE00]) as well as criticism, mostly because they neglect correlations produced by the small number of ions that can fit into a single channel. Correlations can be included in the derivation of PNP (cf. [SNE01, NHE03]), and certain types of correlations can be analyzed and included in generalizations of PNP with some success (cf. [GNE02, GNE03, Maetal03, SNE01, XWGM06]). Much more work is needed in this regard, and it remains to be seen how well extensions of PNP can deal with the entire set of correlations present in particle-based simulations. Despite similar limitations, continuum models are used very widely in many fields: For example, in computational electronics, continuum models are widely used because they are typically much faster than particle-based simulations (cf. [Se84, JaLu89]).

In our work, we use the extended form of PNP as found in [GNE02, GNE03], understanding that we will need to refine and replace this model as it is improved. The following sections show how our inversion approach can easily be updated to possibly improved forward models.

We assume that the total number of different species is M, but we distinguish between the free species and the species confined to the channel, which create the permanent charge of the channel. For simplicity, we restrict ourselves to a single confined species, denoted with index M, but extensions to multiple confined species

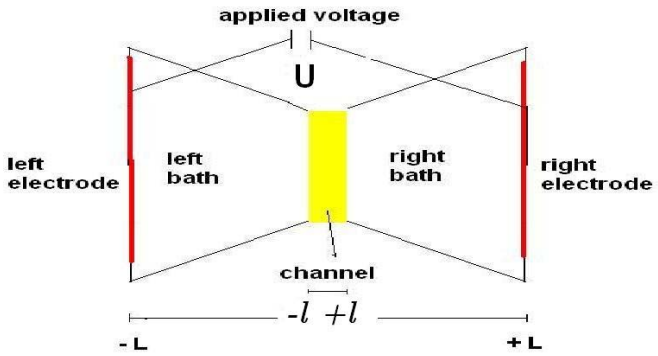


FIG. 1. *Two-dimensional sketch of the computational domain* Ω *modelling the bath-channel system.*

are possible. We mention that selectivity in an ion channel can occur only if $M \geq 4$, since one needs at least two free species with charges of the same sign in order to have selectivity of one over the other, as well as a confined species (permanent charge) and a free species of opposite sign to achieve charge neutrality in the bath. Since the bath and channel in practice always includes water, the number of densities should satisfy $M \geq 5$. Indeed, in the case of an L-type Ca channel we study in further detail below, the number of species is exactly equal to five.

The concentrations have to be computed in a domain Ω that describes the bath and channel. A schematic setup of Ω is depicted in Figure 1.

The electric potential is computed from the Poisson equation with a source term equal to the charge generated by the ions, including the permanent charge. For the continuum description of ion transport, the Nernst–Planck (NP) equations are used, which involve a diffusion term as well as a drift term caused by the electric field (ideal electrostatic potential), an external confining potential, and the excess electrochemical potential. A computational model is a coupled system of the form (after suitable scaling)

$$-\lambda^2 \Delta V = \sum_k z_k \rho_k, \tag{2.1}$$

$$-\nabla \cdot (m_j \rho_j \nabla \mu_j[\rho_1, \ldots, \rho_M; V]) = 0, \qquad j = 1, \ldots, M. \tag{2.2}$$

Here z_k denotes a relative charge of the kth species, m_j is the mobility, and λ is a scaled variable depending on the dielectric coefficient, elementary charge, and typical values of the concentrations ρ_k. The potentials μ_k are computed as variations of an energy functional, i.e.,

$$\mu_k = \frac{\partial}{\partial \rho_k} E[\rho_1, \ldots, \rho_M; V], \tag{2.3}$$

which is of the form

$$E[\rho_1,\dots,\rho_M;V] = \int_\Omega \left(-\lambda^2|\nabla V|^2 + \sum_k (z_k V\rho_k + c_k\rho_k \log\rho_k + \mu_k^0\rho_k) \right) dx + E^{ex}[\rho_1,\dots,\rho_M]. \tag{2.4}$$

The functional E includes electrostatic interaction via the electric field (the first two terms), diffusion (the logarithmic term), external forces via potentials μ_k^0, and direct electric and chemical interactions. Note that the Poisson equation (2.1) can be seen as an equilibrium condition for this energy, i.e.,

$$0 = \frac{\partial}{\partial V} E[\rho_1,\dots,\rho_M,V]. \tag{2.5}$$

Besides the specific exchange terms in energy and potentials, the PNP equations (2.1), (2.2) are a standard model for electrodiffusion of charged species (cf. [Ru90]), which has well-known applications to semiconductors (cf. [VR50, MRS90]). A major difference between electrodiffusion of ions and semiconductors is that it is easy to control the concentrations of the different species in the bath independently of the applied potential, while it is not easy (or even usually possible) to control the concentration of holes or electrons independent of the contact potential. Boundary conditions for the ion channel problem are of the form

$$\begin{aligned}
&V = U && \text{on } \Gamma_D, && \\
&\rho_j = \eta_j && \text{on } \Gamma_D, && j = 1,\dots,M-1, \\
&\frac{\partial \mu_M}{\partial n} = 0 && \text{on } \Gamma_D, && \\
&\frac{\partial V}{\partial n} = 0 && \text{on } \Gamma_N, && \\
&\frac{\partial \mu_j}{\partial n} = 0 && \text{on } \Gamma_N, && j = 1,\dots,M.
\end{aligned} \tag{2.6}$$

Here the boundary is split into $\partial\Omega = \Gamma_D \cup \Gamma_N$, where Γ_N is the insulated part and $\frac{\partial .}{\partial n}$ denotes the normal derivative. Since there are usually two baths, Γ_D will consist of two separated components, and the boundary values are typically constant on each component. The potential U (or, rather, the difference of U between the left and right bath) denotes an applied voltage, and η_j are the bath concentrations of the free species, which are constrained by the charge neutrality condition

$$\sum_{j=1}^{M-1} z_j\eta_j = 0. \tag{2.7}$$

Note that the confined species is usually modeled at equilibrium, which is equivalent to the zero flux boundary condition (for the constrained ions) on the whole boundary. The total number of confined particles N_M needs to be specified to determine ρ_M, giving

$$\int_\Omega \rho_M \, dx = N_M. \tag{2.8}$$

The (measured) output of a channel is the current flowing out on one side, given by

$$I = \sum_{k=1}^{M-1} \int_{\Gamma_0} z_k J_k \cdot dn, \tag{2.9}$$

where $\Gamma_0 \subset \Gamma_D$ is one of the connected components of Γ_D and J_k denotes the flux of species k given by

$$J_k = -\rho_k \nabla \mu_k = -c_k \nabla \rho_k - z_k \rho_k \nabla V - \rho_k \nabla \mu_k^0 - \rho_k \nabla \mu_k^{ex}, \tag{2.10}$$

where the excess potential is defined as $\mu_k^{ex} = \frac{\partial E^{ex}}{\partial \rho_k}$. The current can also be measured and computed from the charge induced on surrounding (Dirichlet) boundaries using the Shockley–Ramo theorem (cf. [NPGE04]).

We mention that the nondimensionalization and scaling of (2.1), (2.2), (2.6) can be performed in an analogous way to the drift-diffusion model for semiconductors (cf. [MRS90]), and for typical values one also has to expect that λ is small; i.e., the Poisson equation (2.1) becomes a singularly perturbed problem.

The system just described has to be coupled to some model for the excess potentials. The excess electrochemical potentials (obtained as variations of the excess energy with respect to the particle densities) include the direct interactions between the ions, usually obtained from hard-sphere or Lennard-Jones models. The external confining potential describes the external forces produced by the structure of the channel on the ionic groups of the protein that make up the permanent charge. This confined permanent charge produces the selectivity of the channel. For our test computations detailed below, we use a specific model of the other components of the excess potential based on density-functional theory (DFT), as described in [GNE02, GNE03, NCE00]. Other models of the excess electrochemical potential require similar computational schemes and lead to the same kind of inverse problems. For a detailed statement of all equations used in the computation of the excess potentials we refer to the appendix of [BEE06].

3. Inverse problems in ion channels. As in many inverse problems, we consider two classes of inverse problems in ion channels, which have different practical motivations:

- *Identification problems* consist in determining properties of a "real" channel (permanent charge and structure), given measurements of the channel output (the total current, in a standard experimental setting) at various different conditions (applied voltages, bath concentrations of the ions).
- *Design problems* consist in determining properties of a "synthetic" channel—either a modification of a natural channel (cf. [Mietal06]) or an abiotic analogue of a biological channel (cf. [Sietal06])—such that optimal characteristics are obtained with respect to some criterion (e.g., selectivity with respect to certain ion species). The medical and technological effects of improved selectivity can be very important. For example, improving Ca selectivity in the L-type Ca channel (by using a drug that changes permanent charge in a way mathematics suggests, if such a drug can be made) would be medically relevant.

The unknowns to be identified or designed are related to the permanent charge, i.e., the ion species confined to the channel. First, an important number is the total

amount of permanent charge, i.e., the number N_M of charged particles confined to the channel. A second important quantity determining the permanent charge is the external confining potential μ_M^0, which represents the forces acting on the permanent charge and encodes the channel structure. In the absence of an electrical field and of electrochemical interaction with other ions, the permanent charge density is given by

$$\rho_M = \gamma_M N_M \exp\left(-\mu_M^0 / z_k\right) \tag{3.1}$$

with a constant γ_M determined from the condition (2.8). Hence, the number N_M and the confining potential μ_M^0 determine the permanent charge density and, subsequently, the selectivity properties of the channel. If the sensitivity of the permanent charge density ρ_M with respect to voltages and bath concentrations in the measured range appears to be negligible, one can also try to directly infer ρ_M from the measurements, ignoring the NP equation for ρ_M. The total charge N_M is a single positive number for which a lower bound (zero) and an upper bound (since too large permanent charges would destroy the channel) are available, and thus it could even be determined by sampling all its possible values. The ill-posedness plays no significant role in the determination of N_M. The confining potential μ_M^0 (and also the density ρ_M as an alternative) is a function of space, so that the inverse problem of determining the confining potential is infinite-dimensional. Since ill-posedness in the sense of discontinuous dependence on data arises only for infinite-dimensional problems and numerical instability becomes more severe as the number of unknowns/design parameters in the inverse problems increases (cf. [EHN96]), instability effects are expected to be more significant for determining the confining potential than for determining the total charge. As a consequence of the ill-posedness, suitable regularization methods have to be used to compute stable approximations of the confining potential, as explained in the previous sections. In the following, we will describe the computational solution of the inverse problems of determining total charge and confining potential in detail, both in the cases of identification and of design.

3.1. Identification. The aim of the identification problem is to find the total charge and/or the confining potential from measurements of the outflow current I taken at different bath concentrations η_j (boundary values of the densities ρ_j) and at different applied voltages U (boundary values of the electric potential V). The measured current I is one real number for each combination of voltage and bath concentrations. In general, I can be seen as a functional of voltage and bath concentrations. The underlying forward model creates a relation between the input P and the output I, which can be modeled via a nonlinear operator $F : P \mapsto I$ between function spaces. Note that the evaluation of the operator F for a specific value of P involves the solution of forward problems with given P for each combination of voltage and bath concentrations (in the idealized setting an infinite number of forward problems). In this setup, the identification problem can be formulated as the operator equation

$$F(P) = I^\delta, \tag{3.2}$$

where I^δ denotes the noisy version of the current obtained from measurements.

We mention that this identification problem has many similarities to the identification of *doping profiles* (i.e., permanent charges) in semiconductor devices from electrical measurements, a problem which has been investigated in detail previously (cf. [BEMP01, BEM02, BELM04, LMZ06, Wo06, WB06]). Since the underlying differential operators appearing in the forward model are exactly the same, one may

expect similar mapping properties of the forward operator F. In particular, this analogy suggests that the identification problem in ion channels (concerning the permanent charge density or the associated constraining potential) is severely ill-posed, as was found in some cases for semiconductors (cf. [BEMP01]). Below we shall also provide analytical arguments for a simplified model and numerical ones for the full model confirming the severe ill-posedness.

We also want to highlight some important differences between the identification problem for the permanent charge of ion channels and the already known identification of doping profiles in semiconductors. First, the forward models include additional effects such as the higher number of species, the excess electrochemical potentials, the different boundary conditions, and the model for the permanent charge density depending on the constraining potential. The second and most important difference is the amount of data that can be used. For semiconductors, only the voltage can be varied, but the boundary concentrations (of electrons and holes) are fixed. As a consequence, the amount of data is not enough to produce a unique solution of the inverse problem using current measurements (cf. [BEM02, Wo06]). On the other hand, one can also measure capacitances in the case of semiconductors (i.e., variations of the total charge with respect voltage change), which can significantly improve the quality of reconstructions in the case of unipolar devices (cf. [Wo06]), but measurements of nonlinear capacitance in biological systems are not analogous (cf. [BeSt98]). However, even with additional capacitance measurements, there are examples of nonuniqueness for the identification of doping profiles in bipolar devices due to an inherent antisymmetry caused by the special boundary values in semiconductors (cf. [Wo06]). Boundary values cover a wide range in an ion channel, and so this antisymmetry is broken, and uniqueness in the identification becomes more likely.

For semiconductors, it has already been shown that very demanding problems such as the inverse conductivity problem with a measured Dirichlet-to-Neumann map arise as special cases, and the full inverse dopant profiling is even more complex (cf. [BEMP01]). Since the measured currents and capacitances are functions of a single variable—the voltage—in semiconductors, the evaluation of the corresponding forward map F involves significantly fewer numerical solutions ("solves") than in the case of ion channels, where the PNP system (2.1), (2.2) has to be solved for varying bath concentrations as well as voltage. Consequently, the computational complexity of the identification problem is even higher for ion channels and seems to be one of the most challenging inverse problems with respect to this issue. Because of the high number of solves of the PNP system, it is of fundamental importance to use efficient numerical schemes for the forward problem. Here we use a mixed finite element scheme with a novel symmetric linearization, which allows an efficient and robust solution of PNP systems with input parameters that cover a wide range of values (cf. [BW07]). The fact that currents are measured for many different setups in ion channel experiments is of crucial importance for the quality of reconstructions. Since the data set is richer than for semiconductors, one can actually achieve more ambitious goals in the inverse problem, e.g., unique reconstruction of the permanent charge density as a function of space (as we shall see below).

3.2. Design. The general remarks and notation of section 3.1 are also valid here. However, in the case of (optimal) design, there is an objective to be achieved instead of an object to be determined. In the applications to ion channels we have in mind, the primary objective is always to increase selectivity of one species over another. As discussed in detail in [GE02], selectivity has to be defined by experimental

results, and several different selectivity measures are available. A selectivity measure S_j of a species can be defined as a functional of ion densities and fluxes (possibly at varying voltage; cf. [GE02]). Since the densities and fluxes depend implicitly on the unknowns P related to the permanent charge (total number of charges or the constraining potential), the selectivity measure can also be rewritten as a functional $S_j = S_j(P)$ of these parameters. If the aim is to increase selectivity of species a over b, then one can minimize a relative selectivity measure

$$Q(S_a(P), S_b(P)) \to \min_P. \tag{3.3}$$

A simple widely used choice which we also use in our computational experiments is the selectivity quotient $Q(S_a, S_b) = -\frac{S_a}{S_b}$ (note that minimizing the negative quotient is equivalent to the original aim of maximizing the relative sensitivity). Analogous treatment is possible for other choices of Q, e.g., $Q(S_a, S_b) = \frac{S_b}{S_a}$ or $Q(S_a, S_b) = -S_a + S_b$.

In practice, to achieve a design task does not mean to actually maximize the functional Q, but usually one is satisfied if a significant improvement with respect to the criterion described by Q, e.g., the channel selectivity, is achieved.

The optimal design problem shares many of the problems of instability and ill-posedness with the identification problem. In the optimal design problem, however, there are no input data, but only a goal to be achieved, so that noise in the input data is not relevant. However, if one minimizes a functional Q as part of the solution of the design problem, as was done previously in the identification problem, then first of all the minimizer might not exist, which means that the norm of P tends to infinity in the associated minimization algorithm. Even if a minimizer exists, it might not be robust with respect to small perturbations of the problem (modeling errors, numerical errors, small changes of applied voltage and concentrations, etc.), so that a computed solution becomes useless in practice. Due to these instabilities, we have used regularization approaches to solve the design problem similar to those used for the identification problem (section 5).

We finally mention that optimal design problems for PNP systems have also been investigated before in semiconductor applications (cf. [HP02a, HP02b, BP03]), but again there are many significant differences in applications to ion channels. Besides all the differences in the forward problem, the optimal design of semiconductors (and, in particular, the objective functional) is always related to currents. In semiconductors, only holes and electrons carry charge, and so there is no analogue to selectivity measures of ion channels. Hence, the optimal design task for ion channels is a quite new problem that connects only loosely to previous literature.

4. Analysis of a simplified model. In order to obtain further insight into the structure of the inverse problems, we study a simplified model case for a spatially one-dimensional setup, i.e., $\Omega = (-L, L)$, with the channel being the subregion $(-\ell, \ell)$. We ignore all direct interactions; i.e., we set $E^{ex} \equiv 0$, and, moreover, we set $\mu_k^0 \equiv 0$. Hence, we arrive at the one-dimensional PNP model

$$-\lambda^2 V'' - \sum_{j=1}^{M} z_j \rho_j = 0, \tag{4.1}$$

$$J_k' = 0, \qquad k = 1, \dots, M-1, \tag{4.2}$$

$$J_k - \rho_k z_k V' - c_k {\rho_k}' = 0, \qquad k = 1, \dots, M-1, \tag{4.3}$$

with boundary conditions

$$V(-L) = 0, \qquad V(L) = U, \qquad \rho_k(\pm L) = \eta_k^{\pm}, \qquad k = 1, \dots, M-1.$$

The equations simplify after an exponential transform to a new set of variables (also called *Slotboom variables*; cf. [MRS90]) $u_k = e^{-\beta_k V}\rho_k$, where $\beta_k = -\frac{z_k}{c_k}$. We obtain

$$-\lambda^2 V'' - \sum_{j=1}^{M-1} z_j e^{\beta_j V} u_k = z_M \rho_M, \tag{4.4}$$

$$J_k' = 0, \qquad k = 1, \dots, M-1, \tag{4.5}$$

$$J_k - c_k e^{\beta_k V} u_k' = 0, \qquad k = 1, \dots, M-1, \tag{4.6}$$

with boundary values $V(-L) = 0$, $V(L) = U$, $u_k(-L) = \eta_k^-$, and $u_k(L) = e^{-\beta_k U}\eta_k^+$. Starting from this transformation, (4.5) and (4.6) can be integrated to obtain the solution

$$\rho_k(x) = \left(\eta_k^- + (e^{-\beta_k U}\eta_k^+ - \eta_k^-)\frac{G_k(x)}{G_k(L)}\right) e^{\beta_k V(x)} \tag{4.7}$$

with the function

$$G_k(y) := \int_{-L}^{y} e^{-\beta_k V(x)}\, \mathrm{d}x.$$

Inserting the explicit solution for the concentrations from the NP equations into the Poisson equation, we obtain a single nonlinear integro-differential equation for the electric potential as

$$-\lambda^2 V'' - \sum_{k=1}^{M-1} \mathcal{R}_k[V] = z_M \rho_M \tag{4.8}$$

with the nonlinear operators $\mathcal{R}_k$ given by

$$\mathcal{R}_k[V](x) = z_k \left(\eta_k^- + (e^{-\beta_k U}\eta_k^+ - \eta_k^-)\frac{G_k(x)}{G_k(L)}\right) e^{\beta_k V(x)}. \tag{4.9}$$

The fluxes J_k can be computed as

$$J_k = \frac{(e^{-\beta_k U}\eta_k^+ - \eta_k^-)}{\int_{-L}^{L} e^{-\beta_k V(x)}\, \mathrm{d}x}. \tag{4.10}$$

Since the fluxes J_k are constant in spatial dimension one (and $J_M = 0$), we obtain the current globally as

$$I = \sum_{k=1}^{M-1} z_k J_k. \tag{4.11}$$

From a computational viewpoint, it seems attractive to consider a setup around (thermodynamic) equilibrium, since the solution of the forward model can be approximated by the simpler linearization around the equilibrium state. An equilibrium situation is obtained if the fluxes of all species vanish, which means in the one-dimensional

setting that $\eta_k^- = e^{-\beta_k U}\eta_k^+$, since the flux of every ionic species vanishes in this case. Note that one can always find suitable combinations of the bath concentrations that satisfy the above equilibrium condition as well as charge neutrality (e.g., vanishing bath concentration will always be an equilibrium case), so that we obtain a family of equilibria, still freely parameterized by the voltage U. This is an important particular feature of PNP systems in channels and will allow us to study some new effects. On the other hand, this also highlights possible redundancy in the data, since there are several parameter combinations that produce zero fluxes and even more that produce zero current, i.e., data without information content for the inverse problem.

The equilibrium electric potential parameterized by U will satisfy

$$-\lambda^2 V_{0,U}'' - \sum_{k=1}^{M-1} z_k \eta_k^- e^{\beta_k V_{0,U}} = z_M \rho_M, \qquad V_{0,U}(-L) = 0, V_{0,U}(L) = U. \tag{4.12}$$

Since β_k and $-z_k$ have the same sign and since η_k^- is nonnegative, the nonlinear terms $z_k\eta_k^- e^{\beta_k V_{0,U}}$ in the Poisson equation depend monotonically on $V_{0,U}$, so that the existence and uniqueness of the solution can be seen easily, as well as the stable dependence on ρ_M.

Now consider the linearization of the problem around the equilibrium values of $\eta_k^\pm$, i.e., the first-order change (in ϵ) of the output I with respect to perturbations of the form $\eta_k^\pm + \epsilon\hat{\eta}_k^\pm$ that still satisfy charge neutrality. The first-order expansion of the integral term in (4.10) disappears, since the numerator vanishes at equilibrium, and hence the linearized output is given by

$$\hat{I}(U) = \sum_{k=1}^{M-1} z_k \frac{(e^{-\beta_k U}\hat{\eta}_k^+ - \hat{\eta}_k^-)}{\int_{-L}^{L} e^{-\beta_k V_{0,U}(x)}\,\mathrm{d}x}.$$

We mention that the use of $\hat{I}_U$ instead of I produces only a restriction of the data set. It is not a simplifying assumption because $\hat{I}_U$ can be computed from the measurements of currents I in a full range of parameters around their equilibrium values.

Now assume that $M \geq 4$, so that we have at least three different mobile species. Then one can always find values $\hat{\eta}_k^\pm$ satisfying charge neutrality such that $\hat{\eta}_k^- = e^{-\beta_k U}\hat{\eta}_k^+$, $k \neq m$, and $\hat{\eta}_m^- \neq e^{-\beta_m U}\hat{\eta}_m^+$ for some $m \in 1, \ldots, M-1$. Hence, for this choice,

$$\hat{I}(U) = z_m \frac{(e^{-\beta_m U}\hat{\eta}_m^+ - \hat{\eta}_m^-)}{\int_{-L}^{L} e^{-\beta_m V_{0,U}(x)}\,\mathrm{d}x},$$

and therefore one can directly infer the knowledge of $M(U) = \int_{-L}^{L} e^{-\beta_m V_{0,U}(x)}\,\mathrm{d}x$ from the knowledge of $\hat{I}_U$. Since the equilibrium Poisson equation can be solved uniquely for fixed U and given ρ_M, the forward map can be related to a (nonlinear) integral operator, and the identification of the permanent charge density corresponds to a nonlinear integral equation of the first kind (the unknown appears only under the integral sign), which is a classical ill-posed problem (cf. [En97, Gr84]).

The analysis particularly simplifies for the equilibrium case of small bath concentrations, i.e., a perturbation of $\eta_k^\pm \equiv 0$. In this situation, we can compute $V_{0,U} = V_{0,0} + \frac{x+L}{2L}U$, where $V_{0,0}$ solves

$$-\lambda^2 V_{0,0}'' = z_M \rho_M, \qquad V_{0,0}(\pm L) = 0.$$

Noticing that there is a one-to-one dependence between ρ_M and the function $f := e^{-\beta_m V_{0,0}}$, we can rephrase the integral equation as

$$\tilde{M}(\sigma) = \int_{-L}^{L} e^{-\sigma x} f(x)\, \mathrm{d}x$$

with $\tilde{M}(\sigma) = e^{\frac{\beta_m U}{2}} M(U)$ and $\sigma = -\frac{\beta_m}{2L} U$. Varying the voltage U in an interval $(-U_{max}, U_{\max})$ is then equivalent to varying $\sigma \in (-\sigma_{max}, \sigma_{max})$. Hence, we arrive at a Fredholm integral equation of the first kind for f, with an analytic kernel, a problem which is known to be severely ill-posed (see the analysis below). The standard classification of ill-posedness we refer to divides into mildly ill-posed problems with an error amplification that grows like a polynomial with increasing frequency and severely ill-posed problems with faster growing error amplification (usually exponentially). The remaining step of computing ρ_M from f is another nonlinear ill-posed problem, which involves the application of a logarithm and two differentiations to compute

$$\rho_M = \frac{\lambda^2}{z_M \beta_m} (\log f)''$$

and is therefore mildly ill-posed.

Identifiability, i.e., uniqueness of the reconstruction from the given data set, can be guaranteed in this case independent of the size of U_{max} (respectively, σ_{max}). Assume that $M(\sigma)$ is known in an arbitrarily small interval around $\sigma = 0$; then, in particular, all derivatives

$$(-1)^p \frac{d^p}{d\sigma^p} \tilde{M}(0) = \int_{-L}^{L} x^p f(x)\, \mathrm{d}x, \qquad p = 0, 1, \ldots,$$

and hence all moments of f are known. Since a function is uniquely determined from its moments, we conclude the uniqueness of the reconstruction of f and subsequently of ρ_M. Note that we have used only a subset of the data to show identifiability, and so one might argue that the full inverse problem is actually overdetermined.

In order to gain some quantitative information about the instability present in the identification problem, we investigate the singular values of the operator

$$K : L^2([-L, L]) \to L^2([-\sigma_{max}, \sigma_{max}]), \qquad f \mapsto \int_{-L}^{L} e^{-\sigma x} f(x)\, \mathrm{d}x.$$

Note that K is a symmetric positive semidefinite operator, and hence the singular values and eigenvalues are equal. As mentioned above, the fact that the integral kernel is analytic implies that the eigenvalues decay faster than any polynomial (cf. [We68]). One actually expects exponential decay. This is confirmed by a numerical computation of the spectrum (with 1025 grid points) displayed in Figure 2, where we plot the singular values (rescaled so that the leading one is equal to one) for different values of σ_{max} and $L = 1$ fixed. (We do not consider the change of L, since its change can be related to the change of σ_{max} by a simple rescaling.) Since the error amplification factor at each frequency equals the inverse of the singular value, this problem is indeed severely ill-posed. The influence of the maximal value σ_{max} is seen by comparing the four results in the figure. For a smaller value of σ_{max}, the decay of singular values is faster, which implies a more significant loss of information. Since σ_{max} is proportional to the maximal value of the applied voltage U, this result shows that one should make measurements at as large a voltage as possible to reduce the instability in the reconstruction as much as possible.

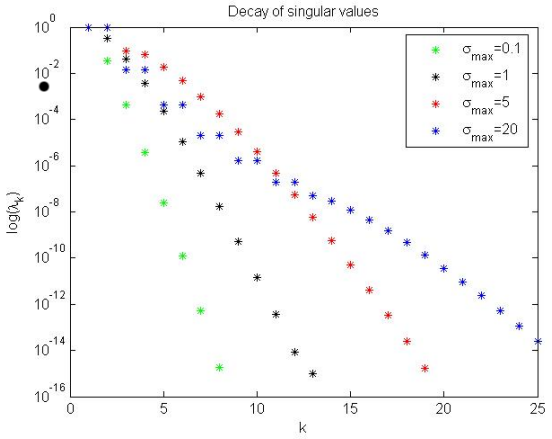


FIG. 2. *Leading singular values of the linear operators K for different values of σ_{max}.*

5. The full inverse problems. In the following, we shall discuss the *forward problem*, namely the solution of the PNP-DFT model for given data, and the map F to the output, namely current-voltage curves for different bath concentrations. This map will be of fundamental use in the mathematical formulation and solution of the inverse problems. As a first step, we analyze the existence and uniqueness of solutions, which hold at least for small bath concentrations of the free species.

5.1. Properties of the forward operators. In the following, we provide an analysis of the PNP model, including the excess free energy. We mention that an extensive analysis of PNP systems is available for applications to semiconductor devices (cf. [MRS90] and the references therein), but the inclusion of the excess free energy in the ion channel model prevents a direct extension of these available results. We shall consider only a particular case of small bath concentrations in the following, in order to make sure that the forward operator can indeed be well-defined at least in some parameter range. To clarify, we state the system we consider, namely the solution of (2.1), (2.10) and, as an equivalent statement of (2.2),

$$\nabla \cdot J_k = 0, \qquad k = 1, \dots, M, \tag{5.1}$$

together with (2.6) and (2.8). In order to show the specific dependence on the applied voltage U, the vector of bath concentrations $\eta = (\eta_k)_{k=1,\dots,M-1}$, the number N_M of confined particles, and the confining potential μ_M^0, we introduce the following nomenclature:

> We denote by $\mathcal{P}(U, \eta; N_M, \mu_M^0)$ the problem of solving (2.1), (2.10), (5.1), (2.6), (2.8) for the unknowns $(V, \rho_1, \dots, \rho_M)$.

We shall assume that $U \in H^{\frac{1}{2}}(\partial\Omega_D) \cap L^\infty(\Omega_D)$ and that E^{ex} is twice continuously differentiable on $H^1(\Omega)^M \cap L^\infty(\Omega)^M$. For the sake of simplicity, we also assume that $E^{ex}(\rho_1, \dots, \rho_{M-1}, \cdot)$ is a convex functional of the last variable if ρ_k, $k = 1, \dots, M-1$, is sufficiently small. Consequently, the map between the density ρ_M and the confining potential μ_M^0 is monotone in this range.

We start our analysis in the case of zero bath concentrations, i.e., $\eta_k = 0$, for $k = 1, \dots, M-1$. In this case, there is obviously no flow, and we can easily construct a solution.

LEMMA 5.1. *Under the above assumptions, there exists a solution*

$$(V, \rho_1, \dots, \rho_M) \in H^1(\Omega)^{M+1} \cap L^\infty(\Omega)^{M+1}$$

of problem $\mathcal{P}(U, 0; N_M, \mu_M^0)$*, which satisfies* $\rho_k \equiv 0$ *for* $k = 1, \dots, M-1$.

Proof. The functions $\rho_k \equiv 0$ satisfy the boundary conditions as well as (2.10), (5.1). We now look for a solution of the remaining problem

$$-\lambda^2 \Delta V = z_M \rho_M, \qquad \rho_M = \gamma_M N_M \exp\left(-\frac{z_M V + \mu_M^0 + \mu_M^{ex}}{c_M}\right),$$

with the boundary conditions remaining for ρ_M and V. Using the monotone dependence of μ_M^{ex} on ρ_M it is straightforward to show that for each $V \in H^1(\Omega) \cap L^\infty(\Omega)$ there exists a unique solution $\rho_M \in H^1(\Omega) \cap L^\infty(\Omega)$ of the second equation. Moreover, the specific exponential dependence on V implies that the map $\mathcal{F} : V \mapsto -z_M \rho_M$ is monotone and continuously Fréchet-differentiable, too. Hence, we can perform a further reduction to a problem of the form

$$-\lambda^2 \Delta V + \mathcal{F}(V) = 0 \quad \text{in } \Omega$$

with Neumann and Dirichlet boundary conditions on the respective parts of $\partial\Omega$. Finally, a standard result for elliptic equations with monotone operators implies existence and uniqueness of this remaining problem (cf. [Sh96]). ∎

In order to proceed to small positive bath concentrations, we shall perform a linearization around zero concentrations. The formal linearization of the PNP system around a given state $(V, \rho_1, \dots, \rho_M)$ is given by

$$-\lambda^2 \Delta \hat{V} - \sum z_k \hat{\rho}_k = f_0, \tag{5.2}$$

$$\nabla \cdot \left(c_k \nabla \hat{\rho}_k + z_k \rho_k \nabla \hat{V} + z_k \rho_k \nabla \left(\sum_j \frac{\partial \mu_k^{ex}}{\partial \rho_j} \hat{\rho}_j \right) + z_k \hat{\rho}_k \nabla (V + \mu_k^0 + \mu_k^{ex}) \right) = \nabla \cdot (\rho_k \nabla f_k), \tag{5.3}$$

$$c_M \frac{\hat{\rho}_M}{\rho_M} + \hat{V} + \sum_j \frac{\partial \mu_M^{ex}}{\partial \rho_j} \hat{\rho}_j = f_M \tag{5.4}$$

with right-hand sides $f_j \in L^\infty(\Omega) \cap H^1(\Omega)$, $j = 0, \dots, M$, to be solved for $\hat{V}$ and $\hat{\rho}_k$. The left-hand side of (5.2), (5.3) is indeed a Fréchet derivative of the left-hand side in the PNP system. We are going to prove that this linearization defines a continuously invertible linear operator $(f_0, \dots, f_M) \mapsto (\hat{V}, \hat{\rho}_1, \dots, \hat{\rho}_M)$ around zero bath concentrations, i.e., for $(V, \rho_1, \dots, \rho_M)$ being the solution of problem $\mathcal{P}(U, 0; N_M, \mu_M^0)$ from Lemma 5.1. The implicit function theorem in Banach spaces (cf. [De85, Theorem 15.1]) then yields the local existence and uniqueness of solutions as well as the well-posedness of the linearized problems for small bath concentrations.

LEMMA 5.2. *Let* $(V, \rho_1, \dots, \rho_M)$ *be the solution of problem* $\mathcal{P}(U, 0; N_M, \mu_M^0)$*, as in Lemma* 5.1*. Then, for any* $f_j \in L^\infty(\Omega) \cap H^1(\Omega), \dots$*, there exists a unique solution*

$$(\hat{V}, \hat{\rho}_1, \dots, \hat{\rho}_M) \in H^1(\Omega)^{M+1} \cap L^\infty(\Omega)^{M+1}$$

of (5.2)*,* (5.3)*, which depends continuously on the data.*

Proof. Due to $\rho_k \equiv 0$ for $k = 1, \dots, M-1$, the NP equations (2.2) simplify to

$$\nabla \cdot \left(c_k \nabla \hat{\rho}_k + z_k \hat{\rho}_k \nabla (V + \mu_k^0 + \mu_k^{ex}) \right) = 0$$

and, in particular, become scalar equations decoupled from the other variables. After a change of variables to $u_k := \hat{\rho}_k \exp(-\beta_k(V + \mu_k^0 + \mu_k^{ex}))$, with $\beta_k = -\frac{z_k}{c_k}$, we obtain the equation

$$\nabla \cdot \left(c_k \exp(-\beta_k(V + \mu_k^0 + \mu_k^{ex}))\nabla u_k\right) = 0,$$

whose well-posedness can be analyzed by standard techniques for elliptic equations due to the absence of convective terms. Using also the equilibrium boundary conditions for $\hat{\rho}_M$, we obtain the remaining problem

$$-\lambda^2 \Delta \hat{V} - z_M \hat{\rho}_M = \tilde{f}_0, \qquad c_M \frac{\hat{\rho}_M}{\rho_M} + \hat{V} + \frac{\partial \mu_M^{ex}}{\partial \rho_M}\hat{\rho}_M = \tilde{f}_M,$$

now with the given right-hand sides $\tilde{f}_0 = \sum_{k=1}^{M-1} z_k \hat{\rho}_k + f_0$ and $\tilde{f}_M = \sum_{k=1}^{M-1}(f_k - \frac{\partial \mu_M^{ex}}{\partial \rho_k}\hat{\rho}_k)$. For the remaining problem to compute $\hat{V}$ and $\hat{\rho}_M$, exactly the same arguments as in Lemma 5.1 apply, so that we can conclude the well-posedness of the linearization. ∎

We now have collected the necessary prerequisites to prove the well-posedness of the problem for small bath concentrations.

THEOREM 5.3. *Let $\|\eta_k\|_{H^{1/2}(\Gamma_D)}$ and $\|\eta_k\|_{L^\infty(\Gamma_D)}$ be sufficiently small. Then, for each $U \in H^{1/2}(\Gamma_D)$, there exists a locally unique solution*

$$(V, \rho_1, \ldots, \rho_M) \in H^1(\Omega)^{M+1} \cap L^\infty(\Omega)^{M+1}$$

of problem $\mathcal{P}(U, \eta; N_M, \mu_M^0)$, and the linearized problem (5.2), (5.3) is well-posed.

Proof. In the lemmas above, we have shown that the problem for $\eta \equiv 0$ is well-posed and its Fréchet-derivative exists with continuous inverse in the respective function spaces. Moreover, the equation operator is Fréchet differentiable, so that we can apply the implicit function theorem in Banach spaces (cf. [De85, Theorem 15.1]) to conclude that a locally unique solution of problem $\mathcal{P}(U, \eta; N_M, \mu_M^0)$ exists around $\eta \equiv 0$ and that the linearized problems are well-posed for small η. ∎

As a direct consequence of the above result, we can verify the well-definedness and even differentiability of the map from the relevant input data related to the permanent charge to the output current.

COROLLARY 5.4. *Let $\|\eta_k\|_{H^{1/2}(\Gamma_D)}$ and $\|\eta_k\|_{L^\infty(\Gamma_D)}$ be sufficiently small. Then, for each $U \in H^{1/2}(\Gamma_D)$, the map*

$$\begin{aligned} G(\cdot; U, \eta): \quad \mathbb{R}^+ \times (H^1(\Omega) \cap L^\infty(\Omega)) &\to \mathbb{R}, \\ (N_M, \mu_M^0) &\mapsto I(U, \eta) = \int_{\Gamma_D} \sum z_k J_k \, d\sigma \end{aligned} \tag{5.5}$$

is well-defined, compact, and continuously Fréchet differentiable.

5.2. Regularization. In practice, one has to discretize the function I of the bath concentrations and voltages, so that one computes only a finite number K of function evaluations, denoted by $I_1, \ldots, I_K$, and the operator F can be written in the form $F = (F_1, \ldots, F_K)$. The evaluation of a single part F_j amounts to a single solution of the forward problem for a specific combination of the bath concentrations and the applied voltage and the subsequent computation of the outflow current from the solution. The linearization is then of the form $F' = (F_1', \ldots, F_K')$, and its adjoint is of the form $F'(P)^* = \sum_{j=1}^K F_j'(P)^*$. Note that the operators F_j are of the form $F_j(P) =$

$G(H(P); U^j, \eta^j)$, where H is the affine linear operator mapping the parameter P to the pair (N_M, μ_0^M). If both N_M and μ_0^M are the unknowns in the inverse problem, then H is just the identity. If one of them is known, then H is the operator mapping the other one to the pair (N_M, μ_0^M). The well-definedness and compactness of the operators F_j and subsequently of F can directly be inferred from Corollary 5.4, and one can even conclude the existence of Fréchet derivatives of F.

Due to the instability of the inverse problems, regularization methods should be used for their solution. One of the most frequently used classes of regularization methods for nonlinear problems is variational methods (cf. [EHN96, EKN89, SV89]), where the inverse problem (3.2) is approximated by the variational problem

$$J_\alpha(P) := \|F(P) - I^\delta\|^2 + \alpha R(P) \to \min_P \tag{5.6}$$

with a suitable regularization functional R (e.g., $R(P) = \|P - P^*\|^2$ for Tikhonov regularization) and a positive real regularization parameter α. An alternative is iterative regularization methods (cf. [KNS06, ES00, OBGXY05]), based on an iteration procedure of the form

$$P_{n+1} = P_n - G_n(F(P_n) - I^\delta) \tag{5.7}$$

with a linear or even nonlinear operator G_n (depending on P_n in general). Such an iterative scheme becomes a regularization method with the appropriate choice of a stopping index n_* at which the iteration is stopped. A common choice of stopping rule—due to its computational simplicity—is the discrepancy principle; i.e., the iteration is stopped when the residual reaches the order of the noise level. We mention that with the properties of the operator F and its linearization F' derived above, the existing theory of variational and iterative regularization methods can be applied (cf. [EHN96, EKN89, KNS06, ES00]) to our case. We can then guarantee the regularizing properties and convergence of the methods we apply to inverse problems in ion channels.

We mention that an analogous iteration method to (5.7) can (and should) be used to solve the variational problem appearing in variational methods. In our test examples detailed below, we carried out a gradient-based method, which is an iteration procedure of the form

$$P_{n+1} = P_n - \tau_n \left[F'(P_n)^*(F(P_n) - I^\delta) + \alpha R'(P_n)\right] = P_n - \tau_n J'_\alpha(P_n) \tag{5.8}$$

which can be interpreted as a minimization method for the variational problem (5.6) or, with $\alpha = 0$ and an appropriate choice of the stopping index, as an iterative regularization method of the form (5.7). Here F', R' denote the derivatives of the operator F and the functional R, respectively, in the appropriate function spaces. Moreover, $F'(P_n)^*$ is the adjoint of the derivative (which is a linear operator between these function spaces).

The simplest, but already quite significant, inverse problem to be solved in this context is to determine the number N_M, characterizing the total permanent charge (i.e., $P = N_M$ in the above setting). As noticed above, this problem is one-dimensional as an inverse problem (although, of course, the direct problem is still a system of partial differential equations), and hence the instability does not appear. Also, this problem is *not* very challenging with respect to the optimization algorithm, which is fortunate because this problem is particularly important biologically. The main issue in the optimization is the evaluation of the functional J_α (respectively, the operator

F) and its derivative, which involves the solution of several forward problems. The derivatives can be computed via the adjoint operator $F_j'(P_n)^*$ or approximated simply by finite differencing, which typically creates a higher computational effort but needs no further implementations than those already used to evaluate the forward operator. Since the aim is to identify a single real number only, it seems reasonable that this is possible for rather low values of K, and indeed our computational experiments indicate that this is possible with high accuracy already for $K = 10$ and even for $K = 5$.

The next level of complexity is the identification of the confining potential μ_M^0 or the identification of the permanent charge ρ_M. By analogy to the simplified problems considered above, we have to expect that these identifications are severely ill-posed so that regularization is of fundamental importance. The computational complexity of this inverse problem is much higher also because a much higher number K of different setups is needed in order to obtain a reasonable reconstruction of the confining potential or the permanent charge density. It is interesting that numerical exploration of the forward problem suggests that the details of the distribution of permanent charge, and thus the details of the constraining potential, are much less important for biological function than the total amount of that charge as long as the charge is of one sign and also is not too small.

Also for the design tasks introduced above, one can define variational regularization methods by just changing the objective functional to $Q + \alpha R$. In our computational tests, we specifically use a variational method of the form

$$Q(S_a(P), S_b(P)) + \alpha \|P - P^*\|^2 \to \min_P, \tag{5.9}$$

where P^* is a favored initial design. In a synthetic ion channel, this a priori guess could introduce additional criteria into the minimization; e.g., P^* can represent a total charge or a confining potential that is easy to manufacture, so that the regularization term would introduce a criterion for the minimizer to be close to easily manufacturable states. In this way, robustness is introduced in the problem, which can also be observed in the results of our computational experiments.

From a computational viewpoint, the minimization of the regularized variational problem (5.9) is an analogous task to the one appearing in identification problems. The main steps are the evaluation of the objective functional (by solving forward problems and subsequently evaluating selectivity measures) and the computations of gradients of the objective functional with respect to P. The latter task can again be carried out by finite differencing, which reduces to additional solves of the forward problem and creates a high computational effort, or by solving appropriate adjoint problems. The total computational effort for solving optimal design problems is usually much less than for solving identification problems, since the selectivity measure is computed only for very few different combinations of bath concentrations and voltages. Significantly fewer forward problems have to be solved for evaluating the objective functional than in the case of identification.

6. Case study: An L-type Ca channel. In this section, we report on a case study performed for an L-type Ca channel (LCC), for which we performed the identification and design tasks as described above. A sketch of the LCC is provided in Figure 3.

We choose the LCC because it is of enormous importance as the regulator of the contraction of skeletal and cardiac muscle, and it has received extensive attention in

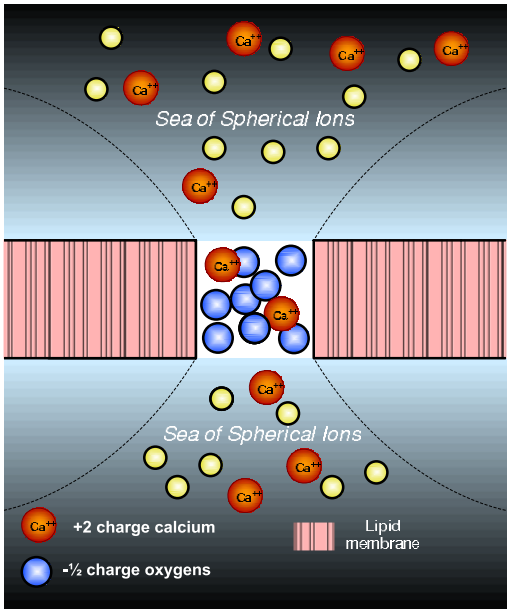


FIG. 3. *Illustration of the LCC when filled with Ca.*

TABLE 1
Parameter settings for the LCC example, using elementary charge $e = 1.602 \times 10^{-19}\,C$.

k	1	2	3	4	5
Species	Ca^{2+}	Na^{+}	Cl^{-1}	H_2O	$O^{-1/2}$
Charge z_k	$2e$	e	$-e$	0	$-\frac{e}{2}$
$\rho_k(L)$	6 mM	12 mM	24 mM	55 M	0 M
$\rho_k(-L)$	var	var	var	55 M	0 M

the biophysics literature for that reason (cf., e.g., [KMS83, Hetal92, SMC03]). Recent work shows quite clearly that many properties of two types of calcium channels can be quantitatively described by extended versions of the PNP model (cf. [Betal06, GNE02, Mietal06, Mietal04, Waetal05, XWGM06]). Given the importance of calcium channels and the demonstrated ability of PNP-type models to explain current voltage relations and selectivity over large ranges of concentration of many types of ions, it is natural to use this system in our investigation of inverse problems.

6.1. Forward model. The forward model of the LCC involves the electrical potential V and five densities ρ_k modeling the three mobile ion species Ca^{2+}, Na^{+}, Cl^{-}, a neutral mobile species H_2O, and half-charged oxygens $O^{-1/2}$ corresponding to the permanent charge. This means that each forward problem consists of a coupled system of six partial differential equations, the Poisson equation (2.1) and five NP equations (2.2) for the densities $\rho_1, \ldots, \rho_5$ (see Table 1 for the assignment of densities to the species).

The channel is modeled as cylindrical with diameter 0.4 nm ($y - z$ plane) and length $2\ell = 1$ nm (x-direction), embedded in two baths both of length 1.7 nm. This yields a total length of 4.4 nm for the system, and therefore the computational domain is chosen as $(-L, L)$ with $L = 2.2$ nm.

From the geometry of the system, it is rather obvious that the flow arises in

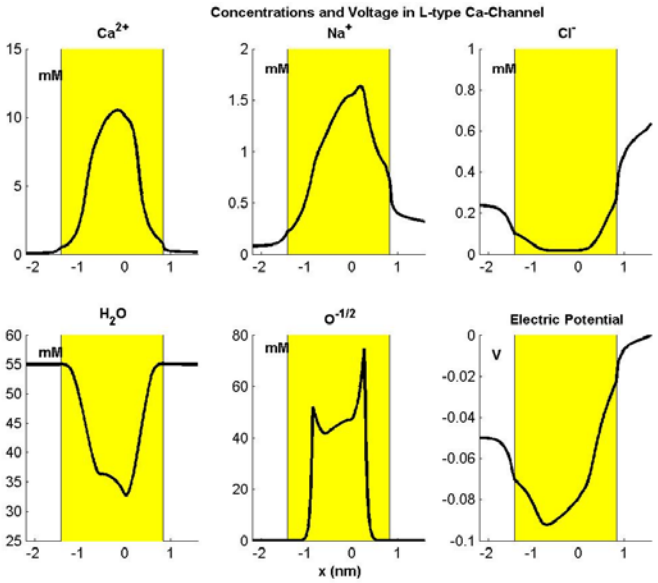


FIG. 4. *Plot of ion densities and electr potential as functions of spatial location for an LCC with applied voltage* 50 *mV. The illuminated region is the channel which is scaled in the* x*-direction by a factor five compared to the bath regions.*

the x-direction, and the model can be reduced by averaging in the $y-z$ plane to a one-dimensional problem with single spatial variable x; but note that our procedures are in no way restricted to the one-dimensional case. In this averaging procedure, the shape of the channel has to be taken into account, which yields some spatially dependent coefficients in the reduced system of one-dimensional differential equations. The details of the averaging and an exact statement of the equations to be solved for the LCC can be found in [GNE02, GNE03, NCE00, BEE06].

We solve the forward problem on a grid with $n = 1251$ (for data generation) and $n = 1000$ cells (for the inverse problem) with a standard conforming finite element discretization of the electric potential and the Poisson equation and a mixed finite element discretization of the continuity equations for the ions. Since we have the electric potential and five different species (Ca^{2+}, Na^+, Cl^-, H_2O, and $O^{-1/2}$), this yields $1252 + 5 \times 1251 = 7507$ degrees of freedom (for data generation), respectively, $1001 + 5 \times 1000 = 6001$ (for the inverse problem) degrees of freedom, in the forward problem.

The measurements are the currents, taken as functions of the voltage and of the left bath concentrations $\rho_k(-L)$ for $k = 1, 2$, whereas the right bath concentrations $\rho_k(L)$ are kept fixed. The water concentration ("osmolarity") is fixed in both baths, and $\rho_5(\pm L) = 0$, because of the confinement of permanent charge to the channel. The concentrations $\rho_3(\pm L)$ are finally determined from the charge neutrality $\sum_k z_k \rho_k(\pm L) = 0$. The parameter settings for the boundary values are given in Table 1, where var means that the values are varied in the identification process.

The solution of the forward model for an LCC with the above settings—applied voltage $U = 50$ mV, $N_M = 8$ confined oxygens, and confining potential μ_M^0 plotted as the exact value in Figure 8—is illustrated in Figure 4. The illuminated region corre-

sponds to the channel, while the white region to the left and right correspond to the bath. In this example, one observes many typical effects, in particular the selectivity properties of the channel. Due to the negative permanent charge (oxygens), there is an attractive electrical force on the positively charged ions (Na and Ca) and a repulsive force on the negatively charged ions (Cl). Moreover, the additional "chemical" forces arising from the finite volume of the ions produce an additional decrease of the densities in the channel region. These excluded volume forces are particularly important because of the narrow cross section of the channel. This decrease in densities can be observed in particular in the plot of the water density, since it is the only force acting on this species. (There are no electrical interactions with water in our system due to neutrality of our model of water.)

6.2. Identification I: Reconstruction of the total charge. In this case, one assumes that the structure of the channel is known, but the total charge of the crowded elements in the selectivity filter is unknown. The inverse identification problem consists of identifying the total charge based on measurements of the total current for different bath concentrations of the ions. As noticed before, the reconstruction of the total charge is the simplest case of an inverse problem for ion channels, so that we expect more accurate results than for the more complicated inverse problems in the sections below.

This inverse problem is a finite-dimensional one. We try only to identify a single real number from a finite number of measurements. As mentioned above, this inverse problem is not ill-posed in the classical sense of inverse problems theory, cf. [EHN96], because of the low dimension. The only possible instability is due to nonlinearity effects, but such effects seemed not to appear in the various computational tests.

For a test of the inverse problem technique, we generated synthetic data for the setup as used in the LCC [GNE03], i.e., a crowded charge consisting of eight half-charged oxygens. This means we solve the forward problem with the finer grid and then compute the resulting currents. Subsequently, we perturb the synthetic measurements by noise and use them as data to solve the inverse problem. (The same technique is also used for the other inverse problems below.) In this way, we have a known reference solution, and we can check to see if the algorithm yields reasonable reconstructions in a stable way.

The reconstructions are carried out by a gradient method for the associated least-squares functional describing the residual. The gradients are approximated by finite differences. This is for illustration only. More efficient ways are possible to approximate the gradient for this and related problems, e.g., via adjoint problems.

In this case, one obtains very accurate reconstructions of the exact total charge even for noisy data and even for a rather low number of measurements, allowing us to deal effectively with this quite significant biological problem. The pessimism of early analysis can be removed if the problem is posed with PNP equations and solved with the methods of inverse problems (cf. [At79, AJ78]); see below. A typical setup consists of three different applied voltages (0.1V, 0V, -0.1 V) and two different concentrations for Na and Ca (2 mM and 4 mM) in the left bath. With all combinations, this gives $3 \times 2 \times 2 = 12$ measured values; i.e., the problem is already overdetermined. An illustration of the reconstruction process in this situation is given in Figure 5. Here the reconstructed mass of the crowded particles (scaled by the mass of the eight half-charged oxygens in the real structure) are plotted versus the number of iterations in the optimization method. In this case, a standard stopping criterion would stop the calculation after some 90 to 100 iterations. (The reconstruction does not change

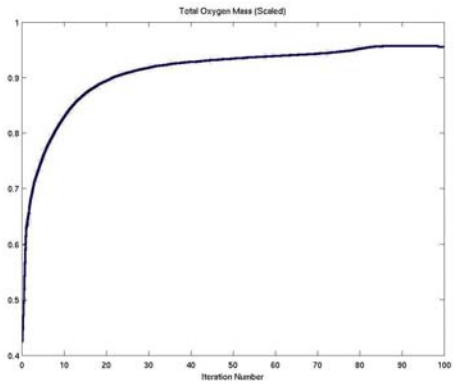


FIG. 5. *Plot of the total charge (relative to the exact value) during the iterations of the gradient method.*

significantly with more computation.) The difference between the scaled mass of the real total charge and the reconstructed one is less than 5%, although the initial value is quite far away from the solution. Similar behavior was found also in other tests with different initial values and parameter settings.

6.3. Identification II: Reconstruction of the structure. The second inverse problem is related to the reconstruction of the structure of the channel. This is done indirectly by identifying the confining potential acting on the crowded ions (oxygens in our example), which models the way the structure interacts with the channel. More specifically, the confining potential models the forces that keep the charged oxygens of the channel inside the selectivity filter.

The unknown in the above setting is given by $P = \mu_5^0$. Now the inverse problem is to find a space-dependent function on the channel region, which is really an infinite-dimensional problem. In an idealized setting, the unique reconstruction of the confining potential (as a function of space) would require an infinite number of measurements. Therefore, any measurement realized in practice (where, of course, only a finite number of measurements can be taken) has to be interpreted as a discretization of the problem with an infinite number of measurements. It therefore seems obvious that a higher number of measurements yields better reconstruction, and this is also confirmed by all our tests. On the other hand, a much higher number of measurements forces an extremely high computational effort.

The variation of the confining potential μ_5^0 has a significant influence only in the channel region, since outside it will just take some very large values that cause the confinement of the permanent charge species. In the solution of the identification problem, we use this a priori knowledge and approximate μ_5^0 by a constant function in the baths. Note that due to the large values of μ_5^0 in the bath regions, the concentration ρ_5 is almost zero there in any case.

As representative examples of the behavior of the reconstructions, we illustrate the results for

(a) four applied voltages and different left bath concentrations for Na and Cl, for a total of $4 \times 2 \times 2 = 16$ measurements (voltages ± 10 mV, ± 5 mV and concentrations 2 mM, 4 mM),
(b) six applied voltages and different left bath concentrations for Na and Cl, for

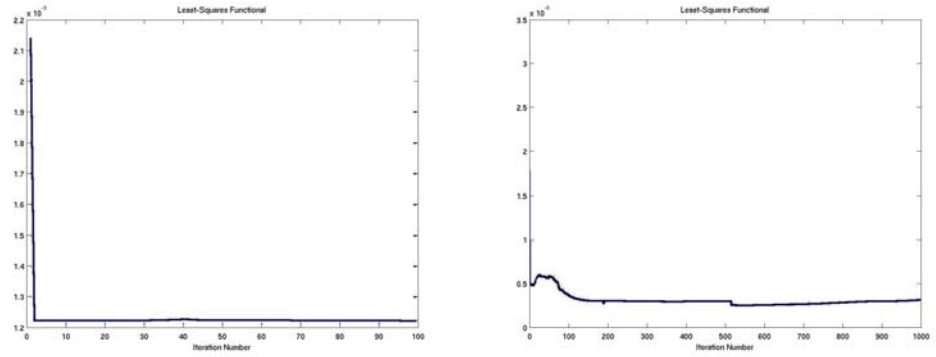


FIG. 6. *Plot of the squared residual* $\|F(P_n) - I^\delta\|2$ *as a function of the iteration number for* $4 \times 2 \times 2$ *measurements (left) and* $6 \times 3 \times 3$ *measurements (right).*

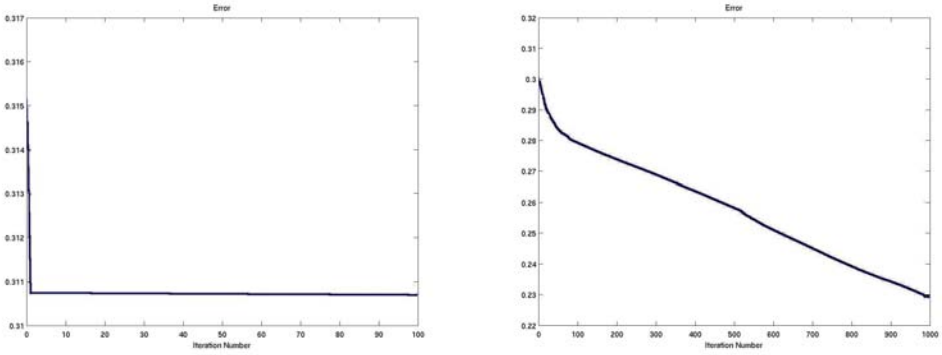


FIG. 7. *Plot of the identification error* $\|P_n - P^\dagger\|$ *as a function of the iteration number for* $4 \times 2 \times 2$ *measurements (left) and* $6 \times 3 \times 3$ *measurements (right).*

a total of $6 \times 3 \times 3 = 54$ measurements (voltages ± 10 mV, ± 6.6 mV, ± 3.3 mV and concentrations 2 mM, 4 mM, 6 mM),

obtained with 0.1% noise. The resulting evolution of the least-squares functional during the iteration is plotted in Figure 6 (left for case (a) and right for case (b))—one observes that they are quite similar in the two cases, and the residual decreases to some value around the size of the noise level. As has to be expected for iterative regularization methods (cf. [EHN96, KNS06]), the evolution of the reconstruction error, however, is completely different, as one can see in the plots of Figure 7 (left for case (a) and right for case (b)). In the first case (16 measurements), the reconstruction error is hardly reduced, while in the second case, one already obtains a very significant decrease before the noise level is reached. This can also be seen from the final reconstructions obtained with a stopping of the iteration dependent on the noise, which are shown (here plotting the negative potentials for illustration purpose) in Figure 8 (left for case (a) and right for case (b)). The initial guess used in both cases is shown in Figure 9. One observes that the second reconstruction is already rather close to the real potential, in particular in the left part of the channel. The reason for the better reconstruction in the left part is that the concentrations are varied in the left bath, and so there is more sensitivity with respect to the data in this region.

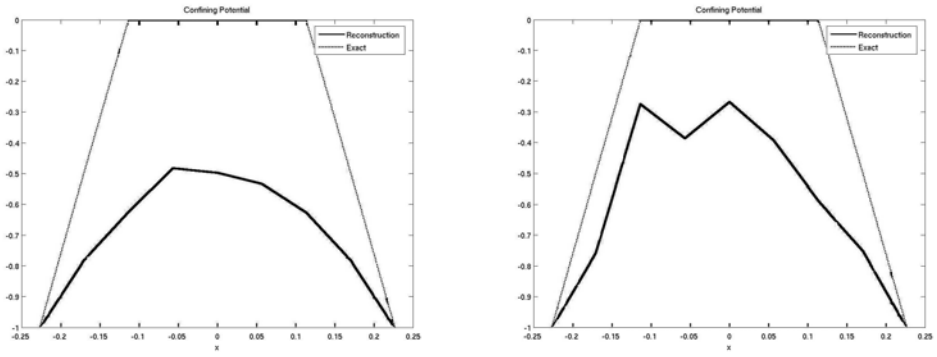


FIG. 8. *Final reconstructions P_{n_*} obtained at the stopping index determined by the discrepancy principle for $4 \times 2 \times 2$ measurements (left) and $6 \times 3 \times 3$ measurements (right).*

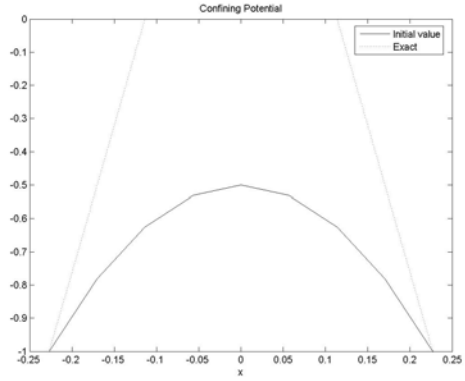


FIG. 9. *Initial value P_0 used for all reconstructions of potentials.*

These results clearly indicate that the reconstructions will improve for an increasing number of measurements. For a very high number of measurements, the computational complexity of the inverse problem dramatically increases and will be necessary to implement very efficient methods to compute reconstructions, including faster forward solvers (cf. [BW07]), adjoint methods for computing derivatives (cf. [GP00]), and multiscale versions of the regularization methods (cf. [Sch98, BM02]).

The instability of the identification problem in this case is illustrated in the plots of Figure 10. Here we use the same setup as before ($6 \times 3 \times 3$ measurements) but a slightly higher noise level (1%). We start with an initial guess where the residual is in the order of the noise level; in such a situation, a stopping rule for an iterative regularization such as the discrepancy principle would immediately stop the iteration. If one iterates further (which one would do when using a standard optimization stopping criterion based on the gradient of the residual), then the error starts to increase (and then possibly oscillates), although the residual is still decreasing. This situation is illustrated in Figure 10, where the least-squares functional and the error are plotted as functions of the iteration number. One observes that in this case the least-squares functional is still decreasing, but the error between the reconstruction and the exact solution can increase, which demonstrates the ill-posedness of the problem. Note that

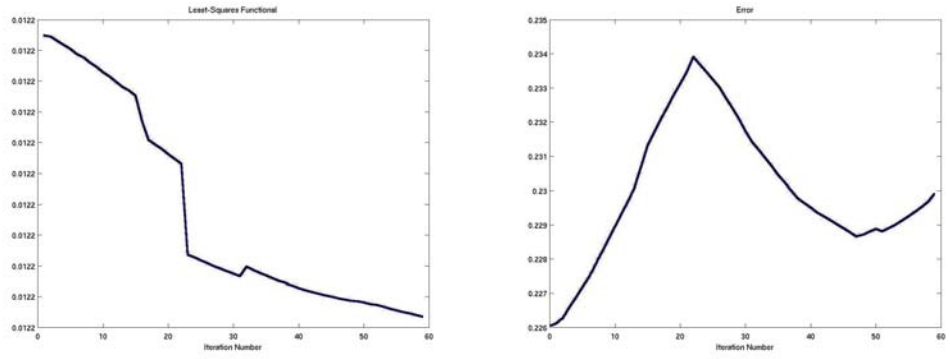


FIG. 10. *Plot of the residual (left) and identification error* $\|P_n - P^\dagger\|$ *(right) as a function of the iteration number without regularizing stopping criterion.*

this effect did not appear in the examples with a stopping criterion based on regularization theory as described in section 5, which again illustrates the importance of regularization.

6.4. Identification III: Reconstruction of the permanent charge density for PNP. As a final step in our study of identification problems, we consider the reconstruction of the permanent charge density ρ_M in a pure PNP model; i.e., the forward model consists in solving (2.1), (2.2), and (2.6) for $k = 1, \dots, M-1$ with given ρ_M and $E^{ex} \equiv 0$. Apart from the elimination of the equation for ρ_M, the discretization and numerical schemes used to solve the PNP system are the same as in the previous section; in particular, we use the Landweber iteration as a regularization method.

In this case, we numerically implemented adjoint solvers to compute derivatives, which results in improved accuracy and lower computational effort even for finer discretizations (in this case, we use 21 grid points) of the unknown in the inverse problem. We refer the reader to [GP00] for a general overview of adjoint methodology and to [BEMP01, Wo06] for the derivation of adjoint problems in related semiconductor applications.

For the reconstruction, we used five different values of the voltage U and eight different bath concentrations of Na and Ca, which results in a total number of $5 \times 16 \times 16 = 1280$ measured values. With this amount of data, very reasonable reconstructions can be obtained even in the presence of noise. The development of the residual and error $\|P_n - P^\dagger\|$ are illustrated in Figures 11 (without noise) and 12 (with noise level 3%). One observes that the residual and error both decrease in a monotone way in the noiseless case, whereas a minimum of the error is reached after some iteration number in the presence of noise. However, at this iteration number the relative residual is already very close to the noise level, so that a stopping rule like the discrepancy principle would stop already slightly earlier. These results have been obtained with an initial value $P_0 \equiv 3$ and an exact value $P^\dagger \equiv 5$, but qualitatively similar results have been computed also with other choices of P_0 and $P^\dagger$.

The quality of the reconstructions is illustrated in Figure 13 for permanent charge $P^\dagger$ being constant (left) and of a sinusoidal shape (right). In both cases, the starting value P_0 is dashed, the exact solution $P^\dagger$ is dotted, and the reconstruction is the solid line. One observes that with the amount of data we use, it is possible to reconstruct the constant solution very accurately, while there remains some visible deviation for

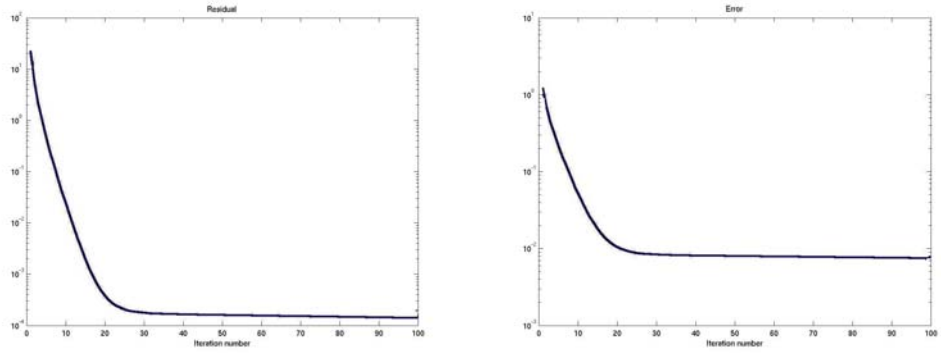


FIG. 11. *Plot of the residual (left) and identification error* $\|P_n - P^\dagger\|$ *(right) as a function of the iteration number in the absence of noise.*

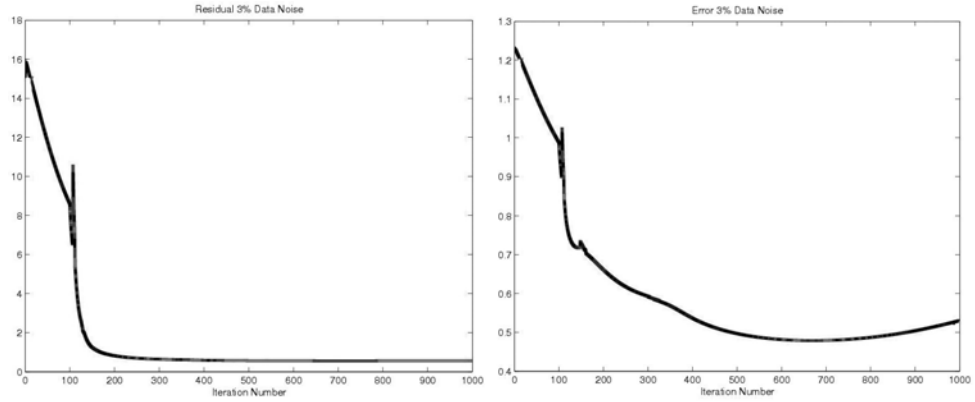


FIG. 12. *Plot of the residual (left) and identification error* $\|P_n - P^\dagger\|$ *(right) as a function of the iteration number for* 3% *data noise.*

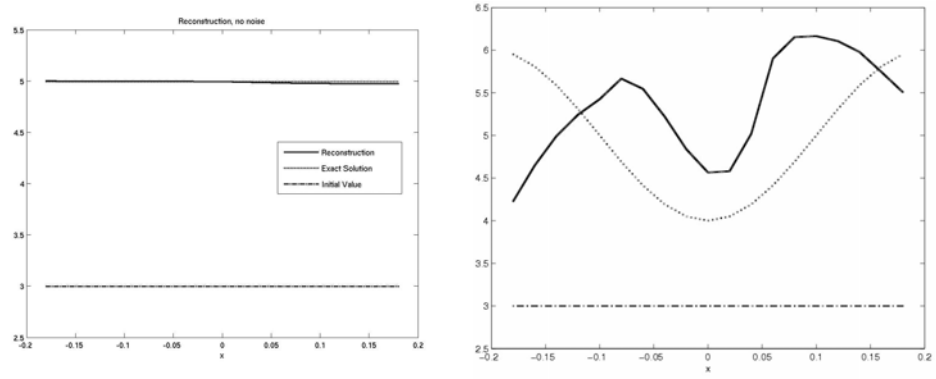


FIG. 13. *Plot of reconstructions in the absence of noise for a spatially constant permanent charge (left) and a spatially varying permanent charge (right).*

the more complicated shape. However, the magnitude of values as well as the principal shape (a valley in the middle) could also be reconstructed in the more difficult sinusoidal second case. Reconstructions of this moderate quality have to be expected in a severely ill-posed problem even for a larger number of measurements, and even these are quite tricky to achieve. Situations in which the permanent charge changes sign or reaches nearly zero are likely to pose even more problems.

6.5. Design: Maximizing selectivity. The final inverse problem we consider is an optimal design problem, which aims at designing in silico channels with optimal sensitivity properties (or at least improved sensitivity compared to a given initial design but possibly also close to this one, which can be used as a constraining criterion).

As a test case, we use one of the three selectivity measures from [GE02], the so-called permeability ratio, at equal concentrations for all ions in the left and right bath. (For this sake, we use the bath concentrations $\rho_k(\pm L) = 20$ mM for $k = 1, 2$ and $\rho_3(\pm L) = 60$ mM.) More precisely, the selectivity measure is the permeability ratio for Na and Ca, where the permeabilities on the right side of the channel are computed (detailed formulas for the computations of the permeabilities S_a are given in the appendix of [BEE06] and in [GE02]). The unknown to be designed is again related to the structure of the channel; i.e., we set $P = \mu_5^0$ and use the same discretization as in the previous section. Since our design goal is to maximize or at least significantly increase the selectivity, we should minimize the negative permeability ratio. It turns out that formulating the negative permeability ratio $-\frac{S_{Na}(P)}{S_{Ca}(P)}$ as the objective functional for selectivity, one ends up with a very unstable problem (which is also expected from the arguments in section 3.2). Moreover, the computed designs seem not really useful for practical construction due to various oscillations. Therefore, we use an additional regularization term as proposed in (5.9),

$$J_\alpha(P) := -\frac{S_{Na}(P)}{S_{Ca}(P)} + \alpha\|P - P^*\|2 \to \min_P, \tag{6.1}$$

where $P = \mu_5^0$ is the confining potential to be optimized and P^* is the favored initial design of the confining potential (the one used in the simulations in [GE02]). Besides its regularizing effect, the second term in the objective functional favors solutions as close as possible to the initial design, which helps to obtain potentials that can be realized in practice.

The objective functional is then minimized with a gradient method and suitable step-size selection to guarantee decrease of the objective, and the gradients are again approximated by finite differences (see above for a discussion of this point).

A special design case (with parameter $\alpha = 200$) is illustrated in the plot in Figure 14 (left), which shows the evolution of the objective functional (black) as well as its first part, the negative permeability (i.e., selectivity) ratio $-\frac{S_{Na}(P)}{S_{Ca}(P)}$, during the iteration until convergence. One observes that an increase in the selectivity measure of more than 100% is achieved by the optimization. The initial value used for the optimization and the final result are plotted in Figure 15. One observes that the two potentials are still very close, and so the structure has not been changed completely.

For comparison (and illustration of instabilities), the classical approach of just minimizing the negative permeability (i.e., selectivity) ratio is also illustrated with the same initial value and parameter settings but with the objective functional $J(P) := -\frac{S_{Na}(P)}{S_{Ca}(P)}$. Again, the right plot in Figure 14 displays the objective functional during the iterations; the optimal solution is plotted in Figure 16. One observes that in this

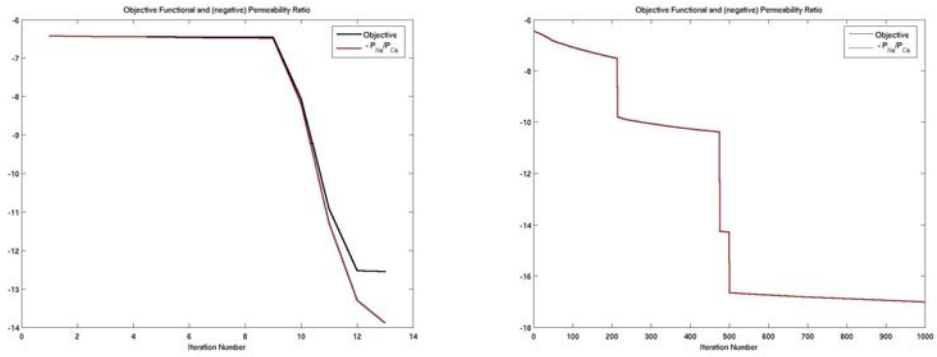


FIG. 14. *Objective functional $J_\alpha(P_n)$ (black) and negative permeability ratio (grey; red online) as a function of the iteration number for $\alpha = 200$ (left) and $\alpha = 0$ (right).*

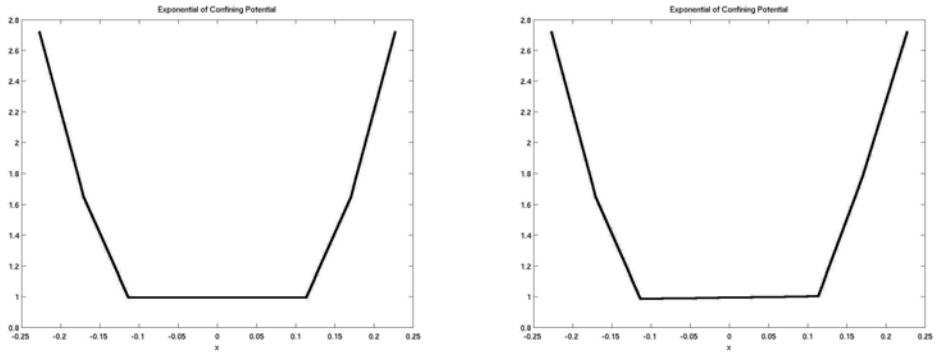


FIG. 15. *Initial value (left) and computed optimal confining potential (right) for the functional J_α with $\alpha = 200$.*

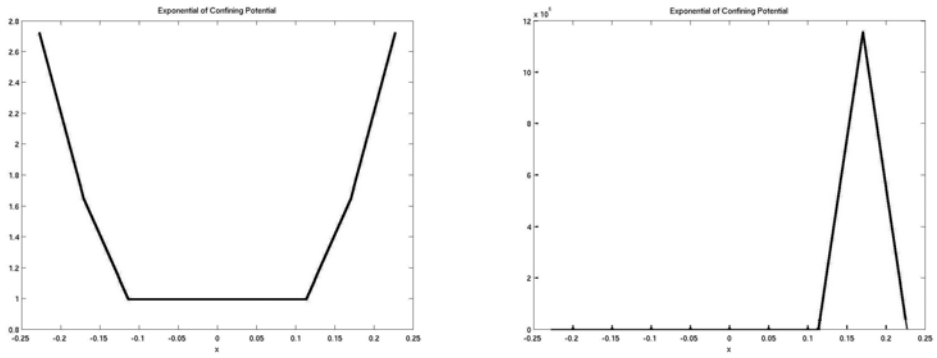


FIG. 16. *Initial value (left) and computed optimal confining potential (right) for the functional J.*

case, the gradient method needs many more iterations than with penalization but does not yield a dramatic increase of selectivity ratio (around 17 instead of 14 for the penalized case). However, just one look at the optimal confining potential in the

unpenalized case (Figure 16) shows that the (small) increase in the ratio is caused by a blowup in the confining potential (notice the vertical scale of 10^6!). Obviously, such extremely high forces will not be easy to realize, and the resulting channel will not be useful in practice, which is another point in favor of our regularization approach. The regularization parameter α can control the balance between increasing the selectivity and "practicability," namely, remaining close enough to the initial design that the new channel can actually be built. If α is very large, then the minimizer of J_α will remain close to the initial guess. For $\alpha \to 0$, the permeability ratio can be increased further, but also the optimal confining potential will increase more and more (until it reaches the one computed for J in the limit). So, regularization gives (in addition to the advantages discussed) even more flexibility in finding a compromise between different design goals.

We summarize by stating that our examples show that both the identification and the design goals can be achieved in a stable and efficient way by our approach based on regularization, as illustrated by the special case using Tikhonov regularization with an iterative minimization of the Tikhonov functional, and that such results are not possible by standard approaches due the ill-posed nature of the inverse problems considered. It will be interesting to see how regularization methods help in the solution of a range of problems in ion channels and proteins.

REFERENCES

[Aletal94] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, Garland, New York, 1994.

[As99] F. M. Ashcroft, *Ion Channels and Disease*, Academic Press, New York, 1999.

[At79] D. Attwell, *Problems in the interpretation of membrane current-voltage relations*, in Membrane Transport Processes, Vol. 3, C. F. Stevens and R. W. Tsien, eds., Raven Press, New York, 1979, pp. 21–41.

[AJ78] D. Attwell and J. J. B. Jack, *The interpretation of membrane current voltage relations: A Nernst-Planck analysis*, Prog. Biophys. Mol. Biol., 34 (1978), pp. 81–107.

[BeSt98] F. Bezanilla and E. Stefani, *Gating currents*, Methods Enzymol., 293 (1998), pp. 331–352.

[Betal06] D. Boda, M. Valisko, B. Eisenberg, W. Nonner, D. Henderson, and D. Gillespie, *Effect of protein dielectric coefficient on the ionic selectivity of a calcium channel*, J. Chem. Phys., 125 (2006), 034901.

[BEE06] M. Burger, B. Eisenberg, and H. W. Engl, *Mathematical Design of Ion Channel Selectivity via Inverse Problems Technology*, U.S. Patent Application, Serial Number 60/791,185.

[BELM04] M. Burger, H. W. Engl, A. Leitao, and P. Markowich, *On inverse problems for semiconductor equations*, Milan J. Math., 72 (2004), pp. 273–314.

[BEM02] M. Burger, H. W. Engl, and P. Markowich, *Inverse doping problems for semiconductor devices*, in Recent Progress in Computational and Applied PDEs, T. F. Chan, Y. Huang, T. Tang, J. A. Xu, and L. A. Ying, eds., Kluwer, Boston, Dordrecht, London, 2002, pp. 39–54.

[BEMP01] M. Burger, H. W. Engl, P. Markowich, and P. Pietra, *Identification of doping profiles in semiconductor devices*, Inverse Problems, 17 (2001), pp. 1765–1795.

[BM02] M. Burger and W. Mühlhuber, *Numerical approximation of an SQP-type method for parameter identification*, SIAM J. Numer. Anal., 40 (2002), pp. 1775–1797.

[BP03] M. Burger and R. Pinnau, *Fast optimal design of semiconductor devices*, SIAM J. Appl. Math., 64 (2003), pp. 108–126.

[BW07] M. Burger and M. T. Wolfram, *Symmetric Discretization of PNP-Systems*, manuscript.

[CE93] D. P. Chen and R. S. Eisenberg, *Charges, currents and potentials in ionic channels of one conformation*, Biophys. J., 64 (1993), pp. 1405–1421.

[CER90] D. Colton, R. Ewing, and W. Rundell, *Inverse Problems in Partial Differential*

Equations, SIAM, Philadelphia, 1990.

[De85] K. Deimling, *Nonlinear Functional Analysis*, Springer, Berlin, 1985.

[Ei90] R. S. Eisenberg, *Channels as enzymes*, J. Memb. Biol., 115 (1990), pp. 1–12.

[Ei98] R. S. Eisenberg, *Ionic channels in biological membranes. Electrostatic analysis of a natural nanotube*, Contemp. Phys., 39 (1998), pp. 447–466.

[En97] H. Engl, *Integralgleichungen*, Springer, Wien, 1997.

[EHN96] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, 1996.

[EKN89] H. W. Engl, K. Kunisch, and A. Neubauer, *Convergence rates for Tikhonov regularization of nonlinear ill-posed problems*, Inverse Problems, 5 (1989), pp. 523–540.

[ER95] H. W. Engl and W. Rundell, eds., *Inverse Problems in Diffusion Processes*, SIAM, Philadelphia, 1995.

[ES00] H. W. Engl and O. Scherzer, *Convergence rate results for iterative methods for solving nonlinear ill-posed problems*, in Surveys on Solution Methods for Inverse Problems, D. Colton, H. W. Engl, A. K. Louis, J. McLaughlin, and W. F. Rundell, eds., Springer, Vienna, New York, 2000, pp. 7–34.

[GE02] D. Gillespie and R. Eisenberg, *Physical descriptions of experimental selectivity measurements in ion channels*, Eur. Biophys. J., 31 (2002), pp. 454–466.

[GNE02] D. Gillespie, W. Nonner, and R. Eisenberg, *Coupling Poisson-Nernst-Planck and density functional theory to calculate ion flux*, J. Phys.: Condens. Matter, 14 (2002), pp. 12129–12145.

[GNE03] D. Gillespie, W. Nonner, and R. Eisenberg, *Density functional theory of charged, hard-sphere fluids*, Phys. Rev. E (3), 68 (2003), 031503.

[GP00] M. B. Giles and N. A. Pierce, *An introduction to the adjoint approach to design*, Flow, Turbulence and Combustion, 65 (2000), pp. 393–415.

[Gr84] C. W. Groetsch, *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Pitman, London, 1984.

[Hetal92] S. H. Heinemann, H. Terlau, W. Stuhmer, K. Imoto, and S. Numa, *Calcium channel characteristics conferred on the sodium channel by single mutations*, Nature, 356 (1992), pp. 441–443.

[Hi01] B. Hille, *Ionic Channels of Excitable Membranes*, Sinauer Associates, Sunderland, MA, 2001.

[HP02a] M. Hinze and R. Pinnau, *Optimal control of the drift-diffusion model for semiconductor devices*, in Optimal Control of Complex Structures, K. H. Hoffmann, I. Lasiecka, G. Leugering, and J. Sprekels, eds., Birkhäuser, Basel, Berlin, 2002, pp. 95–106.

[HP02b] M. Hinze and R. Pinnau, *An optimal control approach to semiconductor design*, Math. Models Methods Appl. Sci., 12 (2002), pp. 89–107.

[IR02a] W. Im and S. Roux, *Ion permeation and selectivity of OmpF porin: A theoretical study based on molecular dynamics, Brownian dynamics, and continuum electrodiffusion theory*, J. Mol. Biol., 322 (2002), pp. 851–869.

[IR02b] W. Im and S. Roux, *Ions and counterions in a biological channel: A molecular dynamics simulation of OmpF porin from Escherichia coli in an explicit membrane with* 1 *M KCl aqueous salt solution*, J. Mol. Biol., 319 (2002), pp. 1177–1197.

[IS05] V. Isakov, *Inverse Problems for Partial Differential Equations*, 2nd ed., Springer, New York, 2005.

[JaLu89] C. Jacoboni and P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation*, Springer, New York, 1989.

[KNS06] B. Kaltenbacher, A. Neubauer, and O.Scherzer, *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*, manuscript, Johannes Kepler University, Linz, Austria, 2007.

[KMS83] P. G. Kostyuk, S. L. Mironov, and Y. M. Shuba, *Two ion-selective filters in the calcium channel of the somatic membrane of mollusc neurons*, J. Memb. Biol., 76 (1983), pp. 83–93.

[LHJR00] F. Lehmann-Horn and K. Jurkat-Rott, *Channelopathies*, Elsevier Science, New York, 2000.

[LMZ06] A. Leitao, P. A. Markowich, and J. Zubelli, *On the inverse doping profile problems for the voltage-current map*, Inverse Problems, 22 (2006), pp. 1071–1088.

[Maetal03] A. B. Mamonov, R. D. Coalson, A. Nitzan, and M. G. Kurnikova, *The role of the dielectric barrier in narrow biological channels: A novel composite approach to modeling single-channel currents*, Biophys. J., 84 (2003), pp. 3646–3661.

[MRS90] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer, Wien, New York, 1990.

[Mietal04] H. Miedema, A. Meter-Arkema, J. Wierenga, J. Tang, B. Eisenberg, W. Nonner, H. Hektor, D. Gillespie, and W. Meijberg, *Permeation properties of an engineered bacterial OmpF porin containing the EEEE-locus of Ca2+ channels*, Biophys. J., 87 (2004), pp. 3137–3147.

[Mietal06] H. Miedema, M. Vrouenraets, J. Wierenga, R. S. Eisenberg, D. Gillespie, W. Meijberg, and W. Nonner, *Ca2+ selectivity of a chemically modified OmpF with reduced pore volume*, Biophys J., 91 (2006), pp. 4392–4400.

[NHE03] B. Nadler, U. Hollerbach, and R. S. Eisenberg, *Dielectric boundary force and its crucial role in gramicidin*, Phys. Rev. E Stat. Nonlin. Soft Matter Phys., 68 (2003), 021905.

[Na06] F. Natterer, *Imaging and Inverse Problems of Partial Differential Equations*, preprint, University of Münster, Münster, Germany, http:/wwwmath1.uni-muenster.de/num/Preprints/2006/artikel.pdf (2006).

[NCE00] W. Nonner, L. Catacuzzeno, and R. Eisenberg, *Binding and selectivity in L-type Ca channels: A mean spherical approximation*, Biophys. J., 79 (2000), pp. 1976–1992.

[NPGE04] W. Nonner, A. Peyser, D. Gillespie, and B. Eisenberg, *Relating microscopic charge movement to macroscopic currents: The Ramo-Shockley theorem applied to ion channels*, Biophys. J., 87 (2004), pp. 3716–3722.

[OBGXY05] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, *An iterative regularization method for total variation-based image restoration*, Multiscale Model. Simul., 4 (2005), pp. 460–489.

[Ru90] I. Rubinstein, *Electro-diffusion of Ions*, SIAM, Philadelphia, 1990.

[SMC03] W. A. Sather and E. W. McCleskey, *Permeation and selectivity in calcium channels*, Annu. Rev. Physiol., 65 (2003), pp. 133–159.

[Sch98] O. Scherzer, *An iterative multi-level algorithm for solving nonlinear ill-posed problems*, Numer. Math., 80 (1998), pp. 579–600.

[SNE01] Z. Schuss, B. Nadler, and R. S. Eisenberg, *Derivation of PNP equations in bath and channel from a molecular model*, Phys. Rev. E (3), 64 (2001), 036111.

[SV89] T. I. Seidman and C. R. Vogel, *Well-posedness and convergence of some regularization methods for nonlinear ill-posed problems*, Inverse Problems, 5 (1989), pp. 227–238.

[Se84] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer, New York, 1984.

[Sietal06] Z. S. Siwy, M. R. Powell, A. Petrov, E. Kalman, C. Trautmann, and R. S. Eisenberg, *Calcium-induced voltage gating in single conical nanopores*, Nano Lett., 6 (2006), pp. 1729–1734.

[Sh96] R. E. Showalter, *Nonlinear Partial Differential Equations and Monotone Operators in Banach Space*, AMS, Providence, RI, 1996.

[TBSS01] D. P. Tieleman, P. C. Biggin, G. R. Smith, and M. S. Sansom, *Simulation approaches to ion channel structure-function relationships*, Q. Rev. Biophys., 34 (2001), pp. 473–561.

[VR50] W. R. Van Roosbroeck, *Theory of flow of electrons and holes in germanium and other semiconductors*, Bell System Tech. J., 29 (1950), pp. 560–607.

[Waetal05] Y. Wang, L. Xu, D. Pasek, D. Gillespie, and G. Meissner, *Probing the role of negatively charged amino acid residues in ion permeation of skeletal muscle ryanodine receptor*, Biophys. J., 89 (2005), pp. 256–265.

[We68] H. Weyl, *Gesammelte Abhandlungen*, Springer, Berlin, 1968.

[Wo06] M. T. Wolfram, *Semiconductor Inverse Dopant Profiling from Transient Measurements*, SFB Technical Report 2006-4 (SFB F013), Johannes Kepler University, Linz, Austria, 2006, submitted.

[WB06] M. T. Wolfram and M. Burger, *Inverse dopant profiling for highly doped semiconductor devices*, in Proceedings of the 5th MATHMOD Conference, F. Breitenecker and I. Troch, eds., Vienna, Austria, 2006.

[XWGM06] L. Xu, Y. Wang, D. Gillespie, and G. Meissner, *Two rings of negative charges in the cytosolic vestibule of type-1 ryanodine receptor modulate ion fluxes*, Biophys. J., 90 (2006), pp. 443–453.

SIAM J. APPL. MATH.
Vol. 67, No. 4, pp. 990–1015

THE METHOD OF MOMENTS FOR NONLINEAR SCHRÖDINGER EQUATIONS: THEORY AND APPLICATIONS*

VÍCTOR M. PÉREZ-GARCÍA†, PEDRO J. TORRES‡, AND GASPAR D. MONTESINOS§

Abstract. The method of moments in the context of nonlinear Schrödinger equations relies on defining a set of integral quantities, which characterize the solution of this partial differential equation and whose evolution can be obtained from a set of ordinary differential equations. In this paper we find all cases in which the method of moments leads to closed evolution equations, thus extending and unifying previous works in the field of applications. For some cases in which the method fails to provide rigorous information we also develop approximate methods based on it, which allow us to get some approximate information on the dynamics of the solutions of the nonlinear Schrödinger equation.

Key words. nonlinear Schrödinger equations, methods of moments, nonlinear optics, Bose–Einstein condensates

AMS subject classifications. 35Q55, 78M50, 35B34, 78M05, 78A60

DOI. 10.1137/050643131

1. Introduction.

Nonlinear Schrödinger (NLS) equations appear in a great array of contexts [1] as, for example, in semiconductor electronics [2, 3], optics in nonlinear media [4], photonics [5], plasmas [6], fundamentation of quantum mechanics [7], dynamics of accelerators [8], mean-field theory of Bose–Einstein condensates [9], or biomolecule dynamics [10]. In some of these fields and many others, the NLS equation appears as an asymptotic limit for a slowly varying dispersive wave envelope propagating in a nonlinear medium [11].

The study of these equations has served as the catalyst for the development of new ideas or even mathematical concepts such as solitons [12] or singularities in partial differential equations [13, 14].

One of the most general ways to express an NLS equation is

$$i\frac{\partial u}{\partial t} = -\frac{1}{2}\Delta u + V(x,t)u + g(|u|^2,t)u - i\sigma(t)u, \tag{1.1}$$

where $\Delta = \partial^2/\partial x_1^2 + \cdots + \partial^2/\partial x_n^2$ and u is a complex function which describes some physical wave. We shall consider here the solution of (1.1) on $\mathbb{R}^n$ and therefore $u : \mathbb{R}^n \times [0,T] \to \mathbb{C}$, with initial values $u(x,0) = u_0(x) \in X$, X being an appropriate functional space ensuring finiteness of certain integral quantities to be defined later.

The family of NLS equations (1.1) contains many particular cases, depending on the specific choices of the nonlinear terms $g(|u|^2,t)$, the potentials $V(x,t)$, the

*Received by the editors October 19, 2005; accepted for publication (in revised form) November 17, 2006; published electronically May 1, 2007. The research of the first and third authors was supported by grants FIS2006-04190 (Ministerio de Educación y Ciencia) and PAI-05-001 (Consejería de Educación y Ciencia de la Junta de Comunidades de Castilla-La Mancha).

http://www.siam.org/journals/siap/67-4/64313.html

†Departamento de Matemáticas, Universidad de Castilla-La Mancha, E.T.S.I. Industriales, and Instituto de Matemática Aplicada a la Ciencia y la Ingeniería, Avd. Camilo José Cela, s/n, Ciudad Real, E-13071, Spain (victor.perezgarcia@uclm.es).

‡Departamento de Matemáticas, Universidad de Granada, Campus de Fuentenueva s/n, Granada 18071, Spain (p.torres@ugr.es). This author's research was supported by grant MTM2005-03483, Ministerio Ciencia y Tecnología, Spain.

§Departamento de Matemáticas, Universidad de Castilla-La Mancha, E.T.S.I. Industriales, Avd. Camilo José Cela, s/n, Ciudad Real, E-13071, Spain (gaspard.montesinos@uclm.es).

dissipation $\sigma(t)$, and the dimension of the space n. The best known cases are those of power type, $g(|u|^2) = \alpha|u|^p$, or polynomial, $g(|u|^2) = \alpha_1|u|^{p_1} + \alpha_2|u|^{p_2}$, nonlinearities.

The potential term $V(x,t)$ models the action of an external force acting upon the system and may have many forms. Finally, we include in (1.1) a simple loss term arising in different applications [15, 16]. In many cases these losses are negligible, i.e., $\sigma = 0$.

The description of the dynamics of initial data ruled by this equation is of great interest for applications. Nevertheless, gathering information about the solutions of a partial differential equation that is nonlinear like (1.1) constitutes a problem that is a priori nearly unapproachable. For this reason, most studies about the dynamics of this type of equation are exclusively numerical. The rigorous studies carried out to date concentrate on (i) properties of stationary solutions [17], (ii) particular results on the existence of solutions [18, 19], and (iii) asymptotic properties [13, 14].

Only when $n = 1$, $g(|u|^2, t) = |u|^2$, $V(x,t) = 0$, $\sigma = 0$ it is possible to arrive at a solution of the initial value problem by using the inverse scattering transform method [12]. In this paper we develop the so-called method of moments, which tries to provide qualitative information about the behavior of the solutions of NLS equations. Instead of tackling the Cauchy problem (1.1) directly, the method studies the evolution of several integral quantities (the so-called *moments*) of the solution $u(x,t)$. In some cases the method allows one to reduce the problem to the study of systems of coupled ordinary nonlinear differential equations. In other cases the method provides a foundation for making approximations in a more systematic (and simpler) way than other procedures used in physics, such as those involving finite-dimensional reductions of the original problem, namely, the averaged Lagrangian, collective coordinates, or variational methods [20, 21]. In any case the method of moments provides information which is very useful for the applied scientist, who is usually interested in obtaining as much information as possible characterizing the *dynamics* of the solutions of the problem.

It seems that the first application of the method of moments was performed by Talanov [22] in order to find a formal condition of sufficiency for the blowup of solutions of the NLS equation with $g(|u|^2) = -|u|^2$ and $n = 2$. Since then, the method has been applied to different particular cases (mainly solutions of radial symmetry in two spatial dimensions), especially in the context of optics, where many equations of NLS type arise [23].

In previous research, the method of moments has been studied in a range of specific situations, but in all such cases the success of the method is unrelated to a more general study. In this paper we try to consider the method systematically and solve a number of open questions: (i) to find the most general type of nonlinear term and potentials in (1.1) for which the method of moments allows us to get conclusions, and (ii) to develop approximate methods based on it for situations in which the moment equations do not allow us to obtain exact results.

2. Preliminary considerations. Let us define the functional space $Q(H)$ as the space of functions for which the so-called energy functional,

$$E(u) = (u, Hu)_{L^2(\mathbb{R}^n)} + \int_{\mathbb{R}^n} G(|u|^2, t)dx, \tag{2.1}$$

is finite, G being a function such that $\partial G(|u|^2,t)/\partial|u|^2 = g(|u|^2,t)$, $(\cdot,\cdot)_{L^2(\mathbb{R}^n)}$ denoting the usual scalar product in $L^2(\mathbb{R}^n)$, and $H = -\frac{1}{2}\Delta + V(x,t)$. For the case

$V(x,t) = 0$ and g independent of time, several results on existence and uniqueness were given by Velo [18].

As regards the case $V \neq 0$, which is the one in which we are mainly interested in our work, the best documented case in the literature is that of potentials with $|D^\alpha V|$ bounded in $\mathbb{R}^n$ for every $|\alpha| \geq 2$; that is, potentials with at most quadratic growth. Oh [24] proved the local existence of solutions in $L^2(\mathbb{R}^n)$ and in $Q(H)$ for nonlinearities of the type $g(u) = -|u|^p$, $0 \leq p < 4/n$. However, the procedure used allows one to substitute this nonlinearity with other more general ones. It also seems possible to extend the results to the case in which the potential depends on t.

Therefore, from now on we shall suppose that the nonlinear term satisfies the conditions set by Velo and that $|D^\alpha V|$ is bounded in space for all $|\alpha| \geq 2$. Under these conditions it is possible to guarantee at least local existence of solutions of (1.1) in appropriate functional spaces.

2.1. Formal elimination of the loss term. In the first place we carry out the transformation [15]

$$\hat{u}(x,t) = u(x,t)e^{\int_0^t \sigma(\tau)d\tau}, \tag{2.2}$$

which is well defined for any bounded function $\sigma(t)$ (this includes all known realistic cases arising in the applications). The equation satisfied by $\hat{u}(x,t)$ is obtained from the following direct calculation:

$$\begin{aligned} i\frac{\partial \hat{u}}{\partial t} &= i\frac{\partial u}{\partial t}e^{\int_0^t \sigma(\tau)d\tau} + i\sigma(t)ue^{\int_0^t \sigma(\tau)d\tau} \\ &= \left[-\frac{1}{2}\Delta u + V(x,t)u + g(|u|^2,t)u - i\sigma(t)u\right] e^{\int_0^t \sigma(\tau)d\tau} + i\sigma(t)ue^{\int_0^t \sigma(\tau)d\tau} \\ &= -\frac{1}{2}\Delta\hat{u} + V(x,t)\hat{u} + \hat{g}(|\hat{u}|^2,t)\hat{u}, \end{aligned}$$

where

$$\hat{g}(|\hat{u}|^2,t) = g(|u|^2,t) = g(e^{-2\int_0^t \sigma(\tau)d\tau}|\hat{u}|^2,t). \tag{2.3}$$

From here on we shall consider, with no loss of generality, that $\sigma(t) = 0$ in (1.1), assuming that this choice might add an extra time-dependence to the nonlinear term.

3. The method of moments: Generalities.

3.1. Definition of the moments. Let us define the following quantities:

$$I_{k,j}(t) = \int x_j^k |u(x,t)|^2 dx = \left\| x_j^{k/2} u \right\|^2_{L^2(\mathbb{R}^n)}, \tag{3.1a}$$

$$V_{k,j}(t) = 2^{k-1} i \int x_j^k \left(u(x,t)\frac{\partial u(\bar{x},t)}{\partial x_j} - \bar{u}\frac{\partial u(x,t)}{\partial x_j} \right) dx, \tag{3.1b}$$

$$K_j(t) = \frac{1}{2}\int \left|\frac{\partial u(x,t)}{\partial x_j}\right|^2 dx = \frac{1}{2}\left\|\frac{\partial u}{\partial x_j}\right\|^2_{L^2(\mathbb{R}^n)} \tag{3.1c}$$

$$J(t) = \int G(|u(x,t)|^2,t)dx, \tag{3.1d}$$

with $j = 1,\ldots,n$ and $k = 0,1,2,\ldots$, which we will call *moments* of $u(x,t)$ in analogy with the moments of a distribution. From now on and also in (3.1) it is understood

that all integrals and norms refer to the spatial variables $x \in \mathbb{R}^n$ unless otherwise stated. In (3.1) we denote by $\bar{u}$ the complex conjugate of u.

In some cases we will make specific reference to which solution u of (1.1) is used to calculate the moments by means of the notation: $I^u_{k,j}$, etc.

The moments are quantities that have to do with *intuitive* properties of the solution $u(x,t)$. For example, the moment $I_{0,0}$ is the squared $L^2(\mathbb{R}^n)$-norm of the solution and therefore measures the *magnitude*, *quantity*, or *mass* thereof. Depending on the particular application context, this moment is denominated mass, charge, intensity, energy, number of particles, etc. The moments $I_{1,j}(t)$ are the coordinates of the *center* of the distribution u, giving us an idea of the overall position thereof. The quantities $I_{2,j}$ are related to the *width* of the distribution defined as $W_j = (\int_{\mathbb{R}^n}(x_j - I_{1,j})^2|u|^2dx)^{1/2} = (I_{2,j} + I^2_{1,j}I_{0,0} - 2I^2_{1,j})^{1/2}$, which is also a quantity with an evident meaning.

The evolution of the moments is determined by that of the function $u(x,t)$. From now on we will assume that the initial datum $u_0(x)$ and the properties of the equation guarantee that the moments are well defined for all time. This excludes explicitly certain classes of initial data such as plane waves, etc., which do not decay at infinity. Thus our results will be relevant for studying the evolution of initially localized and regular enough initial data.

3.2. First *conservation law*. It is easy to prove formally that the moment $I_{0,0}$ is invariant during the temporal evolution by just calculating

$$\begin{aligned}
\frac{d}{dt}I_{0,0}(t) &= \int_{\mathbb{R}^n}\left(\frac{d}{dt}|u|^2\right)dx = \int_{\mathbb{R}^n}\left(\bar{u}\frac{\partial}{\partial t}u + u\frac{\partial}{\partial t}\bar{u}\right)dx \\
&= \int_{\mathbb{R}^n} i\left(\frac{1}{2}\bar{u}\Delta u - V(x,t)|u|^2 - g(|u|^2,t)|u|^2\right)dx \\
&\quad + \int_{\mathbb{R}^n} i\left(-\frac{1}{2}u\Delta\bar{u} + V(x,t)|u|^2 + g(|u|^2,t)|u|^2\right)dx \\
&= \int_{\mathbb{R}^n}\frac{i}{2}\left(\bar{u}\Delta u - u\Delta\bar{u}\right)dx \\
&= \frac{i}{2}\left(\int_{\mathbb{R}^n}|\nabla u|^2dx - \int_{\mathbb{R}^n}|\nabla u|^2dx\right) = 0, \qquad (3.2)
\end{aligned}$$

where we have performed integration by parts and used that the function u and its derivatives vanish at infinity.

Obviously the above *demonstration* is formal in the sense that a regularity, which we do not know for certain, has been used for u. Nevertheless, this type of proof can be formalized by making a convolution of the function u with a regularizing function. The details of these methodologies can be seen in [18] or [24, 25]. In this paper we will limit ourselves to formal calculations.

4. General results for harmonic potentials.

4.1. Introduction. From this point onward we will focus on the particular case of interest for this study when $V(x,t)$ is a harmonic potential of the type $V(x,t) = \frac{1}{2}(x, \Lambda(t)x)$, where Λ is a real matrix of the form $\Lambda_{ij}(t) = \lambda_i^2(t)\delta_{ij}$, with $\lambda_i \geq 0$ for $i = 1, \ldots, n$, and δ_{ij} is the Kronecker delta. Bearing in mind the results of section 2.1,

the NLS equation under study is then

$$i\frac{\partial u}{\partial t} = -\frac{1}{2}\Delta u + \frac{1}{2}\left(\sum_{j=1}^{n}\lambda_j^2 x_j^2\right)u + g(|u|^2,t)u. \tag{4.1}$$

This equation appears in a wide variety of applications such as propagation of waves through optical transmission lines with online modulators [26, 27, 28, 29], propagation of light beams in nonlinear media with a gradient of the refraction index [30, 31], or dynamics of Bose–Einstein condensates [9]. Generically it can provide a model for studying some properties of the solutions of NLS equations localized near a minimum of a general potential $V(x)$.

4.2. First moment equations. If we differentiate the definitions of the moments $I_{1,j}$ and $V_{0,j}$, we obtain, after some calculations, the evolution equations

$$\frac{dI_{1,j}}{dt} = V_{0,j}, \tag{4.2a}$$

$$\frac{dV_{0,j}}{dt} = -\lambda_j^2 I_{1,j}, \tag{4.2b}$$

so that $I_{1,j}$, $j = 1,\dots,n$, satisfy

$$\frac{d^2 I_{1,j}}{dt^2} + \lambda_j^2 I_{1,j} = 0 \tag{4.3}$$

with initial data $I_{1,j}(0), \dot{I}_{1,j}(0) = V_{0,j}(0)$. These expressions are a generalization of the Ehrenfest theorem of linear quantum mechanics to the NLS equation and particularized for the potential that concerns us [24, 32].

This result has been discussed previously in many papers and is physically very interesting. It indicates that the evolution of the center of the solution is independent of the nonlinear effects and of the evolution of the rest of the moments and depends only on the potential parameters.

4.3. Reduction of the general problem to the case $I_{1,j} = V_{0,j} = 0$. We shall begin by stating the following lemma [33].

LEMMA 4.1. *Let $u(x,t)$ be a solution of* (4.1) *with the initial datum $u(x,0) = u_0(x)$. Then the functions*

$$u_R(x,t) = u(x - R(t),t)e^{i\theta(x,t)}, \tag{4.4a}$$

where

$$\theta(x,t) = \left(x,\dot{R}\right) + \int_0^t \left[(\dot{R}(t'),\dot{R}(t')) - (R(t'),\Lambda(t')R(t'))\right]dt' \tag{4.4b}$$

and

$$\frac{d^2 R}{dt^2} + \Lambda R = 0 \tag{4.4c}$$

for any set of initial data $R(0), \dot{R}(0) \in \mathbb{R}^n$, are also solutions of (4.1).

Proof. All we have to do is substitute (4.4a), (4.4b), and (4.4c) into (4.1). ∎

One noteworthy conclusion is that, given a solution of (4.1), we can *translate* it initially by a constant vector and obtain another solution. In the case of stationary states, defined as solutions of the form

$$u(x,t) = \varphi_\mu(x) e^{i\mu t}, \tag{4.5}$$

which exist in the autonomous case (i.e., $d\lambda/dt = 0$) and whose dynamics is trivial, this result implies that under displacements the only dynamics acquired is one of the movement of the center given by (4.4c). The coincidence of the evolution laws (4.3) and (4.4c) allows us to state the following theorem, which is an immediate consequence of the above lemma.

THEOREM 4.2. *If $\psi(x,t)$ is a solution of (4.1) with nonzero $I^\psi_{1,j}$ or $V^\psi_{0,j}$, then there exists a unique solution $u(x,t) = \psi(x + \{I^\psi_{1,j}(t)\}_j, t) e^{i\theta(x,t)}$ with*

$$\theta(x,t) = -\sum_j x_j V^\psi_{0,j} + \sum_j \left[\int_0^t V^\psi_{0,j}(t')^2 - \lambda_j(t')^2 I^\psi_{1,j}(t')^2 \right] dt'$$

such that $I^u_{1,j} = 0$ and $V^u_{0,j} = 0$.

The important conclusion of this theorem is that it suffices to study solutions with $I_{1,j}$ and $V_{0,j}$ equal to zero, as those that have one of these coefficients different from zero can be obtained from previous ones, by means of translation and multiplication by a linear phase in x. From a practical standpoint, what is most important is that $I_{1,j}$ be null without any loss of generality, as then we can establish a direct link between the widths and the moments $I_{2,j}$ (see the discussion in the third paragraph of section 3).

4.4. Moment equations. Assuming that all of the moments can be defined at any time t, we can calculate their evolution equations by means of direct differentiation. The results are gathered in the next theorem.

THEOREM 4.3. *Let $u_0(x)$ be an initial datum such that the moments $I_{2,j}$, $V_{1,j}$, K_j, and J are well defined at $t = 0$. Then*

$$\frac{dI_{2,j}}{dt} = V_{1,j}, \tag{4.6a}$$

$$\frac{dV_{1,j}}{dt} = 4K_j - 2\lambda_j^2 I_{2,j} - 2\int_{\mathbb{R}^n} D(\rho,t)dx, \tag{4.6b}$$

$$\frac{dK_j}{dt} = -\frac{1}{2}\lambda_j^2 V_{1,j} - \int_{\mathbb{R}^n} D(\rho,t)\frac{\partial^2 \phi}{\partial x_j^2}\, dx, \tag{4.6c}$$

$$\frac{dJ}{dt} = \int_{\mathbb{R}^n} D(\rho,t)\Delta\phi\, dx + \int_{\mathbb{R}^n} \frac{\partial G(\rho,t)}{\partial t} dx, \tag{4.6d}$$

where $D(\rho,t) = G(\rho,t) - \rho g(\rho,t)$, $\rho = |u(x,t)|^2$.

Proof. The demonstration of the validity of (4.6) can be carried out from direct calculations, performing integration by parts, and using the decay of u and ∇u at infinity.

To demonstrate (4.6a) it is easier to work with the modulus-phase representation

of u, $u=\rho^{1/2}e^{i\phi}$ (with $\rho>0$). Then

$$\begin{aligned}\frac{dI_{2,j}}{dt}&=\int x_j^2\dot{\rho}=-\int x_j^2(\nabla\rho\cdot\nabla\phi+\rho\Delta\cdot\phi)\\&=-\int x_j^2\nabla\rho\cdot\nabla\phi+\int\nabla(x_j^2\rho)\cdot\nabla\phi\\&=-\int x_j^2\nabla\rho\cdot\nabla\phi+\int x_j^2\nabla\rho\cdot\nabla\phi+\int\nabla(x_j^2)\rho\cdot\nabla\phi\\&=2\int x_j\rho\frac{\partial\phi}{\partial x_j}=V_{1,j}.\end{aligned}$$

We can also prove (4.6b) as follows:

$$\begin{aligned}\frac{dV_{1,j}}{dt}&=i\int x_j\left(u_t\frac{\partial\bar{u}}{\partial x_j}+u\frac{\partial\bar{u_t}}{\partial x_j}-\bar{u_t}\frac{\partial u}{\partial x_j}-\bar{u}\frac{\partial u_t}{\partial x_j}\right)\\&=\int x_j\left[-\frac{1}{2}\Delta u\frac{\partial\bar{u}}{\partial x_j}+\frac{1}{2}\left(\sum\lambda_k^2x_k^2\right)u\frac{\partial\bar{u}}{\partial x_j}+gu\frac{\partial\bar{u}}{\partial x_j}+u\frac{1}{2}\frac{\partial\Delta\bar{u}}{\partial x_j}\right.\\&\qquad\left.-\lambda_j^2x_j|u|^2-\frac{1}{2}\left(\sum\lambda_k^2x_k^2\right)u\frac{\partial\bar{u}}{\partial x_j}-\frac{\partial g}{\partial x_j}|u|^2-gu\frac{\partial\bar{u}}{\partial x_j}+\text{c.c.}\right]\\&=-2\lambda_j^2\int x_j^2|u|^2-2\int x_j|u|^2\frac{\partial g}{\partial x_j}\\&\qquad-\frac{1}{2}\int x_j\left(\Delta u\frac{\partial\bar{u}}{\partial x_j}+\Delta\bar{u}\frac{\partial u}{\partial x_j}-u\frac{\partial\Delta\bar{u}}{\partial x_j}+\bar{u}\frac{\partial\Delta u}{\partial x_j}\right),\end{aligned}\tag{4.7}$$

where c.c. indicates the complex conjugate. Operating on the above integrals, we have

$$\begin{aligned}&\int x_j\left(\Delta u\frac{\partial\bar{u}}{\partial x_j}+\Delta\bar{u}\frac{\partial u}{\partial x_j}-u\frac{\partial\Delta\bar{u}}{\partial x_j}-\bar{u}\frac{\partial\Delta u}{\partial x_j}\right)\\&=-2\int x_j\left(\nabla u\cdot\frac{\partial\nabla\bar{u}}{\partial x_j}+\nabla\bar{u}\cdot\frac{\partial\nabla u}{\partial x_j}\right)-4\int\left|\frac{\partial u}{\partial x_j}\right|^2-2\int|\nabla u|^2\\&=2\int|\nabla u|^2-4\int\left|\frac{\partial u}{\partial x_j}\right|^2-2\int|\nabla u|^2=-4\int\left|\frac{\partial u}{\partial x_j}\right|^2\end{aligned}$$

and

$$\int x_j\frac{\partial g}{\partial x_j}\rho=-\int g\rho-\int x_j\frac{\partial G}{\partial x_j}=-\int g\rho+\int G=\int D.$$

By substitution into (4.7), we arrive at the desired result:

$$\begin{aligned}\frac{dV_{1,j}}{dt}&=-2\lambda_j^2\int x_j^2|u|^2+2\int\left|\frac{\partial u}{\partial x_j}\right|^2-2\int D\\&=-2\lambda_j^2I_{2,j}+4K_j-2\int D.\end{aligned}$$

Let us now prove (4.6c):

$$\begin{aligned}\frac{dK_j}{dt}&=\frac{1}{2}\int\frac{d}{dt}\left(\frac{\partial u}{\partial x_j}\frac{\partial\bar{u}}{\partial x_j}\right)=\frac{1}{2}\int\left(\frac{\partial u_t}{\partial x_j}\frac{\partial\bar{u}}{\partial x_j}+\frac{\partial u}{\partial x_j}\frac{\partial\bar{u_t}}{\partial x_j}\right)\\&=\frac{1}{2}\int\frac{\partial}{\partial x_j}\left[\frac{i}{2}\Delta u-\frac{i}{2}\left(\sum_{k=1}^n\lambda_k^2x_k^2\right)u-igu\right]\frac{\partial\bar{u}}{\partial x_j}+\text{c.c.};\end{aligned}$$

then

$$\frac{dK_j}{dt} = -\frac{i}{2}\lambda_j^2 \int x_j \left(u\frac{\partial \bar{u}}{\partial x_j} - \bar{u}\frac{\partial u}{\partial x_j}\right) - \frac{i}{2}\int \frac{\partial g}{\partial x_j}\left(u\frac{\partial \bar{u}}{\partial x_j} - \bar{u}\frac{\partial u}{\partial x_j}\right)$$
$$+ \frac{i}{4}\int \left(\frac{\partial \Delta u}{\partial x_j}\frac{\partial \bar{u}}{\partial x_j} - \frac{\partial u}{\partial x_j}\frac{\partial \Delta \bar{u}}{\partial x_j}\right) = -\frac{1}{2}\lambda_j^2 V_{1,j} + \int g\frac{\partial}{\partial x_j}\left(\rho\frac{\partial \phi}{\partial x_j}\right)$$
$$= -\frac{1}{2}\lambda_j^2 V_{1,j} + \int g\left(\rho\frac{\partial^2 \phi}{\partial x_j^2} + \frac{\partial \rho}{\partial x_j}\frac{\partial \phi}{\partial x_j}\right),$$

and, using the definition of G, we obtain

$$\frac{dK_j}{dt} = -\frac{1}{2}\lambda_j^2 V_{1,j} + \int g\rho\frac{\partial^2 \phi}{\partial x_j^2} + \int \frac{\partial G}{\partial x_j}\frac{\partial \phi}{\partial x_j}$$
$$= -\frac{1}{2}\lambda_j^2 V_{1,j} + \int g\rho\frac{\partial^2 \phi}{\partial x_j^2} - \int G\frac{\partial^2 \phi}{\partial x_j^2}$$
$$= -\frac{1}{2}\lambda_j^2 V_{1,j} - \int (G - \rho g)\frac{\partial^2 \phi}{\partial x_j^2} = -\frac{1}{2}\lambda_j^2 V_{1,j} - \int D\frac{\partial^2 \phi}{\partial x_j^2}.$$

Finally, to demonstrate (4.6d) we proceed as follows:

$$\frac{dJ}{dt} = \int \frac{dG(\rho,t)}{dt} = \int \left[\frac{\partial G}{\partial \rho}\left(\frac{\partial \rho}{\partial u}u_t + \frac{\partial \rho}{\partial \bar{u}}\bar{u}_t\right) + \frac{\partial G}{\partial t}\right]$$
$$= \int \left[\frac{\partial G}{\partial \rho}\left(\bar{u}u_t + u\bar{u}_t\right) + \frac{\partial G}{\partial t}\right] = \frac{i}{2}\int \left[g\left(\bar{u}\Delta u - u\Delta \bar{u}\right) + \frac{\partial G}{\partial t}\right]$$
$$= \int \left[-g\nabla\cdot(\rho\nabla\phi) + \frac{\partial G}{\partial t}\right] = \int \left[-g\nabla\rho\cdot\nabla\phi - g\rho\Delta\phi + \frac{\partial G}{\partial t}\right]$$
$$= \int \left[-\nabla G\cdot\nabla\phi - g\rho\Delta\phi + \frac{\partial G}{\partial t}\right] = \int (G - g\rho)\,\Delta\phi + \int \frac{\partial G}{\partial t}$$
$$= \int D\Delta\phi + \int \frac{\partial G}{\partial t}. \qquad \square$$

A direct consequence of the theorem is the following.

COROLLARY 4.4. *Let $u(x,t)$ be a stationary solution of* (4.1). *Then*

$$K_j = \frac{1}{2}\lambda_j^2 I_{2,j} + \frac{1}{2}\int D(\rho)dx. \tag{4.8}$$

5. Solvable cases of the method of moments. In this section we will study several particular situations of practical relevance in which the method of moments thoroughly provides exact results.

5.1. The linear case $g(\rho,t) = 0$. In this case, (3.1d) and (4.6d) tell us that $J(t) = 0$ for all t, and then the moment equations (4.6) become

$$\frac{dI_{2,j}}{dt} = V_{1,j}, \tag{5.1a}$$

$$\frac{dV_{1,j}}{dt} = 4K_j - 2\lambda_j^2 I_{2,j}, \tag{5.1b}$$

$$\frac{dK_j}{dt} = -\frac{1}{2}\lambda_j^2 V_{1,j}. \tag{5.1c}$$

That is, in the linear case the equations for the moments along each direction j of the physical space $\mathbb{R}^n$ are uncoupled. This property was known in the context of optics for $n = 2$ and constant λ_j [34]. Here we see that this property holds for any number of spatial dimensions, time dependence $\lambda(t)$, and even for nonsymmetric initial data.

5.2. Condition of closure of the moment equations in the general case. Equations (4.6) do not form a closed set, and therefore to obtain, in general, their evolution we would need to continue obtaining moments of a higher order, which would provide us with an infinite hierarchy of differential equations. Given the similarity among the terms that involve second derivatives of the phase of the solution in (4.6), it is natural to wonder whether it would be possible to somehow close the system and thus obtain information about the solutions.

From this point on, and for the rest of the section, we will limit ourselves to the case $\lambda_j(t) = \lambda(t)$, $j = 1, \dots, n$, which is the most realistic one, and which includes as a particular case the situation without external potentials $\lambda_j = 0$. Let us define the following quantities:

$$\mathcal{I} = \sum_{j=1}^{n} I_{2,j}, \quad \mathcal{V} = \sum_{j=1}^{n} V_{1,j}, \quad \mathcal{K} = \sum_{j=1}^{n} K_j. \tag{5.2}$$

Differentiating (5.2) and using (4.6), we have

$$\frac{d\mathcal{I}}{dt} = \mathcal{V}, \tag{5.3a}$$

$$\frac{d\mathcal{V}}{dt} = 4\mathcal{K} - 2\lambda^2\mathcal{I} - 2n\int_{\mathbb{R}^n} D(\rho,t)dx, \tag{5.3b}$$

$$\frac{d\mathcal{K}}{dt} = -\frac{1}{2}\lambda^2\mathcal{V} - \int_{\mathbb{R}^n} D(\rho,t)\Delta\phi dx, \tag{5.3c}$$

$$\frac{dJ}{dt} = \int_{\mathbb{R}^n} D(\rho,t)\Delta\phi + \int_{\mathbb{R}^n} \frac{\partial G(\rho,t)}{\partial t}dx. \tag{5.3d}$$

If we add up (5.3c) and (5.3d), we arrive at

$$\frac{d\mathcal{I}}{dt} = \mathcal{V}, \tag{5.4a}$$

$$\frac{d\mathcal{V}}{dt} = 4\left[\mathcal{K} - \frac{n}{2}\int_{\mathbb{R}^n} D(\rho,t)dx\right] - 2\lambda^2\mathcal{I}, \tag{5.4b}$$

$$\frac{d\left(\mathcal{K}+J\right)}{dt} = -\frac{1}{2}\lambda^2\mathcal{V} + \int_{\mathbb{R}^n} \frac{\partial G}{\partial t}dx. \tag{5.4c}$$

In order that equations (5.4) form a closed system, they must fulfill $-\frac{n}{2}\int_{\mathbb{R}^n} D(\rho,t)dx = J = \int_{\mathbb{R}^n} G(\rho,t)dx$ and that $\int_{\mathbb{R}^n} \frac{\partial G(\rho,t)}{\partial t}dx$ can be expressed in terms of the other known quantities. The former condition requires that

$$0 = \int_{\mathbb{R}^n}\left[\frac{n}{2}D(\rho,t) + G(\rho,t)\right]dx = \int_{\mathbb{R}^n}\left[\left(1+\frac{n}{2}\right)G(\rho,t) - \frac{n}{2}\frac{\partial G(\rho,t}{\partial\rho}\rho\right]dx.$$

As $G(\rho,t)$ does not depend explicitly on x, this condition is verified when

$$\frac{\partial G(\rho,t)}{\partial\rho} = \left(1+\frac{2}{n}\right)\frac{G(\rho,t)}{\rho}, \tag{5.5}$$

that is, if

$$G(\rho,t) = g_0(t)\rho^{1+2/n}, \tag{5.6}$$

or, equivalently, if

$$g(\rho,t) = g_0(t)\rho^{2/n}, \tag{5.7}$$

where $g_0(t)$ is an arbitrary function that indicates the temporal variation of the nonlinear term. Then

$$\int_{\mathbb{R}^n} \frac{\partial G(\rho,t)}{\partial t}dx = \frac{1}{g_0}\frac{dg_0}{dt}\int_{\mathbb{R}^n} G(\rho,t)dx = \frac{1}{g_0}\frac{dg_0}{dt}J(t). \tag{5.8}$$

To close the equations it is necessary that $g_0(t)$ be constant in order to cancel the last term of this expression. Then, the nonlinearities for which it is possible to find closed results are

$$g(\rho) = g_0\rho^{2/n} = g_0|u|^{4/n}, \tag{5.9}$$

with $g_0 \in \mathbb{R}$, remembering that in the case $g_0 < 0$ there may be problems of blowup in finite time. Fortunately, these nonlinearities for $n = 1, 2, 3$ correspond to cases of practical interest. For instance, the case $n = 1$ with quintic nonlinearity has been studied in [35, 36, 37] and the case $n = 2$, with cubic nonlinearity, corresponds probably to the most relevant instance of the NLS equation, i.e., the cubic one in two spatial dimensions [30, 31]. For $n = 3$ the nonlinearity given by (5.9) appears in the context of the Hartree–Fock theory of atoms.

5.3. Simplification of the moment equations. Defining a new quantity $\mathcal{E} = \mathcal{K} + J$, (5.4) becomes

$$\frac{d\mathcal{I}}{dt} = \mathcal{V}, \tag{5.10a}$$

$$\frac{d\mathcal{V}}{dt} = 4\mathcal{E} - 2\lambda^2(t)\mathcal{I}, \tag{5.10b}$$

$$\frac{d\mathcal{E}}{dt} = -\frac{1}{2}\lambda^2(t)\mathcal{V}. \tag{5.10c}$$

These equations form a set of nonautonomous linear equations for the three averaged moments: $\mathcal{E}(t)$, $\mathcal{V}(t)$, and $\mathcal{I}(t)$. To continue our analysis, we note that

$$\mathcal{Q} = 2\mathcal{E}\mathcal{I} - \frac{1}{4}\mathcal{V}^2 \tag{5.11}$$

is a dynamical invariant of (5.10). We finally define $X = |\mathcal{Q}|^{-1/4}\mathcal{I}^{1/2}$, which is proportional to the *mean width* of u. A simple calculation allows us to corroborate that the equation that $X(t)$ satisfies is

$$\frac{d^2X}{dt^2} + \lambda^2(t)X = \frac{\operatorname{sgn}(\mathcal{Q})}{X^3}. \tag{5.12}$$

Solving (5.12) allows us to calculate $\mathcal{V}$ and $\mathcal{E}$ by simple substitution in (5.10). This equation is similar to that obtained in [30] for solutions of radial symmetry in the case $n = 2$ and $g(u) = |u|^2$. Here we find that it is possible to obtain a more general

result for solutions without specific symmetry requirements, and for any combination of dimension and nonlinearity $g(u) = |u|^p$ satisfying the condition $np = 4$. The case with $\mathrm{sgn}(\mathcal{Q}) = -1$ corresponds to collapsing situations [13, 38]. In what follows we consider mostly the case $\mathcal{Q} > 0$.

Equation (5.12) was studied by Ermakov in 1880 [39], although since then it has been *rediscovered* many times (see, e.g., [40]). It is a particular case of the so-called Ermakov systems [41, 42, 43], for which it is possible to give fairly complete results. Especially easy, though tedious to demonstrate, is the following claim.

THEOREM 5.1 (Ermakov, 1880). *Let $X(t)$ be the solution of* (5.12) *with initial data $X(0) = X_0$, $\dot{X}(0) = \dot{X}_0$. Then, if $\chi_1(t)$ and $\chi_2(t)$ are solutions of the differential equation*

$$\frac{d^2\chi}{dt^2} + \lambda^2(t)\chi = 0 \tag{5.13a}$$

satisfying the initial data $\chi_1(0) = X_0$, $\dot{\chi}_1(0) = \dot{X}_0$ and $\chi_2(0) = 0$, $\dot{\chi}_2(0) \neq 0$, then

$$X(t) = \sqrt{\chi_1^2(t) + \frac{1}{w^2}\chi_2^2(t)}, \tag{5.13b}$$

where w is the constant $w = \chi_1\dot{\chi}_2 - \chi_2\dot{\chi}_1$.

Equation (5.13b) is often called *the principle of nonlinear superposition*. Equation (5.13a) is the well-known Hill's equation [44] which models a parametrically forced oscillator and which has been studied in depth. In the following, we shall study a couple of special situations in view of their physical interest.

It is remarkable that the complex dynamics of a family of nonlinear partial differential equations can be understood in terms of a simple equation such as Hill's.

If we suppose that the function $\lambda^2(t)$ depends on a parameter ε in the way $\lambda^2(t) = 1+\tilde{\lambda}_\varepsilon(t)$, $\tilde{\lambda}_\varepsilon(t)$ being a periodic function with maximum value ε (not necessarily small), there exists a complete theory that describes the intervals of values of ε for which the solutions of (5.13a) are bounded (intervals of stability) and the intervals for which the solutions are unbounded (intervals of instability) [44].

5.4. Connection of the method of moments with variational methods. In the physical literature devoted to the study of applications of the NLS equations there is a widely used method which receives different names depending on the specific field of application: time-dependent variational method, collective coordinates method, or method of averaged Lagrangians. There is a huge literature on the applications of this method to different problems (see, e.g., the reviews [20] and [21] for two specific application fields).

The idea of the method is to write the Lagrangian density corresponding to the NLS equation

$$\mathcal{L} = \frac{i}{2}\left(u\frac{\partial\bar{u}}{\partial t} - \bar{u}\frac{\partial u}{\partial t}\right) - \frac{1}{2}|\nabla u|^2 + V(x,t)|u|^2 + G(|u|), \tag{5.14}$$

and to transform the problem of solving the NLS equation into the problem of finding $u(x,t)$ such that the action

$$S(u,\bar{u}) = \int \mathcal{L}(x,t)dx\,dt \tag{5.15}$$

has an extremum.

This new problem is as difficult to handle as the equation itself. The idea of the heuristic method of averaged Lagrangians (or variational method or collective coordinates method) is to restrict the analysis of this variational problem to a particular family of trial functions *which are not the true solutions*, i.e., finding the extremum over a prescribed family of trial functions. Taking a particular form of the trial function depending on a few parameters $u(x,t) = \varphi(x, p_1(t), \dots, p_S(t))$ leads to an averaged finite-dimensional Lagrangian

$$L(t) = \int_{\mathbb{R}^n} \mathcal{L}(x,t)dx. \tag{5.16}$$

From (5.16), using the Euler–Lagrange equations

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{p_j}}\right) - \frac{\partial L}{\partial p_j} = 0, \tag{5.17}$$

one obtains evolution equations for the parameters $p_j(t)$.

Since the trial functions (sometimes called "solutions") must be incorporated from the very beginning in the treatment (i.e., one must choose their specific form to be either Gaussian, sech, etc.), the information provided by this method is the "approximate" evolution of the parameters $p_j(t)$, and since nobody knows how far the solution is from the trial function, it is not clear what the word "approximate" means in that context. Usually one can choose φ based on experience or qualitative considerations.

In this sense the moment method, when it works, provides a much more convenient and rigorous way to obtain the evolution of the relevant parameters without assuming an (incorrect) specific form of the solution. Moreover, since there are no error bounds for the estimates of the method of averaged Lagrangians, one must at the end simulate numerically the full NLS equation in order to validate the predictions of the time-dependent variational method. Within the framework of the method of moments these simulations are not necessary, since the equations are exact.

6. Applications.

6.1. Dynamics of laser beams in GRIN media. When a laser beam propagates in a medium with a gradex refraction index (GRIN medium) with a specific profile quadratic in the transverse coordinates, the distribution of intensity $u(x,y,z)$ in the permanent regime is ruled by (4.1) with $g(\rho) = \rho$ and $n = 2$ (in the optical version of the equation $t \leftrightarrow z$), so that we are dealing with the critical case that we know how to solve. Although in principle it would be possible to design fibers with arbitrary profiles, technically the simplest way is to join fibers with different uniform indexes in each section.

In this case, the phenomenon can be modeled by

$$\lambda^2(t) = \begin{cases} a^2, & t \in [0, T_a], \\ b^2, & t \in (T_a, T_a + T_b = T]. \end{cases} \tag{6.1}$$

Equation (5.13a) with $\lambda(t)$ given by (6.1) is known as the Meissner equation, whose solution is trivial, given in each segment by a combination of trigonometric functions.

The solutions to the Meissner equation can be bounded (periodical or quasi-periodical) or unbounded (resonant oscillations). In Figure 6.1 the two types of solutions are shown for a particular choice of parameters.

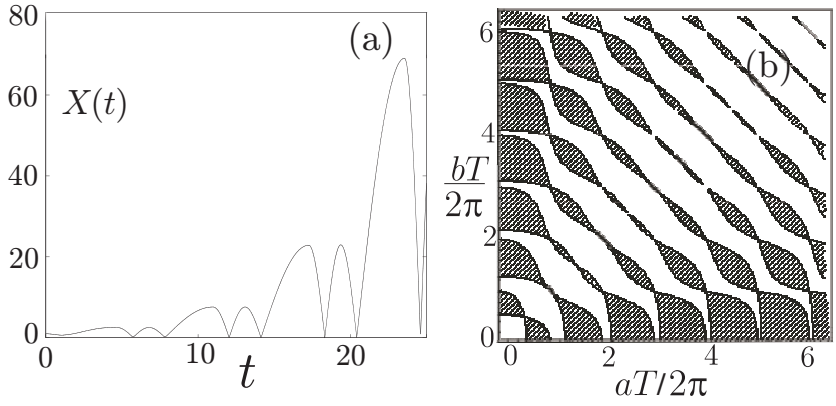


FIG. 6.1. *Solutions of* (5.13a) *with* $\lambda(t)$ *given by* (6.1). (a) *Resonant solution for* $T_a = 10\pi$, $T = 20\pi$, $a = 0.05$, $b = 0.15$ *and* (b) *regions of resonance for* $T_a = T_b = T/2$.

As far as the regions of stability in the space of parameters are concerned, they can be obtained by studying the discriminant of (5.13a), defined as the trace of the monodromy matrix, that is,

$$D(a, b, T_a, T_b) := \phi_1(T) + \phi_2'(T), \tag{6.2}$$

where ϕ_1, ϕ_2 are the solutions of (5.13a) satisfying the initial data $\phi_1(0) = 1$, $\phi_1'(0) = 0$ and $\phi_2(0) = 0$, $\phi_2'(0) = 1$, respectively.

In our case it is easy to arrive at

$$\phi_1(T) = \cos(aT_a)\cos(bT_b) - \frac{a}{b}\operatorname{sen}(aT_a)\operatorname{sen}(bT_b), \tag{6.3a}$$

$$\phi_2'(T) = \cos(aT_a)\cos(bT_b) - \frac{b}{a}\operatorname{sen}(aT_a)\operatorname{sen}(bT_b). \tag{6.3b}$$

Finally, the form of the discriminant is

$$D(a, b, T_a, T_b) = 2\cos(aT_a + bT_b) - \frac{(a-b)^2}{ab}\operatorname{sen}(aT_a)\operatorname{sen}(bT_b). \tag{6.4}$$

The Floquet theory for linear equations with periodical coefficients connects the stability of the solutions of (5.13a) with the value of the discriminant. The regions of resonance correspond to values of the parameters for which $|D| > 2$, whereas if $|D| < 2$, the solutions are bounded [44]. The equations $D(a, b, T_a, T_b) = 2$ and $D(a, b, T_a, T_b) = -2$ are the manifolds that limit the regions of stability in the four-dimensional space of parameters. In reality, defining $\alpha = aT$, $\beta = bT$, $T_a = \gamma T$, $T_b = (1-\gamma)T$, the number of parameters is reduced to three:

$$D(\gamma, \alpha, \beta) = 2\cos(\alpha\gamma + \beta(1-\gamma)) - \frac{(\alpha-\beta)^2}{\alpha\beta}\operatorname{sen}(\alpha\gamma)\operatorname{sen}(\beta(1-\gamma)). \tag{6.5}$$

Therefore, the isosurfaces $D(\gamma, \alpha, \beta) = 2$ and $D(\gamma, \alpha, \beta) = -2$ can be visualized in three dimensions, as is shown in Figure 6.2.

The general study of the regions that appear in Figure 6.2 is complex, which leads us to focus on a few particular cases below.

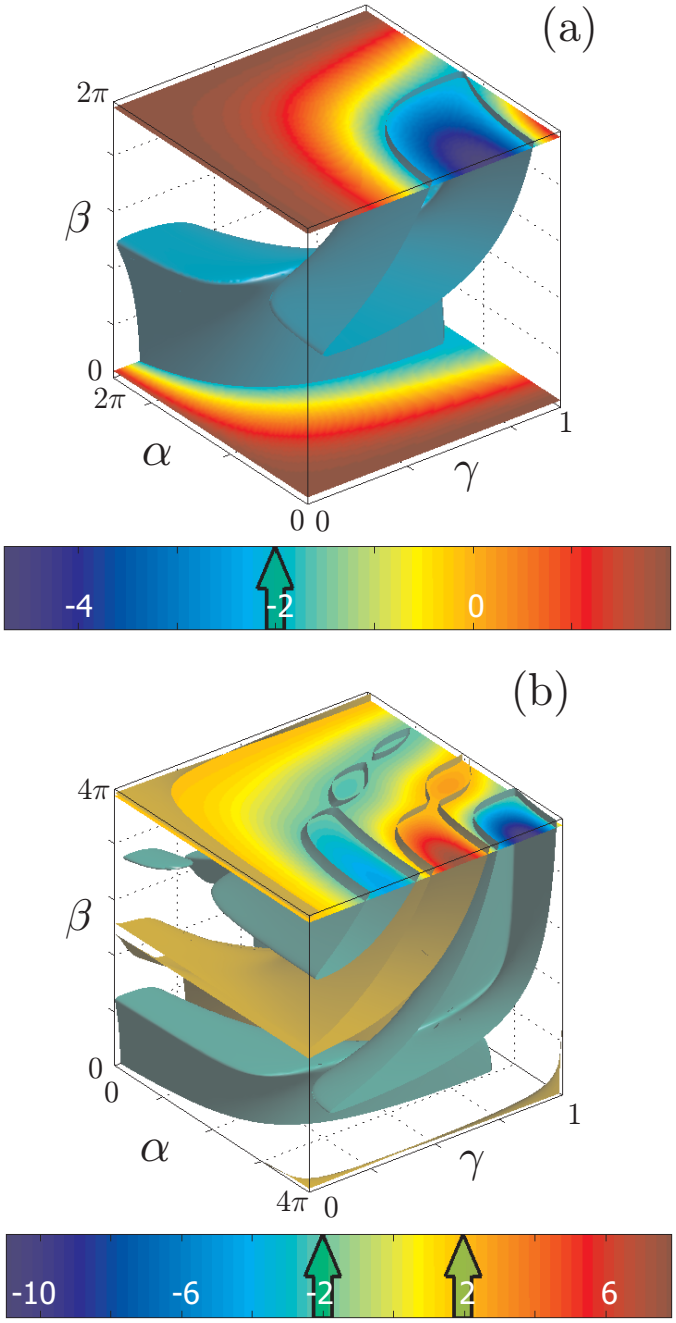


FIG. 6.2. (a) *Isosurfaces corresponding to* $D = -2$ *(in this range of values* $D \leq 2$*) for a limited range of parameters. The regions between the gray surface and the planes limiting the drawing are regions of resonance. Sections are shown for two particular values of* β*, where the bluish tones correspond to the regions of resonance. The color bar indicates the color corresponding to each level of* $D(\gamma, \alpha, \beta)$*, and the arrow indicates the color assigned to the isosurface* $D = -2$*.* (b) *The same as* (a) *but for a larger range of parameters. Isosurfaces* $D = 2$ *and* $D = -2$ *are shown in brown and green, respectively. A section is shown for a particular value of* β *with bluish and reddish tones corresponding to regions of resonances with* $D < -2$ *and* $D > 2$*, respectively. In this case the two values* $D = 2$ *and* $D = -2$ *are indicated by arrows on the color bar.*

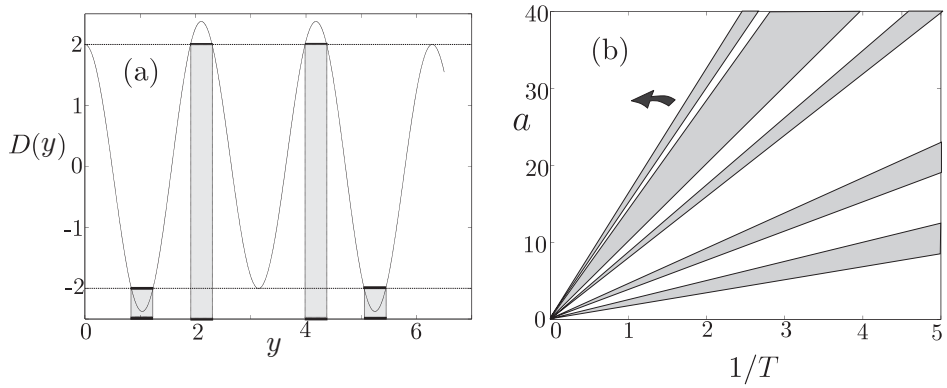


FIG. 6.3. (a) *First four regions of resonance in* y *(shaded) for* $T_a = T/2$, $b = 2a$, *as a function of* $y = aT/2$. (b) *First five regions of resonance in the plane* $a - 1/T$.

For example, in the case in which the two sections have the same length $T_a = T_b$, the discriminant depends only on aT, bT and it is

$$D(a,b) = 2\cos\left(\frac{(a+b)T}{2}\right) - \frac{(aT-bT)^2}{abT^2}\operatorname{sen}\left(\frac{aT}{2}\right)\operatorname{sen}\left(\frac{bT}{2}\right), \tag{6.6}$$

so that now the condition $|D| = 2$ determines curves such as those of Figure 6.1(b).

The structure of the regions of resonance can be explored in more detail, fixing the relative values of the coefficients; for example, taking $b = 2a$,

$$D(a,T) = 2\cos\left(\frac{3aT}{2}\right) - \frac{1}{2}\operatorname{sen}\left(\frac{aT}{2}\right)\operatorname{sen}(aT). \tag{6.7}$$

Defining the variable $y = aT/2$, the discriminant is a function $D(y)$; see Figure 6.3. The so-called characteristic curves are hyperbolas of the form $2y_{\pm}^{(n)} = aT$ with $y_{+}^{(n)}$ and $y_{-}^{(n)}$ being respectively the solutions of the algebraic equations $f_{\pm}(y) = 2\cos 3y - \frac{1}{2}\operatorname{sen} y \operatorname{sen} 2y \mp 2 = 0$. It is easy to demonstrate that the regions of resonance are contained between two consecutive zeros of f_+ or f_- that can be obtained using any elementary numerical method. If we draw a as a function of $1/T$, the regions of resonance are the shaded portions in Figure 6.3(b). Obviously the image is repeated due to the periodicity 2π of $D(y)$, and there are only four basic regions of resonance (together with its harmonics) contained in the intervals (roots of f_+ and f_-): $y \in [0.84, 1.23] \cup [1.91, 2.3] \cup [3.98, 4.37] \cup [5.05, 5.44]$ (see Figure 6.3(a)).

Another case of possible interest is that in which one of the fibers is not of GRIN type, that is, $b = 0$. Then the discriminant is given by the limit of (6.4) when $b \to 0$:

$$D(a,T) = -aT\operatorname{sen}\left(\frac{aT}{2}\right) + 2\cos\left(\frac{aT}{2}\right). \tag{6.8}$$

As in the previous case, the only relevant parameter is $y = aT/2$, the regions of resonance on the plane $a - T$ are hyperbolas, and the relevant quantities are the zeros of f_+ and f_-, which are given by

$$f_{\pm}(y) = -y\operatorname{sen} y + 2\cos y \mp 2 = 0. \tag{6.9}$$

Now there is no exact periodicity in the positions of the zeros any more, but at least it is possible to estimate the location of those of high order. To do this we must bear in mind that for y big enough the dominating term in both cases is $f_\pm(y) \simeq -y \operatorname{sen} y$, so that the zeros will be given by $y = n\pi$. It can be seen with a perturbative argument that the convergence ratio is of the order of $\mathcal{O}(1/n)$. Writing $y_\pm^{(n)} = n\pi + \varepsilon_\pm^{(n)}$ and substituting it into (6.9), it is found that

$$\varepsilon_\pm^{(n)} \simeq (-1)^{n+1} \frac{1 \pm 2}{n\pi}. \tag{6.10}$$

This type of analysis can be extended to any restricted set of parameters.

6.2. Dynamics of Bose–Einstein condensates. There has recently been great interest in the study of the dynamics of Bose–Einstein condensates in a parametrically oscillating potential. Recent experiments (see, e.g., [45, 46]) have motivated a series of qualitative theoretical analyses (the pioneer works on this subject can be seen in [47, 48, 49], although there is a great deal of subsequent literature).

In the models to which we refer, the trap is modified harmonically in time; that is,

$$\lambda^2(t) = 1 + \varepsilon \cos \omega t \tag{6.11}$$

with $\varepsilon > -1$. Equation (5.13a) with $\lambda(t)$ given by (6.11) is called the Mathieu equation. For this equation it is possible, as in the case of the Meissner equation, to carry out the study of the regions of the space of parameters in which resonances occur. In the first place, for any fixed ε, there exist two successions $\{\omega_n\}, \{\omega_n'\}$ with $\omega_n, \omega_n' \overset{n\to\infty}{\longrightarrow} 0$ such that if we take $\omega \in (\omega_n, \omega_n')$, (5.13a) possesses a resonance. In the second place, for fixed ω, the resonances appear when ε is large enough. The boundaries of those regions are the so-called characteristic curves that cannot be obtained explicitly but whose existence can be demonstrated analytically, as in the previous section, by using the discriminant. In the case of the Mathieu equation, it can be proven that the regions of instability begin in frequencies $\omega = 2, 1, 1/2, \ldots, 2/n^2, \ldots$ [44].

As in the previous case, the resonant behavior depends only on the parameters and not on the initial data. With respect to stability, the Massera theorem implies that if (ε, ω) is in a region of stability, then there exists a periodic solution of (5.13a), and by the nonlinear superposition principle, such a solution is stable in the sense of Lyapunov.

7. Approximate methods I: Quadratic phase approximation (QPA).

7.1. Introduction and justification of the QPA. Up to now, the results we have shown for the evolution of the solution moments are exact and in some sense rigorous. Unfortunately, in many situations of practical interest it is not possible to obtain closed evolution equations for the moments. In this section we will deal with an approximate method which is based on the method of moments.

The idea of this method is to approximate the phase of the solution u by a quadratic function of the coordinates, that is,

$$u(x,t) = U(x,t) \exp\left(i \sum_{j=1}^{n} \beta_j x_j^2 \right), \tag{7.1}$$

where $U(x,t)$ is a *real* function.

Why use a quadratic phase? Although there is not a formal justification and we do not know of any rigorous error bounds for the method to be presented here, there are several reasons which can heuristically support the use of this ansatz for the phase for situations where there are no essential shape changes of the solutions during the evolution. First of all, when (4.1) has self-similar solutions, they have exactly a quadratic phase [50]. Second, the dynamics of the phase close to stationary solutions of the classical cubic NLS equation in two spatial dimensions (critical case) is known to be approximated by quadratic phases [14, 16]. Finally, to capture the dynamics of the phase of solutions close to the stationary ones, which have a constant phase, by means of a polynomial fit, the terms of lowest order are quadratic since the linear terms in the phase may be eliminated by using Theorem 4.2.

For NLS equations all commonly used ansatzes in the framework of the previously mentioned variational methods have a quadratic phase, e.g., in applications related to dispersion management [51, 52], Bose–Einstein condensation, etc. Our systematic method provides a more general framework in which other methods can be systematized and understood.

As we will see in what follows, the choice (7.1) allows us to obtain explicit evolution equations and solves the problem of calculating the integrals of the phase derivatives in (4.6).

7.2. Modulated power-type nonlinear terms. Under the QPA, for modulated power-type nonlinearities $g(\rho,t) = g_0(t)\rho^{p/2}$, $p \in \mathbb{R}$, for which $\int_{\mathbb{R}^n} D(\rho)dx = -pJ/2$, the moment equations (4.6) are

$$\frac{dI_{2,j}}{dt} = V_{1,j}, \tag{7.2a}$$

$$\frac{dV_{1,j}}{dt} = 4K_j - 2\lambda_j^2 I_{2,j} + pJ, \tag{7.2b}$$

$$\frac{dK_j}{dt} = -\frac{1}{2}\lambda_j^2 V_{1,j} + p\beta_j J, \tag{7.2c}$$

$$\frac{dJ}{dt} = -p\left(\sum_{j=1}^{n}\beta_j\right)J + \frac{1}{g_0}\frac{dg_0}{dt}J. \tag{7.2d}$$

To these equations we must add the identity $V_{1,j} = 4\beta_j I_{2,j}$, which is directly obtained by calculating $V_{1,j}$. Or, expressed otherwise,

$$\beta_j = \frac{\dot{I}_{2,j}}{4I_{2,j}}. \tag{7.3}$$

Let us now consider the simplest case of solutions with spherical symmetry with $\lambda_j = \lambda(t)$, $j = 1,\dots,n$, for which $\phi(x_1,\dots,x_n) = \beta(t)\left(x_1^2 + \cdots + x_n^2\right)$. Using the same notation as in (5.2), the moment equations become

$$\frac{d\mathcal{I}}{dt} = \mathcal{V}, \tag{7.4a}$$

$$\frac{d\mathcal{V}}{dt} = 4\left(\mathcal{K} + \frac{np}{4}J\right) - 2\lambda^2\mathcal{I}, \tag{7.4b}$$

$$\frac{d\mathcal{K}}{dt} = -\frac{1}{2}\lambda^2\mathcal{V} + np\beta J, \tag{7.4c}$$

$$\frac{dJ}{dt} = -np\beta J + \frac{1}{g_0}\frac{dg_0}{dt}J \tag{7.4d}$$

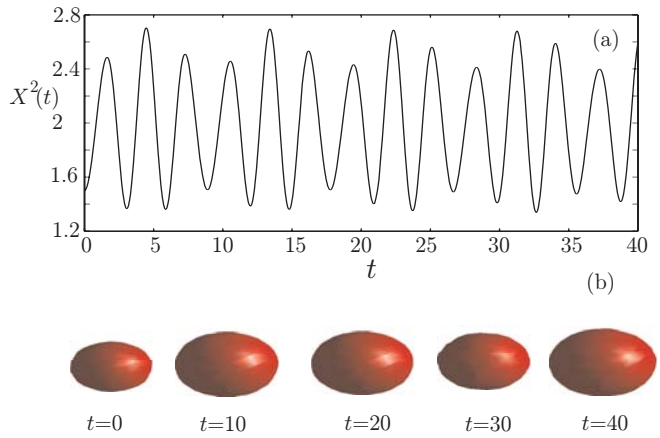


FIG. 7.1. *Solutions of* (4.1) *in three dimensions with* $p = 2$, $\lambda^2(t) = 1 + 0.1\sin(2.8t)$, *and* $g_0 = 10$ *with initial data* $u_0(x) = e^{-x^2/2}/\pi^{3/4}$. (a) $X^2(t)$ *obtained numerically from the solutions on the* 3*D grid.* (b) *Isosurfaces for* $|u|^2 = 0.02$ *on the spatial region* $[-3,3]\times[-3,3]\times[-3,3]$ *and different instants of time showing the oscillations of the solution.*

with $\mathcal{V} = 4\beta\mathcal{I}$. Despite the complexity of the system of equations (7.4) it is possible to find two positive invariants,

$$\mathcal{Q}_1 = 2\mathcal{K}\mathcal{I} - \mathcal{V}^2/4, \tag{7.5a}$$

$$\mathcal{Q}_2 = \frac{np}{2g_0}\mathcal{I}^{np/4}J. \tag{7.5b}$$

The existence of these invariants provides J as a function of $\mathcal{I}$, which allows us to arrive at an equation for $X = \mathcal{I}^{1/2}$,

$$\frac{d^2X}{dt^2} + \lambda^2(t)X = \frac{\mathcal{Q}_1}{X^3} + g_0(t)\frac{\mathcal{Q}_2}{X^{np/2+1}}. \tag{7.6}$$

Again we obtain a Hill's equation with a singular term. Note that in the case $n = 3$, $p = 2$ we have a quartic term in the denominator, which corresponds with the type of powers that appear in the equations which are obtained in the framework of averaged Lagrangian methods [20].

The quadratic phase method provides reasonably precise results that at least describe the qualitative behavior of the solutions of the partial differential equation. Using several numerical methods, we have carried out different tests especially in the most realistic case $np = 6$ in (4.1). For example, in Figure 7.1 we present the results of a simulation of (4.1) with $n = 3$, $p = 2$, $\lambda^2(t) = 1 + 0.1\sin(2.8t)$, and $g_0 = 10$ for an initial datum $u_0(x) = e^{-x^2/2}/\pi^{3/4}$. In this case the simplified equation (7.6) predicts quasi-periodic solutions, which is what we obtain when resolving the complete problem.

In Figure 7.2 we show the results for $\lambda^2(t) = 1 + 0.1\sin(2.1t)$, for which (7.6) predicts resonant solutions. Again, the results of the two models are in good agreement.

Another interesting application of the quadratic phase approximation method is the case of cubic nonlinearity, $g(\rho, t) = g_0(t)\rho = g_0(t)|u|^2$, without potential $\lambda(t) = 0$. In this situation (7.6) becomes

$$\frac{d^2X}{dt^2} = \frac{\mathcal{Q}_1}{X^3} + g_0(t)\frac{\mathcal{Q}_2}{X^{n+1}}, \tag{7.7}$$

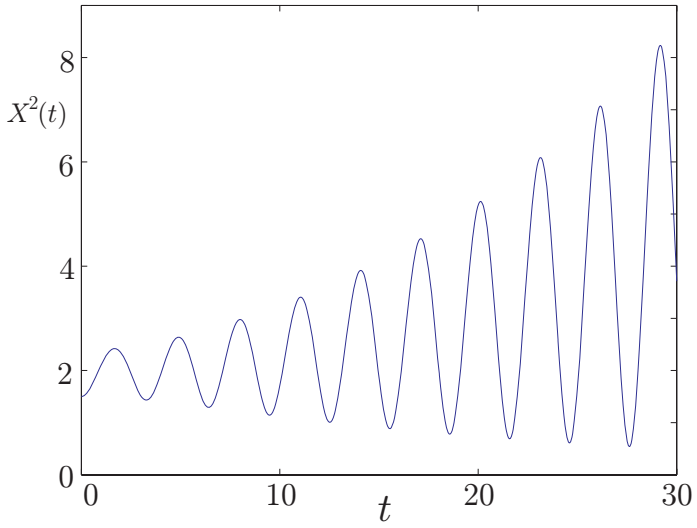


FIG. 7.2. *Results of the simulation of* (4.1) *in three dimensions on a grid of* $64 \times 64 \times 64$ *with* $\Delta t = 0.005$ *for* $p = 2$, $\lambda^2(t) = 1 + 0.1\sin(2.1t)$. $X^2(t)$ *obtained numerically from the solutions on the* $3D$ *grid is shown.*

where the conserved quantities are

$$\mathcal{Q}_1 = 2\mathcal{K}\mathcal{I} - \frac{\mathcal{V}^2}{4}, \tag{7.8a}$$

$$\mathcal{Q}_2 = \frac{n}{g_0}\mathcal{I}^{n/2}J. \tag{7.8b}$$

This model describes the propagation of light in nonlinear Kerr media as well as the dynamics of trapless Bose–Einstein condensates. In this situation the previous equations are used to study the possibility of stabilizing unstable solutions of the NLS equation by means of an appropriate temporal modulation of the nonlinear term, that is, by choosing a suitable function $g_0(t)$, thus providing an alternative to more heuristic treatments [53, 54, 55]. More details can be seen in [56].

7.3. Closure of the equations in other cases. We have just seen that the quadratic phase approximation method allows us to close the moment equations in the case of power-type nonlinear terms. Following those ideas, we have managed to close the equations in more general cases.

We start from the evolution equations for the mean moments (5.4), that after performing the quadratic phase approximation become

$$\frac{d\mathcal{I}}{dt} = \mathcal{V}, \tag{7.9a}$$

$$\frac{d\mathcal{V}}{dt} = 4\mathcal{K} - 2\lambda^2\mathcal{I} - 2n\int_{\mathbb{R}^n} D(\rho,t), \tag{7.9b}$$

$$\frac{d\mathcal{K}}{dt} = -\frac{1}{2}\lambda^2\mathcal{V} - 2n\beta\int_{\mathbb{R}^n} D(\rho,t), \tag{7.9c}$$

$$\frac{dJ}{dt} = 2n\beta\int_{\mathbb{R}^n} D(\rho,t) + \int_{\mathbb{R}^n}\frac{\partial G(\rho,t)}{\partial t}, \tag{7.9d}$$

and $\mathcal{V} = 4\beta\mathcal{I}$.

The idea to close the previous equations is to calculate the evolution of $\int_{\mathbb{R}^n} D(\rho,t)dx$, which is the term that prevents us from closing the equations, and try to write this evolution in terms of the moments. Let us define a new moment $\mathcal{F}$ as

$$\mathcal{F}(t) = \int_{\mathbb{R}^n} D(\rho,t)dx = J - \int_{\mathbb{R}^n} \rho \frac{\partial G(\rho,t)}{\partial \rho} dx. \tag{7.10}$$

Then, the evolution equations are

$$\frac{d\mathcal{I}}{dt} = \mathcal{V}, \tag{7.11a}$$

$$\frac{d\mathcal{V}}{dt} = 4\mathcal{K} - 2\lambda^2\mathcal{I} - 2n\mathcal{F}, \tag{7.11b}$$

$$\frac{d\mathcal{K}}{dt} = -\frac{1}{2}\lambda^2\mathcal{V} - 2n\beta\mathcal{F}, \tag{7.11c}$$

$$\frac{dJ}{dt} = 2n\beta\mathcal{F} + \int_{\mathbb{R}^n} \frac{\partial G}{\partial t} dx, \tag{7.11d}$$

together with the evolution of $\mathcal{F}$

$$\frac{d\mathcal{F}}{dt} = 2n\beta F + 2n\beta \int_{\mathbb{R}^n} \rho^2 \frac{\partial^2 G}{\partial \rho^2} dx + \int_{\mathbb{R}^n} \frac{\partial G}{\partial t} dx - \int_{\mathbb{R}^n} \rho \frac{\partial}{\partial t}\frac{\partial G}{\partial \rho} dx. \tag{7.12}$$

To try to close the system of equations (7.11)–(7.12) we impose that $\int_{\mathbb{R}^n} \rho^2 \frac{\partial^2 G}{\partial \rho^2} dx$ is a linear combination of $\mathcal{F}$ and J

$$\int_{\mathbb{R}^n} \rho^2 \frac{\partial^2 G}{\partial \rho^2} dx = a_{\mathcal{F}}\mathcal{F} + a_J J = (a_{\mathcal{F}} + a_J)\int_{\mathbb{R}^n} G dx - a_{\mathcal{F}} \int_{\mathbb{R}^n} \rho \frac{\partial G}{\partial \rho} dx, \tag{7.13}$$

where $a_{\mathcal{F}}$ and a_J are two arbitrary constants. Then G must verify

$$\int_{\mathbb{R}^n} \left[\rho^2 \frac{\partial^2 G}{\partial \rho^2} + a_{\mathcal{F}} \rho \frac{\partial G}{\partial \rho} - (a_{\mathcal{F}} + a_J) G \right] = 0.$$

Therefore, if the nonlinear term $g(\rho)$ in the NLS equation is such that $G(\rho)$ verifies Euler's equation

$$\rho^2 \frac{\partial^2 G}{\partial \rho^2} + a_{\mathcal{F}} \rho \frac{\partial G}{\partial \rho} - (a_{\mathcal{F}} + a_J) G = 0, \tag{7.14}$$

the evolution equations will close. In that case we can write $G(\rho,t) = g_0(t)G_1(\rho)$, where $g_0(t)$ is an arbitrary function which indicates the temporal variation of the nonlinear term and $G_1(\rho)$ satisfies (7.14). So

$$\int_{\mathbb{R}^n} \frac{\partial G}{\partial t} dx = \frac{dg_0}{dt} \int_{\mathbb{R}^n} G_1(\rho) dx = \frac{1}{g_0}\frac{dg_0}{dt} J(t),$$
$$\int_{\mathbb{R}^n} \rho \frac{\partial}{\partial t}\frac{\partial G}{\partial \rho} dx = \frac{dg_0}{dt} \int_{\mathbb{R}^n} \rho \frac{dG_1}{d\rho} = \frac{1}{g_0}\frac{dg_0}{dt}[J(t) - \mathcal{F}(t)],$$

and the moment equations are written as

$$\frac{d\mathcal{I}}{dt} = \mathcal{V}, \tag{7.15a}$$

$$\frac{d\mathcal{V}}{dt} = 4\mathcal{K} - 2\lambda^2\mathcal{I} - 2n\mathcal{F}, \tag{7.15b}$$

$$\frac{d\mathcal{K}}{dt} = -\frac{1}{2}\lambda^2\mathcal{V} - 2n\beta\mathcal{F}, \tag{7.15c}$$

$$\frac{dJ}{dt} = 2n\beta\mathcal{F} + \frac{1}{g_0}\frac{dg_0}{dt}J, \tag{7.15d}$$

$$\frac{d\mathcal{F}}{dt} = 2n\beta(1+a_{\mathcal{F}})\mathcal{F} + 2n\beta a_J J + \frac{1}{g_0}\frac{dg_0}{dt}\mathcal{F}. \tag{7.15e}$$

By solving (7.14), we obtain specific nonlinear terms for which the quadratic phase approximation allows us to write closed equations for the moments. Depending on the parameter $\delta = (1+a_{\mathcal{F}})^2 + 4a_J$, there exist three families of solutions

$$G_1(\rho) = \begin{cases} C_1\rho^{p_+} + C_2\rho^{p_-}, & \delta > 0, \\ C_1\rho^R + C_2\rho^R\log\rho, & \delta = 0, \\ C_1\rho^R\cos(I\log\rho) + C_2\rho^R\sin(I\log\rho), & \delta < 0, \end{cases} \tag{7.16}$$

where $p_\pm = \left((1-a_{\mathcal{F}}) \pm \delta^{1/2}\right)/2$, $R = (1-a_{\mathcal{F}})/2$, $I = |\delta|^{\frac{1}{2}}/2$.

The most interesting case for applications is $\delta > 0$, the nonlinear term being of the form

$$g_1(\rho) = k_1\rho^{p_+-1} + k_2\rho^{p_--1}, \tag{7.17}$$

where k_1 and k_2 are arbitrary constants and p_+ and p_- are defined through the relations

$$a_{\mathcal{F}} = 1 - p_+ - p_-, \tag{7.18a}$$

$$a_J = -(p_+ - 1)(p_- - 1). \tag{7.18b}$$

Equation (7.17) implies that the quadratic phase approximation allows us to close the moment equations for nonlinear terms, which can be written as a linear combination of two arbitrary powers of $|u|$.

As in the previous subsection, it is possible to find some invariant quantities, namely,

$$Q_1 = 2\mathcal{K}\mathcal{I} - \frac{\mathcal{V}^2}{4}, \tag{7.19a}$$

$$Q_+ = C\frac{n}{f_+}\frac{\mathcal{I}^{a+n}}{g_0}(J + f_+\mathcal{F}), \tag{7.19b}$$

$$Q_- = C\frac{n}{f_+}\frac{\mathcal{I}^{a-n}}{g_0}(J + f_-\mathcal{F}), \tag{7.19c}$$

where

$$a_\pm = \frac{p_\pm - 1}{2}, \quad f_\pm = \frac{1}{p_\mp - 1}, \quad C = \left(1 - \frac{f_-}{f_+}\right)^{-1} = \left(1 - \frac{p_- - 1}{p_+ - 1}\right)^{-1}. \tag{7.20}$$

These conserved quantities allow us to write a differential equation for the dynamical width $X(t) = \mathcal{I}^{1/2}$:

$$\frac{d^2X}{dt^2} + \lambda^2(t)X = \frac{\mathcal{Q}_1}{X^3} + g_0(t)\left(\frac{\mathcal{Q}_-}{X^{2a_-n+1}} - \frac{\mathcal{Q}_+}{X^{2a_+n+1}}\right). \tag{7.21}$$

The most interesting kind of nonlinearity in the form of (7.17) is the so-called cubic-quintic nonlinearity, for which $g_0(t) = 1$, $g_1(\rho) = k_1\rho + k_2\rho^2 = k_1|u|^2 + k_2|u|^4$. Then we have $p_+ - 1 = 2$, $p_- - 1 = 1$, $a_{\mathcal{F}} = -4$, $a_J = -2$, $a_+ = 1$, $f_+ = 1$, $a_- = 1/2$, $f_- = 1/2$, $C = 2$. The invariant quantities are

$$\mathcal{Q}_1 = 2\mathcal{K}\mathcal{I} - \frac{\mathcal{V}^2}{4}, \tag{7.22a}$$

$$\mathcal{Q}_+ = 2n\mathcal{I}^n(J + \mathcal{F}), \tag{7.22b}$$

$$\mathcal{Q}_- = 2n\mathcal{I}^{n/2}\left(J + \frac{\mathcal{F}}{2}\right), \tag{7.22c}$$

and the equation for the width is

$$\frac{d^2X}{dt^2} + \lambda^2(t)X = \frac{\mathcal{Q}_1}{X^3} + \frac{\mathcal{Q}_-}{X^{n+1}} - \frac{\mathcal{Q}_+}{X^{2n+1}}. \tag{7.23}$$

These equations contain a finite-dimensional description of the dynamics of localized solutions of the model and are similar to those found under specific assumptions for the profile $u(x,t)$ (see, e.g., [57, 58, 59]). The main difference is that the method of moments allows us to obtain the equations under minimal assumptions on the phase of the solutions and that depend on general integral quantities related to the initial data $\mathcal{Q}_1, \mathcal{Q}_+, \mathcal{Q}_-$. This is an essential advantage over the averaged Lagrangian methods used in the literature for which the specific shape of the solution must be chosen a priori (see also [20, 50]).

8. Approximate methods II: The Thomas–Fermi limit.

8.1. Concept. In the framework of the application of the NLS equations to Bose–Einstein condensation problems (and thus for nonlinearities of the form $g(\rho) = g_0\rho$), the Thomas–Fermi limit corresponds to the case $g_0 \gg 1$. (Note that this is only one of the many different meanings of "Thomas–Fermi" limit in physics.)

Usually, what is pursued in this context is to characterize the ground state, defined as the stationary solution of the NLS equation given by (4.5) with fixed L^2-norm having minimal energy E. It is also interesting to find the dynamics of the solutions under small perturbations of the ground state solution.

8.2. Physical treatment. Let us consider the problem of characterizing the ground state of (4.1). The usual "physical" way of dealing with this problem consists of assuming that if the nonlinear term is very large, then it would be possible to neglect the Laplacian term in (4.1) (!) and to obtain the ground state solution as

$$\varphi_{TF}(x) = \sqrt{\left(\frac{\mu - \frac{1}{2}\sum \lambda_j^2 x_j^2}{g_0}\right)_+}. \tag{8.1}$$

The value of μ is obtained from the condition of normalization $\|\varphi_{TF}\|_2 = 1$. This procedure provides a solution without nodes, which is then argued to be an approximation to the ground state.

This method is used in many applied works, but unfortunately it is not even self-consistent. Near the zero of the radicand of (8.1) the approximation obtained has divergent derivatives, which contradicts the initial hypothesis of "smallness" of the Laplacian term. Although several numerical results can be obtained using this approximation, its foundation is very weak.

In order to understand the problem better, we rewrite (4.1) making the change of variables $\kappa = \mu/g_0$, $\eta = x/\sqrt{g_0}$, $\psi(\eta) = \varphi\left(x/\sqrt{g_0}\right)$, to give us the equation

$$-\frac{1}{2}\epsilon^2\Delta\psi + \frac{1}{2}\left(\sum_j \lambda_j^2\eta_j^2\right)\psi + |\psi|^2\psi = -\kappa\psi, \tag{8.2}$$

with $\epsilon = 1/g_0$. It is evident that $\epsilon^2\Delta\psi$ is a singular perturbation whose effect may not be trivial.

8.3. The method of moments and the Thomas–Fermi limit. What can be said for the case of power-type nonlinearities in the limit $g \gg 1$ on the basis of the method of moments? Before making any approximations we write an evolution equation for $\mathcal{I}$ as follows. For the sake of simplicity, though it is not strictly necessary, we will consider the case of $\lambda_j = \lambda$ for $j = 1, \ldots, n$ and study the equations for the mean values (5.2).

First, we write (5.3a) and (5.3b) as

$$\frac{d\mathcal{I}}{dt} = \mathcal{V}, \tag{8.3a}$$

$$\frac{d\mathcal{V}}{dt} = 4\left(\mathcal{K} + \frac{np}{4}J\right) - 2\lambda^2\mathcal{I} = (4-np)\mathcal{K} + np\mathcal{H} - \left(2+\frac{np}{2}\right)\lambda^2\mathcal{I}, \tag{8.3b}$$

where $\mathcal{H}$ is the conserved energy. Combining (8.3a) and (8.3b), we arrive at

$$\frac{d^2\mathcal{I}}{dt^2} + \left(2+\frac{np}{2}\right)\lambda^2\mathcal{I} = (4-np)\mathcal{K} + np\mathcal{H}. \tag{8.4}$$

Equation (8.4) is exact.

The fact that the energy functional E reaches a minimum over φ_0 implies, by Lyapunov stability, that initial data $u_0(x) = \varphi_0(x) + \varepsilon\delta(x)$ close to the ground state must remain proximal for sufficiently small values of ε.

The only approximation needed to complete our analysis is to assume that when $g \gg 1$, then $J \gg \mathcal{K}$ for the ground state. Notice that this is a much more reasonable assumption than the direct elimination of the second derivative in the evolution equation. Thus, the energy conservation and the previous considerations allow us to affirm that $J(t) \gg \mathcal{K}(t)$ for all times.

Although these facts can be used to write explicit bounds for $\mathcal{K}$, as a first approximation and just in order to show the power of these ideas we can simply take $\mathcal{K} \simeq 0$. Under this approximation we have

$$\frac{d^2\mathcal{I}}{dt^2} + \left(2+\frac{np}{2}\right)\lambda^2\mathcal{I} \approx np\mathcal{H}, \tag{8.5}$$

whose solutions can be obtained explicitly as

$$\mathcal{I}(t) \simeq \frac{np\mathcal{H}}{\lambda^2\left(2+\frac{np}{2}\right)} + A\cos\left(\lambda t\sqrt{2+\frac{np}{2}}\right) + B\,\mathrm{sen}\left(\lambda t\sqrt{2+\frac{np}{2}}\right). \tag{8.6}$$

The equilibrium point of (8.4) (corresponding to $A = B = 0$) gives us the "size" of the ground state as a function of the physical parameters. Also the frequency of the oscillations around the equilibrium point is immediately obtained from (8.6):

$$\Omega = \lambda\sqrt{2 + \frac{np}{2}}. \tag{8.7}$$

We have performed numerical simulations of the partial differential equations (4.1) to verify this prediction. Specifically, taking $g = 5000, 20000$, $\lambda = 1$, and initial data of the form $u_0(x) = \varphi_0((1+\varepsilon)x)/\sqrt{1+\varepsilon}$ for $\varepsilon = 0.01$ and $\varepsilon = 0.02$, we find a numerical frequency of $\Omega_{\rm num} = 2.26$, which is in excellent agreement with the value provided by our Thomas–Fermi formula $\Omega_{\rm TF} = \sqrt{8} = 2.24$.

9. Summary and conclusions. In this paper we have developed the method of moments for nonlinear Schödinger equations. First we have found the general expressions of the method and classified the nonlinearities for which it allows a closed explicit solution of the evolution of the moments. We have also discussed several applications of the method such as the dynamics of Kerr beams in nonlinear stratified media and the dynamics of parametrically forced Bose–Einstein condensates.

Approximate techniques based on the method of moments have also been discussed in this paper. In particular, the quadratic phase approximation was developed here and applied to different problems, such as the writing of simple equations describing the stabilization of solitonic structures by control of the nonlinear terms and the dynamics of localized structures in cubic-quintic media. Finally, we have also studied the moment equations in the so-called Thomas–Fermi limit.

REFERENCES

[1] L. Vázquez, L. Streit, and V. M. Pérez-García, eds., *Nonlinear Klein-Gordon and Schrödinger Systems: Theory and Applications*, World Scientific, Singapur, 1997.

[2] F. Brezzi and P. A. Markowich, *The three-dimensional Wigner-Poisson problem: Existence, uniqueness and approximation*, Math. Model Methods Appl. Sci., 14 (1991), pp. 35–61.

[3] J. L. López and J. Soler, *Asymptotic behaviour to the 3D Schrödinger/Hartree-Poisson and Wigner-Poisson systems*, Math. Model Methods Appl. Sci., 10 (2000), pp. 923–943.

[4] Y. Kivshar and G. P. Agrawal, *Optical Solitons: From Fibers to Photonic Crystals*, Academic Press, New York, 2003.

[5] A. Hasegawa, *Optical Solitons in Fibers*, Springer-Verlag, Berlin, 1989.

[6] R. K. Dodd, J. C. Eilbeck, J. D. Gibbon, and H. C. Morris, *Solitons and Nonlinear Wave Equations*, Academic Press, New York, 1982.

[7] J. L. Rosales and J. L. Sánchez-Gómez, *Nonlinear Schödinger equation coming from the action of the particle's gravitational field on the quantum potential*, Phys. Lett. A, 166 (1992), pp. 111–115.

[8] R. Fedele, G. Miele, L. Palumbo, and V. G. Vaccaro, *Thermal wave model for nonlinear longitudinal dynamics in particle accelerators*, Phys. Lett. A, 173 (1993), pp. 407–413.

[9] F. Dalfovo, S. Giorgini, L. P. Pitaevskii, and S. Stringari, *Theory of Bose-Einstein condensation in trapped gases*, Rev. Mod. Phys., 71 (1999), pp. 463–512.

[10] A. S. Davydov, *Solitons in Molecular Systems*, Reidel, Dordrecht, The Netherlands, 1985.

[11] A. Scott, *Nonlinear Science: Emergence and Dynamics of Coherent Structures*, Oxf. Appl. Eng. Math. 1, Oxford University Press, Oxford, UK, 1999.

[12] V. E. Zaharov, V. S. L'vov, and S. S. Starobinets, *Turbulence of spin-waves beyond threshold of their parametric excitation*, Sov. Phys. Usp., 17 (1975), pp. 896–941.

[13] C. Sulem and P. Sulem, *The Nonlinear Schrödinger Equation: Self-Focusing and Wave Collapse*, Springer, Berlin, 2000.

[14] G. Fibich and G. Papanicolau, *Self-focusing in the perturbed and unperturbed nonlinear Schrödinger equation in critical dimension*, SIAM J. Appl. Math., 60 (1999), pp. 183–240.

[15] V. M. Pérez-García, M. A. Porras, and L. Vázquez, *The nonlinear Schrödinger equation with dissipation and the moment method*, Phys. Lett. A, 202 (1995), pp. 176–182.

[16] G. Fibich, *Self-focusing in the damped nonlinear Schrödinger equation*, SIAM J. Appl. Math., 61 (2001), pp. 1680–1705.
[17] Z.-Q. Wang, *Existence and symmetry of multibump solutions for nonlinear Schrödinger equations*, J. Differential Equations, 159 (1999), pp. 102–137.
[18] G. Velo, *Mathematical aspects of the nonlinear Schrödinger equation*, in Nonlinear Klein–Gordon and Schrödinger Systems: Theory and Applications, L. Vázquez, L. Streit, and V. M. Pérez-García, eds., World Scientific, Singapore, 1996, pp. 39–67.
[19] D. G. de Figueiredo and Y. H. Ding, *Solutions of a nonlinear Schrödinger equation*, Discrete Contin. Dyn. Syst. Ser. B, 3 (2002), pp. 563–584.
[20] B. Malomed, *Variational methods in nonlinear fiber optics and related fields*, Progr. Optics, 43 (2002), pp. 70–191.
[21] A. Sánchez and A. R. Bishop, *Collective coordinates and length-scale competition in spatially inhomogeneous soliton-bearing equations*, SIAM Rev., 40 (1998), pp. 579–615.
[22] V. I. Talanov, *Focusing of light in cubic media*, JETP Lett., 11 (1970), pp. 199–203.
[23] M. A. Porras, J. Alda, and E. Bernabeu, *Nonlinear propagation and transformation of arbitrary laser beams by means of the generalized ABCD formalism*, Appl. Optim., 32 (1993), pp. 5885–5892.
[24] Y. Oh, *Cauchy problem and Ehrenfest's law of nonlinear Schrödinger equations with potentials*, J. Differential Equations, 81 (1989), pp. 255–274.
[25] E. H. Lieb and M. Loss, *Analysis*, Grad. Stud. Math. 14, AMS, Providence, RI, 1996.
[26] N. Smith, F. M. Knox, N. J. Doran, K. J. Blow, and I. Bennion, *Enhanced power solitons in optical fibres with periodic dispersion management*, Electron. Lett., 32 (1996), pp. 54–55.
[27] I. Gabitov, E. Shapiro, and S. Turitsyn, *Asymptotic breathing pulse in optical transmission systems with dispersion compensation*, Phys. Rev. E, 55 (1997), pp. 3624–3633.
[28] S. Kumar and A. Hasegawa, *Quasi-soliton propagation in dispersion-managed optical fibers*, Opt. Lett., 22 (1997), pp. 372–374.
[29] S. Turitsyn, *Stability of an optical soliton with Gaussian tails*, Phys. Rev. E, 56 (1997), pp. R3784–R3787.
[30] J. J. García-Ripoll, V. M. Pérez-García, and P. Torres, *Extended parametric resonances in nonlinear Schrödinger systems*, Phys. Rev. Lett., 83 (1999), pp. 1715–1718.
[31] V. M. Pérez-García, P. Torres, J. J. García-Ripoll, and H. Michinel, *Moment analysis of paraxial propagation in a nonlinear graded index fiber*, J. Opt. B Quantum Semiclass. Opt., 2 (2000), pp. 353–358.
[32] V. M. Pérez-García, H. Michinel, and H. Herrero, *Bose-Einstein solitons in highly asymmetric traps*, Phys. Rev. A, 57 (1998), pp. 3837–3842.
[33] J. J. García-Ripoll, V. M. Pérez-García, and V. Vekslerchik, *Construction of exact solutions by spatial translations in inhomogenous nonlinear Schrödinger equations*, Phys. Rev. E, 64 (2001), paper 056602.
[34] M. A. Porras, J. Alda, and E. Bernabeu, *Nonlinear propagation and transformation of arbitrary laser beams by means of the generalized ABCD formalism*, Appl. Opt., 32 (1993), pp. 5885–5892.
[35] A. Gammal, T. Frederico, L. Tomio, and F. Kh. Abdullaev, *Stability analysis of the D-dimensional nonlinear Schrödinger equation with trap and two- and three-body interactions*, Phys. Lett. A, 267 (2000), pp. 305–311.
[36] E. B. Kolomeisky, T. J. Newman, J. P. Straley, and X. Qi, *Low dimensional Bose liquids: Beyond the Gross-Pitaevskii approximation*, Phys. Rev. Lett., 85 (2000), pp. 1146–1149.
[37] Yu. B. Gaididei, J. Schjodt-Eriksen, and P. Christiansen, *Collapse arresting in an inhomogeneous quintic nonlinear Schrödinger model*, Phys. Rev. E, 60 (1999), pp. 4877–4890.
[38] M. I. Weinstein, *On the structure and formation of singularities in solutions to nonlinear dispersive evolution equations*, Comm. Partial Differential Equations, 11 (1986), pp. 545–565.
[39] V. P. Ermakov, *Transformation of differential equations*, Univ. Izv. Kiev., 20 (1880), pp. 1–19.
[40] E. Pinney, *The nonlinear differential equation* $y'' + p(x)y + cy^{-3} = 0$, Proc. Amer. Math. Soc., 1 (1950), p. 681.
[41] J. L. Reid and J. R. Ray, *Ermakov systems, nonlinear superposition and solutions of nonlinear equations of motion*, J. Math. Phys., 21 (1980), pp. 1583–1587.
[42] C. Rogers and W. K. Schief, *Multi-component Ermakov systems: Structure and linearization*, J. Math. Anal. Appl., 198 (1990), pp. 194–220.
[43] M. Plum and R. M. Redheffer, *A class of second-order differential equations*, J. Differential Equations, 154 (1999), pp. 454–469.
[44] W. Magnus and S. Winkler, *Hill's Equation*, Dover Publications, New York, 1966.
[45] D. S. Jin, J. R. Ensher, M. R. Matthews, C. E. Wieman, and E. A. Cornell, *Collective*

excitations of a Bose-Einstein condensate in a dilute gas, Phys. Rev. Lett., 77 (1996), pp. 420–423.

[46] M.-O. Mewes, M. R. Andrews, N. J. van Druten, D. M. Kurn, D. S. Durfee, C. G. Townsend, and W. Ketterle, *Collective excitations of a Bose-Einstein condensate in a magnetic trap*, Phys. Rev. Lett., 77 (1996), pp. 988–991.

[47] S. Stringari, *Collective excitations of a trapped Bose-condensed gas*, Phys. Rev. Lett., 77 (1996), pp. 2360–2363.

[48] V. M. Pérez-García, H. Michinel, J. I. Cirac, M. Lewenstein, and P. Zoller, *Low energy excitations of a Bose-Einstein condensate: A time-dependent variational analysis*, Phys. Rev. Lett., 77 (1996), pp. 5320–5323.

[49] V. M. Pérez-García, H. Michinel, J. I. Cirac, M. Lewenstein, and P. Zoller, *Dynamics of Bose-Einstein condensates: Variational solutions of the Gross-Pitaevskii equations*, Phys. Rev. A, 56 (1997), pp. 1424–1432.

[50] V. M. Pérez-García, *Self-similar solutions and collective coordinate methods for nonlinear Schrödinger equations*, Phys. D, 191 (2004), pp. 211–218.

[51] M. Matsumoto, A. Matsumoto, and A. Hasegawa, *Optical Solitons in Fibers*, Springer, New York, 2002.

[52] B. A. Malomed, *Soliton Management in Periodic Systems*, Springer, New York, 2006.

[53] H. Saito and M. Ueda, *Dynamically stabilized bright solitons in a two-dimensional Bose-Einstein condensate*, Phys. Rev. Lett., 90 (2003), paper 040403.

[54] F. Abdullaev, J. G. Caputo, R. A. Kraenkel, and B. A. Malomed, *Controlling collapse in Bose-Einstein condensates by temporal modulation of the scattering length*, Phys. Rev. A, 67 (2003), paper 013605.

[55] I. Towers and B. Malomed, *Stable* $(2+1)$*-dimensional solitons in a layered medium with sign-alternating Kerr nonlinearity*, J. Opt. Soc. Amer. B Opt. Phys., 19 (2002), pp. 537–543.

[56] G. D. Montesinos, V. M. Pérez-García, and P. Torres, *Stabilization of solitons of the multidimensional nonlinear Schrödinger equation: Matter-wave breathers*, Phys. D, 191 (2004), pp. 193–210.

[57] F. Abdullaev, A. Gammal, L. Tomio, and T. Frederico, *Stability of trapped Bose-Einstein condensates*, Phys. Rev. A, 63 (2001), paper 043604.

[58] Z. Jovanoski, *Gaussian beam propagation in d-dimensional cubic-quintic nonlinear medium*, J. Nonlinear Opt. Phys. Materials, 10 (2001), pp. 79–111.

[59] H. Michinel, J. Campo-Taboas, R. García-Fernández, J. R. Salgueiro, and M. L. Quiroga-Teixeiro, *Liquid light condensates*, Phys. Rev. E, 65 (2002), paper 066604.

SIAM J. APPL. MATH.
Vol. 67, No. 4, pp. 1016–1040

THE FORMATION OF RIVER CHANNELS*

A. C. FOWLER†, NATALIA KOPTEVA‡, AND CHARLES OAKLEY†

Abstract. We consider a deterministic model of landscape evolution through the mechanism of overland flow over an erodible substrate, using the St. Venant equations of hydraulics together with the Exner equation for hillslope erosion. A novelty in the model is the allowance for a nonzero bedload layer thickness, which is necessary to distinguish between transport limited and detachment limited sediment removal. It has long been known that transport limited uniform flow is unstable when the hillslope topography is geomorphologically concave (i.e., the center of curvature is above ground). In this paper, we show how finite amplitude development of the consequent channel flow leads to an evolution equation for its depth h of the form $h_t = h^{3/2} + (h^{3/2})_{YY}$, where Y is the cross-stream space variable. We show that solutions of compact support exist but that, despite appearances, blow up does not occur because of an associated integral constraint, and the channel equation admits a unique and apparently globally stable steady state. The consequences for hillslope evolution models are discussed.

Key words. river networks, mathematical geomorphology, channel formation, nonlinear diffusion

AMS subject classifications. 86A99, 35K55, 35K65

DOI. 10.1137/050629264

1. Introduction.

The formation of river networks is one of a class of morphological problems in which fractal structures are generated by an instability in the medium. Other familiar examples are the lungs, blood capillary beds, and underground limestone cave systems. Two questions immediately present themselves in connection with such structures. The first is whether it is possible to explain quantitatively the basic mechanisms which are involved in causing them to form. The second is the consequent deeper issue of whether it is possible to explain and predict the fractal structures which are observed in nature, given that the model will originate as a deterministic set of differential equations. In this paper, we will be concerned with the first of these questions.

The basic way in which landscape evolves under fluvial erosion is this. Tectonic processes cause uplift of mountain belts, and as the mountains are raised, erosion due to rainfall and runoff causes a gradual lowering of the topography. Other processes, such as glacial erosion and landslides, contribute more dramatically: glaciation at high altitudes, and landslides in regions of higher relief. As is evident from Figure 1, this balance between uplift and erosion is unstable, and the runoff is concentrated into small river channels which drain the catchment.

In attempting to formulate a model to describe this process, we identify two variables of importance; these are the surface elevation s and the water depth h (Figure 2). These will be described by evolution equations representing conservation

*Received by the editors April 15, 2005; accepted for publication (in revised form) December 7, 2006; published electronically May 10, 2007.
http://www.siam.org/journals/siap/67-4/62926.html
†Mathematical Institute, Oxford University, 24-29 St Giles', Oxford OX1 3LB, England (fowler@maths.ox.ac.uk, oakley@maths.ox.ac.uk). The first author was supported by the University of Limerick to maintain his position as Adjunct Professor. The third author was supported by the EPSRC via a postgraduate studentship.
‡Department of Mathematics and Statistics, University of Limerick, Limerick, Republic of Ireland (natalia.kopteva@ul.ie).

FIG. 1. *Hillslope topography. Photograph courtesy of Gary Parker.*

of sediment and water, respectively.

Smith and Bretherton (1972) presented such a model and found that while there is a uniform steady state solution, it is unstable to the formation of channel-like features. In particular, they associated instability with concavity of the hillslope, i.e., $s_{xx} > 0$, where x is the downslope direction of flow.

The particular way in which this instability is manifested is curious. The physical mechanism is plain enough, that increasing depth causes increased water flow, which in turn causes increased erosion and thus channel deepening. In their linear stability analysis, Smith and Bretherton found that the mathematical cause of instability was an effective lateral diffusion coefficient for hillslope which was negative. This naturally produces instability, but the resulting growth rate is unbounded at short wavelength, and their model is consequently ill-posed. Unsurprisingly, properly resolved numerical solutions of the Smith–Bretherton model are not available.

Another consequence of this ill-posedness is a suspicious absence of wavelength selection. Loewenherz (1991) addressed this issue by carrying out a formal linear stability analysis using normal modes (something Smith and Bretherton did not do), and she extended this to convex/concave slopes using the asymptotic technique of WKB theory (Carrier, Krook, and Pearson (1966)) at high wave number k. She also considered the problem of regularization as $k \to \infty$, by introduction of a (fairly arbitrary) modification to the sediment transport law.

Later (Loewenherz-Lawrence (1994)), she treated the whole problem again, but now starting from the hydrodynamic theory, which is also the starting point for the model we present below. In this way, she was able to identify the cause of the ill-posedness of the Smith–Bretherton theory, which lies in the assumption of equal water and land surface slopes. The small mismatch between these two allows regularization at high wave number, and therefore also wavelength selection.

A different approach to the issue of wavelength selection was taken by Izumi and Parker (1995, 2000), who used a St. Venant overland flow model together with a finite

threshold stress for the onset of erosion to show that there is a preferred wavelength for instability. Their estimate in the earlier paper was 33 m, comparable to observed headwater spacings of order 100 m. A formal stability analysis in the second paper (of a slightly different problem) yielded plausibly similar values.

The next logical steps in the development of this theory are a nonlinear theory for finite depth channel development, and full numerical solution of the governing equations. Progress in the first of these aims was made by Kramer and Marder (1992), who developed a nonlinear evolution equation for channel depth by seeking particular solutions of their hillslope model, which was similar, but by no means identical to, the Smith–Bretherton model. The main difference between their result and that of the present paper is that their model is partially empirical, and the derivation of the channel model is not placed in the context of a formal asymptotic approximation to the full model. This leads to important differences in the way the channel evolution equation is posed.

Kramer and Marder also sought to implement a direct numerical simulation, but here, in common with other authors, they were stymied. The apparent reason for this is that the governing partial differential equations are very stiff in both space and time. Water flow in channels occurs on much shorter space and time scales than hillslope evolution, and such numerical computational studies as there have been have not been able to overcome this difficulty.

In response to this, they adopted a cellular lattice model, with physically motivated rules at the lattice points determining the evolution of water depth and land surface elevation. Such cellular models do produce networks but evidently lack a theoretically based predictive capacity. To a large extent, they provide the computational model of choice for other researchers also (e.g., Howard (1994), Tucker and Slingerland (1994)).

A variant on this was the model developed by Willgoose, Bras, and Rodríguez-Iturbe (1991), which combined physically based erosion and water flow equations with an artificial equation for an indicator function Y. Essentially, Y would switch from $Y = 0$ (hillslope) to $Y = 1$ (channel) when water flow increased beyond a critical threshold. In this way, Willgoose et al. could simulate network formation but again without a physically based predictive criterion.

In a sequence of papers, Smith and his coworkers have developed a semianalytic theory of hillslope and channel evolution. Their work is actually orthogonal to the present paper but will be discussed in some detail here because of the apparent parallelism with our work. Smith, Birnir, and Merchant (1997a) consider a simplified version of the Smith–Bretherton model, and use it to suggest that large time solutions have separable form, which they are able to characterize in terms of a variational principle. Smith, Birnir, and Merchant (1997b) elaborate this description by suggesting that an initially smooth hillslope develops channels on a small scale through the Smith–Bretherton instability; the channels saturate via nonlinearity and then evolve into the long time separable solutions described earlier. These results are obtained numerically. In order to obtain numerical results for the ill-posed Smith–Bretherton model, Smith, Birnir, and Merchant (1997b) used a coarse grid on a small plot (100 m by 100 m with grid spacing 1 m), together with enough numerical diffusion to stabilize the results. Smith, Merchant, and Birnir (2000) develop a theory for the time evolution of the grade line of both alluvial and bedrock channels; the former is modelled by a nonlinear diffusion equation, and the latter is modelled by a nonlinear first-order wave equation. Both theories ignore hillslope evolution and make heuristic assumptions in order to derive the models. Birnir, Smith, and Merchant (2001) develop the

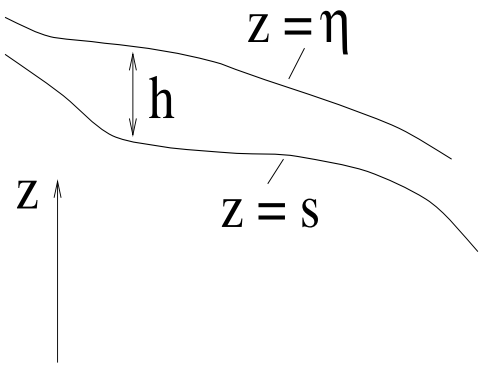


FIG. 2. *Geometry of overland flow.*

ideas originated in the earlier papers by Smith, Birnir, and Merchant (1997a, 1997b). They paint a fairly compelling picture of landscape evolution, which hinges on the twin hypotheses that small scale shock formation in overland flow acts as a seed for white noise to drive the slower hillslope evolution towards a self-similar (separable in time) mature landscape. Crucial to this notion is the assumption that the numerical results are sufficiently detailed to support it. The numerical procedures are improved over those of Smith, Birnir, and Merchant (1997b), but apparently retain the small plot and coarse grid of the earlier calculations, and are therefore open to the same objection, that the coarse grid in particular allows only mildly unstable results by suppressing the high wave number instabilities. The paper by Welsh, Birnir, and Bertozzi (2006) is similar to that of Smith, Merchant, and Birnir (2000), insofar as it uses the Smith–Bretherton model to assess the evolution of the long profile of a river channel. To do this, it assumes a purely one-dimensional model, so that the channel evolves in isolation from the surrounding hillslope.

Our purpose in this paper is to show that a hydrodynamic model similar to those of Loewenherz-Lawrence (1994) and Tucker and Slingerland (1994) leads formally to the derivation of an evolution equation for channel depth (which resembles that of Kramer and Marder). The solution properties of this equation are studied, and it is shown that, despite a similarity of the channel equation to partial differential equations having blow-up properties, there is a unique steady state solution which is stable. This solution may provide an ingredient for future direct numerical simulations of hillslope evolution.

2. A model for sediment and water transport. The geometric situation we consider is portrayed in Figure 2. The vertical coordinate is z, while x and y are horizontal coordinates. The simplest situation is where overland flow occurs down a plane slope, and in this case we take x in the downstream direction and y across stream. The land surface is $z = s(x, y, t)$, the water surface is $z = \eta(x, y, t)$, and the water depth is h, and thus $h = \eta - s$. This relationship is not exact, because the sedimentary surface is further subdivided into a mobile part and a stationary part. A precise statement is given below in (2.9).

The St. Venant equations of hydraulic flow can be written in the form

$$\begin{aligned} &h_t + \boldsymbol{\nabla}.(h\mathbf{u}) = r, \\ &\mathbf{u}_t + (\mathbf{u}.\boldsymbol{\nabla})\mathbf{u} = -g\boldsymbol{\nabla}\eta - \frac{f|\mathbf{u}|\mathbf{u}}{h}. \end{aligned} \tag{2.1}$$

These represent conservation of water mass and momentum and can be derived from the vertically integrated point forms of the equations. r is the source due to rainfall, $\mathbf{u}$ is the mean velocity, and f is a friction factor in a term which represents the bed stress exerted by the flow, assuming this is turbulent. While this is a good parameterization of the bed friction in channelized flow, it is less obviously appropriate for the very thin films which characterize overland flow. We shall comment further on this below, but for the moment we note that consideration of laminar flow at low flow rates would simply have the effect in the model of changing the term $f|\mathbf{u}|$ in $(2.1)_2$ to a constant k, making quantitative but not conceptual difference to the discussion.

Sediment transport. Sediment transport in rivers occurs, for noncohesive sediments with little clay content, when an appropriately dimensionless shear stress (called the Shields stress) delivered by the river exceeds a certain critical value. The turbulent shear stress is taken to be

$$\boldsymbol{\tau} = f\rho_w|\mathbf{u}|\mathbf{u}, \tag{2.2}$$

where ρ_w is water density. If the sediment particles are of diameter D_s (supposed uniform, for simplicity) at the bed, the streamflow exerts a force of approximately $\boldsymbol{\tau}D_s^2$ on it, and it is this force which causes motion. On a slope, there is an additional force due to gravity, approximately $-\Delta\rho g D_s \boldsymbol{\nabla} s$, where $\Delta\rho = \rho_s - \rho_w$ is the density difference between sediment and water, and g is gravitational acceleration. Thus the net effective stress causing motion is actually

$$\boldsymbol{\tau}_e = \boldsymbol{\tau} - \Delta\rho g D_s \boldsymbol{\nabla} s. \tag{2.3}$$

The Shields stress is

$$\mu = \frac{\tau_e}{\Delta\rho g D_s}, \tag{2.4}$$

and particle motion occurs if $\mu \gtrsim \mu_c \approx 0.05$; the critical value depends to some extent on particle size via the particle Reynolds number.

Particle motion occurs in two ways. Larger particles bounce and roll along the bed, and the resultant transport is called bedload transport. Finer particles are lifted up and carried in suspension. In this paper, we will suppose that only bedload transport is relevant. This assumption is made partly for convenience, partly because it corresponds to the choice of Smith and Bretherton (1972), and partly because it may be an unnecessary elaboration to consider suspended load instead or as well.

Various empirical formulae for bedload transport q_b have been proposed. A popular one is that due to Meyer-Peter and Müller (1948), which takes the form

$$q_b = \left(\frac{\rho_s K}{\rho_w^{1/2}\Delta\rho\, g}\right)(\tau_e - \tau_c)_+^{3/2}, \tag{2.5}$$

where Meyer-Peter and Müller chose values of $K = 8$ and $\mu_c = 0.047$, and the critical stress τ_c is defined by

$$\tau_c = \mu_c \Delta\rho g D_s. \tag{2.6}$$

The units of q_b are kg m^{-1} s^{-1}, i.e., mass per unit stream width per unit time.

It is commonly the case that bedload transport is conceived to occur in a layer of zero thickness, if this is considered at all. Although the moving bedload layer

thickness may indeed be small, it is essential to include it in the model (as did Tucker and Slingerland (1994)), because otherwise a relationship such as (2.5) implies that transport occurs even if the substrate is inerodible bedrock. In fact, we must modify (2.5) so that the bedload transport is zero if the bedload layer thickness is equal to zero.

To be specific, we now suppose that $z = s$ describes the interface between stationary bed and moving bedload, and we suppose that the moving bedload layer has thickness a. If the (constant) porosity of the bed (both mobile and immobile) is ϕ and the bedload transport is $\mathbf{q}_b$, then conservation of mobile sediment implies that

$$\rho_s(1-\phi)a_t + \boldsymbol{\nabla}.\mathbf{q}_b = \rho_s(1-\phi)v_A, \tag{2.7}$$

where v_A is the abrasion or entrainment rate of the immobile bed, measured as a velocity.

The Exner equation which describes land surface evolution can now be written in the form

$$\rho_s(1-\phi)s_t = -\rho_s(1-\phi)v_A \ + \ \rho_s(1-\phi)U, \tag{2.8}$$

where U is the velocity of tectonic uplift, or more generally, baselevel fall. The geometric relation between the various depths is seen to be

$$\eta = s + a + h. \tag{2.9}$$

Equations (2.1), (2.7), (2.8), and (2.9) provide five equations for the five variables η, s, a, h, and $\mathbf{u}$; the abrasion rate v_A and bedload transport $\mathbf{q}_b$ need to be prescribed in constitutive relations.

Abrasion and transport rates. It is a fact of observation that the thickness a of the moving bedload layer in a stream is commonly quite small, perhaps only one or two grain thicknesses (Slingerland, Harbaugh, and Furlong (1994, pp. 80–81)). If the stream flow is very rapid, we might expect the consequently rapidly moving grains to mobilize the grains below them. These considerations suggest that the abrasion rate v_A should be a (nonnegative) decreasing function of a which tends to zero at large a and that it should depend on stream flow. With little to guide us, we make the simplest assumption that $v_A = 0$ for a larger than some constant threshold a_0, although it is not difficult to modify this assumption. When $a \geq a_0$, we have conditions of transport limitation, and when $a < a_0$, we have detachment limitation.

We define a bedload velocity (when $a = a_0$)

$$v_b = \frac{q_b}{\rho_s(1-\phi)a_0}, \tag{2.10}$$

and v_b is a function of τ_e. For example, the Meyer-Peter–Müller law (2.5) gives

$$v_b = \left(\frac{K}{\rho_w^{1/2}\Delta\rho\, g(1-\phi)a_0}\right)(\tau_e - \tau_c)_+^{3/2}. \tag{2.11}$$

The constitutive assumptions we will then make for transport and abrasion rates are

$$\begin{aligned} \mathbf{q}_b &= \rho_s(1-\phi)a v_b\,(\tau_e)\,\mathbf{N}, \\ v_A &= k v_b\,(\tau_e)\left[1 - \frac{a}{a_0}\right]_+; \end{aligned} \tag{2.12}$$

the dimensionless constant k would be expected to be extremely small. The direction of bedload transport is given by the unit vector

$$\mathbf{N} = \frac{\boldsymbol{\tau}_e}{\tau_e}. \tag{2.13}$$

Equations (2.7) and (2.8), for mobile and immobile bed surface, respectively, can now be written in the form

$$\begin{aligned} a_t + \boldsymbol{\nabla}.[av_b\mathbf{N}] &= v_A, \\ s_t &= -v_A + U. \end{aligned} \tag{2.14}$$

Nondimensionalization. We choose scales for the variables h, $\mathbf{u}$, η, s, a, τ_e, as well as $\mathbf{x}$ and t, by balancing suitable terms in the governing equations. Suppose that d is a suitable hillslope height scale and l is a suitable horizontal length scale; then we choose

$$\begin{aligned} &r \sim r_D, \quad U \sim U_D, \quad v_b \sim v_D, \quad v_A \sim U_D, \\ &\eta,\ s \sim d, \quad \mathbf{x} \sim l, \quad t \sim [t] = \frac{d}{U_D}, \quad \tau_e \sim [\tau] = f\rho_w\,[u]^2, \\ &\mathbf{u} \sim [u] = \left(\frac{gr_D d}{f}\right)^{1/3}, \quad a \sim a_0, \quad h \sim [h] = l\left(\frac{fr_D^2}{gd}\right)^{1/3}, \end{aligned} \tag{2.15}$$

where square-bracketed terms indicate scales, r_D and U_D are typical precipitation and uplift rates, and for the Meyer-Peter–Müller law (2.11) we would define

$$v_D = \left(\frac{K[\tau]^{3/2}}{\rho_w^{1/2}\Delta\rho\, g(1-\phi)a_0}\right). \tag{2.16}$$

The choice of l is determined by the implied tectonic setting. The simplest conceptual idea is the continuing uplift of an island (or mountain belt), with sea level fixed at prescribed boundaries, and this determines a natural length scale l, the scale of the island. Similarly, crustal folding determines l via the folding wave length. The other length scale d is fixed by the balance of uplift rate with hillslope denudation, which requires (since $v_A \sim U_D$ and also $v_A \sim kv_D$) that

$$U_D = kv_D. \tag{2.17}$$

This determines d through the dependence of v_D on $[\tau]$ and thus $[u]$. For example, if we take v_D to be given by (2.16), then we find

$$d = \left(\frac{\Delta\rho(1-\phi)}{Kf^{1/2}\rho_w}\right)\frac{a_0U_D}{kr_D}. \tag{2.18}$$

The first bracketed term is a constant of $O(1)$, and so we see that the depth scale $d \sim \frac{a_0U_D}{kr_D}$; high mountains are (in this theory) a consequence of high uplift rate and low rainfall, which makes intuitive sense. In addition, the thickness (a_0) and abrasiveness (k) of the bedload layer are crucial in determining d. In practice, we will actually use observed estimates for d to infer suitable values for k.

Using the scaled variables in the model equations (2.1), (2.9), (2.7), and (2.8), we obtain the dimensionless set (where now all the variables refer to the dimensionless

quantities)

$$\begin{aligned} \delta\varepsilon h_t + \boldsymbol{\nabla}.(h\mathbf{u}) &= r, \\ \delta F^2[\,\delta\varepsilon\mathbf{u}_t + (\mathbf{u}.\boldsymbol{\nabla})\mathbf{u}] &= -\boldsymbol{\nabla}\eta - \frac{|\mathbf{u}|\mathbf{u}}{h}, \\ \eta &= s + \delta h + \delta\nu a, \\ \delta\nu\alpha a_t + \boldsymbol{\nabla}.[aV\mathbf{N}] &= \alpha A, \\ s_t &= -A + U, \\ \boldsymbol{\tau}_e &= |\mathbf{u}|\mathbf{u} - \beta\boldsymbol{\nabla}s, \end{aligned} \tag{2.19}$$

where the dimensionless bedload velocity V and abrasion rate A are given, from (2.11) and $(2.12)_2$, by

$$V = [\tau_e - \tau_c^*]_+^{3/2}, \quad A = [1-a]_+ V, \tag{2.20}$$

and the parameters are given by

$$F = \frac{[u]}{(g[h])^{1/2}}, \quad \varepsilon = \frac{U_D}{r_D}, \quad \delta = \frac{[h]}{d},$$

$$\nu = \frac{a_0}{[h]}, \quad \alpha = \frac{kl}{a_0} = \frac{lU_D}{a_0 v_D}, \quad \beta = \frac{\Delta\rho\, D_s}{\rho_w [h]}. \tag{2.21}$$

The dimensionless critical stress can be written in the form

$$\tau_c^* = \frac{\Delta\rho}{\rho_w}\frac{D_s}{[h]}\frac{\mu_c l}{d}, \tag{2.22}$$

which sets out simply how the size of this parameter is determined by the hillslope aspect ratio and by the ratio of water film depth to grain size. μ_c is the dimensionless critical Shields stress, defined in (2.6), and differs from τ_c^* because of the way in which we have nondimensionalized the bed stress.

Parameter estimation. Typical values of precipitation and uplift are $r_D \sim 1$ m y^{-1}, $U_D \sim 10^{-3}$ m y^{-1} (1 km per million years). There is some flexibility in the choice of length scales l and d. Let us suppose that $d \sim 10^3$ m, $l \sim 10^5$ m (i.e., one kilometer uplift over a distance of 100 km) and that $f \sim 0.1$ and $g \sim 10$ m s^{-2}. From these, we find

$$[u] \sim 0.15 \text{ m s}^{-1}, \quad [h] \sim 2.2 \text{ cm}. \tag{2.23}$$

Let us additionally suppose that $a_0 \sim D_s \sim 1$ mm, $\Delta\rho/\rho_w = 2$. It then follows that

$$\begin{gathered} F^2 \sim 0.1, \quad \varepsilon \sim 10^{-3}, \quad \delta \sim 10^{-5}, \\ \alpha \sim 0.1, \quad \beta \sim 0.1, \quad \nu \sim 0.05, \quad \tau_c^* \sim 0.5. \end{gathered} \tag{2.24}$$

It should be emphasized that there is some flexibility in the values of these parameters, but they are all less than one, and in particular ε and δ are very small. It is then legitimate to neglect all the terms proportional to δ in the model. We shall find later that this is a singular approximation, and in order to regularize it we will need at least some of the δ terms to be retained. Apparently, the largest such term is δh in the definition of η, and we therefore choose to retain this term only. It will be easy

to check a posteriori that the neglected terms indeed remain small when the δh term becomes significant.

With the neglect of the terms in δ excluding this excepted term, we derive the reduced model

$$\begin{aligned} \boldsymbol{\nabla}.(h\mathbf{u}) &= r, \\ \mathbf{0} &= -\boldsymbol{\nabla}\eta - \frac{|\mathbf{u}|\mathbf{u}}{h}, \\ \eta &= s + \delta h, \\ \boldsymbol{\nabla}.[aV\mathbf{N}] &= \alpha A, \\ s_t &= -A + U, \\ \boldsymbol{\tau}_e &= |\mathbf{u}|\mathbf{u} - \beta\boldsymbol{\nabla}s. \end{aligned} \tag{2.25}$$

The downslope normal $\mathbf{N}$ is still defined by (2.13).

In order to prescribe boundary conditions for (2.25), consider the uplift of an island continent D with a boundary ∂D; the natural conditions to apply are then

$$\eta = 0 \quad \text{and} \quad \frac{\partial \eta}{\partial n} = 0 \quad \text{on} \quad \partial D. \tag{2.26}$$

These represent the idea that the water surface gradient becomes equal to the ocean gradient (zero) at the coastline. Because the equation for η is essentially elliptic (see the first two equations in (2.25)), the extra condition in (2.26) locates the precise position of the shoreline. Because δ is small, the shoreline position ∂D is essentially known. It will be seen that these conditions are sufficient, together with an initial condition for s, to determine the solution.

A comment on bedload transport. For a given water flow and depth, and thus constant V, the solution for a is $a = 1-\exp(-\alpha x)$, where x is the direction of flow, and assuming that $a = 0$ initially. Thus when $\alpha \gg 1$, we have conditions of transport limitation, and when $\alpha \ll 1$, the transport is detachment limited. The parameter α is the ratio of two small numbers (see (2.21)): k, the ratio of abrasion velocity to bedload velocity (see (2.12)), and a_0/l, the ratio of bedload layer thickness to regional length scale. Its size therefore depends critically on our assumptions about abrasion and bedload. It is plausible that $\alpha \ll 1$ is the more appropriate condition in a regional context over long geological time scales, as suggested by Howard (1994), but this will depend on the friability of the underlying rock. In the laboratory, however, α can be much larger than one because the abrasion coefficient k is likely to be close to one for noncohesive sediments. Simply, noncohesive sediment is eroded and removed rapidly in the field, and over longer time scales, detachment limitation is more appropriate.

A comment on time scales. Although the model and the associated parameter values derived above are consistent with observation, it is unrealistic in the sense that, for example, rainfall is not continuous, and there is no continual overland flow. Rather, erosion actually occurs during severe storms and is virtually absent between them. In a sense, time is not a continuous variable, and it may be more appropriate to switch on the erosional part of the model only during storms. The consequence of this would be a much higher value of r_D, with consequent changes in the parameter values. Despite this, it is still robustly the case that $\delta \ll 1$, and so it seems that the model may be suitable in any case; this, at least, is our assumption.

3. Linear stability. In this section, we review the stability results of Smith and Bretherton (1972) and Loewenherz-Lawrence (1994). We define the downstream flow direction by

$$\mathbf{n} = -\frac{\boldsymbol{\nabla}\eta}{|\boldsymbol{\nabla}\eta|}, \tag{3.1}$$

the stream slope as

$$S = |\boldsymbol{\nabla}\eta|, \tag{3.2}$$

and the water flux as

$$q = h|\mathbf{u}|. \tag{3.3}$$

From (2.25), we then have

$$\boldsymbol{\nabla}.[q\mathbf{n}] = r,$$

$$q = h^{3/2}S^{1/2}, \tag{3.4}$$

and the effective stress is

$$\boldsymbol{\tau}_e = -(h+\beta)\boldsymbol{\nabla}\eta + \delta\beta\boldsymbol{\nabla}h. \tag{3.5}$$

In order to relate our model to those of previous authors, we begin by making corresponding assumptions about bed abrasion and transport. In essence, the prescription of the abrasion rate A in (2.25) is replaced by an assumption that the bedload layer thickness a is constant, $a = 1$. In this case, A is determined by the model, and the bed evolution equation is

$$s_t = U - \frac{1}{\alpha}\boldsymbol{\nabla}.[V\mathbf{N}]. \tag{3.6}$$

This form of the equation is in fact what is obtained in transport limiting conditions when A is prescribed and $\alpha \gg 1$. If A is not prescribed, then the constant k is undefined, so that (2.17) cannot be used to define d. Instead, we define d by choosing $\alpha = 1$, which leads (via (2.21)) to

$$d = \left(\frac{\Delta\rho(1-\phi)}{Kf^{1/2}\rho_w}\right)\frac{lU_D}{r_D}, \tag{3.7}$$

which can be compared with (2.18). For the time being, we assume this to be the case.

Now let us consider the evolution of (one side of) a unidirectional hillslope as shown in Figure 3; that is, we suppose the equations (2.25) are to be solved in the domain $0 < x < 1$, $-L < y < L$, where x is the downslope direction. Suitable boundary conditions are for there to be zero normal flux of sediment and water at the ridge and the two sides, and $\eta = 0$ at $x = 1$. (The extra condition $\eta_x = 0$ at the shoreline is used to locate its precise position near $x = 1$.)

If we take r and U to be constant (more generally, they could be functions of x), then there is a steady state solution for hillslope and water flux; we denote the

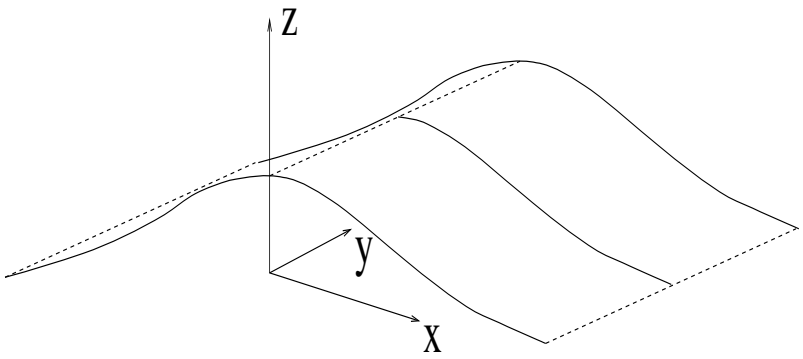


FIG. 3. *One-dimensional hillslope geometry.*

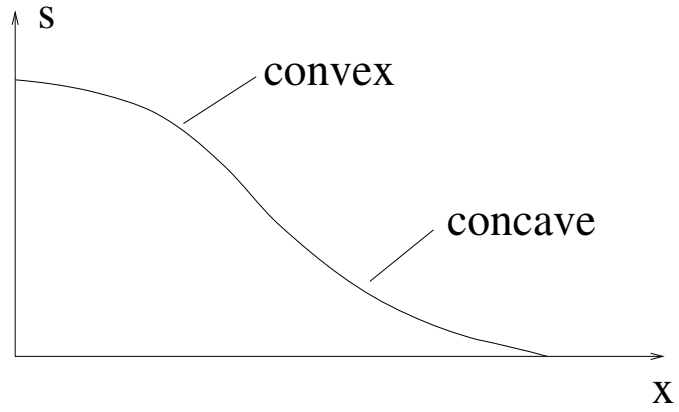


FIG. 4. *Convexity and concavity.*

steady hillslope profile by $\eta = \eta_0(x)$. Smith and Bretherton (1972) showed that for this steady state

$$x\frac{\partial V}{\partial S}S' = V - q\frac{\partial V}{\partial q}, \tag{3.8}$$

where the bedload transport function V is taken to be a function of q and S. (This can be done only if the term in δ is ignored.) Somewhat confusingly, geomorphologists term a slope with $S' < 0$ concave (see Figure 4) or, better, concave upwards, and we shall follow this practice.

As we expect, $\partial V/\partial S > 0$, and this implies that a slope is geomorphologically concave if $\partial V/\partial q > V/q$, and in particular for mathematically convex functions V. We shall find that geomorphologically concave slopes are unstable to channel formation. To leading order in δ, (3.4) and (3.5) imply

$$\tau_e = (qS)^{2/3} + \beta S, \tag{3.9}$$

and so the dimensionless Meyer-Peter–Müller relationship in (2.20), for example, can be written in the form

$$V = [(qS)^{2/3} + \beta S - \tau_c^*]_+^{3/2}. \tag{3.10}$$

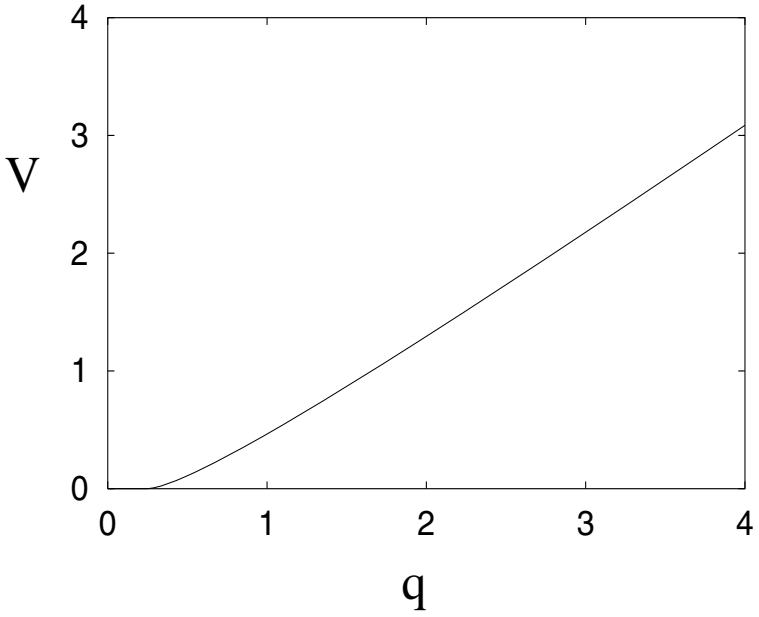


FIG. 5. $V(q,S)$ *given by* (3.10) *for* $\beta = 0.1$, $\tau_c^* = 0.5$.

Figure 5 shows that this relation typically produces a (weakly) mathematically convex function and hence a weakly concave upward hillslope.

Our aim is study perturbations to the steady state $\eta = \eta_0(x)$. Even if the water depth perturbations are large, we can still linearize the geometry of the directions $\mathbf{n}$ and $\mathbf{N}$ by expanding in terms of δ. We do this first. In the one-dimensional steady state, $\mathbf{N} = \mathbf{n} = \mathbf{i}$. We put

$$\eta = \eta_0 + \tilde{\eta}, \tag{3.11}$$

and suppose that $\tilde{\eta}$ is small. We then find

$$\begin{aligned} \boldsymbol{\nabla}\eta &= \eta_0'\mathbf{i} + \boldsymbol{\nabla}\tilde{\eta}, \\ |\boldsymbol{\nabla}\eta| = S &= S_0 - \tilde{\eta}_x + \cdots, \end{aligned} \tag{3.12}$$

where the steady state slope is

$$S_0 = |\eta_0'|. \tag{3.13}$$

Thus

$$\begin{aligned} \mathbf{n} &= \mathbf{i} - \frac{\tilde{\eta}_y}{S_0}\mathbf{j} + \cdots, \\ q &= h^{3/2}S^{1/2}, \end{aligned} \tag{3.14}$$

and in a similar way we find (if also δh is small)

$$\begin{aligned} \tau_e &= (h+\beta)S + \delta\beta h_x + \cdots, \\ \mathbf{N} &= \mathbf{i} - \frac{1}{S_0}\left\{\tilde{\eta}_y - \frac{\delta\beta}{h+\beta}h_y\right\}\mathbf{j} + \cdots. \end{aligned} \tag{3.15}$$

Adopting for the moment only these approximations (that is, we linearize the geometry only), we derive from (3.4) and (3.6) the following approximate model:

$$\frac{\partial q}{\partial x} - \frac{\partial}{\partial y}\left[\frac{q}{S_0}\frac{\partial \tilde{\eta}}{\partial y}\right] = r,$$

$$\frac{\partial \tilde{\eta}}{\partial t} - \delta\frac{\partial h}{\partial t} = U - \frac{\partial V}{\partial x} + \frac{\partial}{\partial y}\left[\frac{V}{S_0}\left\{\frac{\partial \tilde{\eta}}{\partial y} - \frac{\beta\delta}{h+\beta}\frac{\partial h}{\partial y}\right\}\right], \tag{3.16}$$

with q and τ_e defined in (3.14) and (3.15). Notice that this model is still nonlinear.

If the steady solution in which $q_0 = rx$ and $V_0 = Ux$ of this pair of equations is linearized, then what we find is the following. If we put $\delta = 0$ (and thus $V = V(q, S)$), instability occurs if $\partial V/\partial q > V/q$ at any point, as stated above, and the growth rate is unbounded ($\propto k^2$) as the lateral wave number k of modes $\propto e^{iky}$ increases. This implies ill-posedness of the model with $\delta = 0$. If $\delta > 0$ but is small, then the system is stabilized at high wave number. More detailed consideration of the linear eigenvalue problem suggests that instability occurs for k in the range $O\left(\frac{1}{\delta^{1/2}}\right) < k < O\left(\frac{1}{\delta}\right)$, and that maximal growth occurs for $k = O\left(\frac{1}{\delta^{3/4}}\right)$. Oscillations in the x direction are stabilizing.

In dimensional terms, the range of unstable wavelengths l_u is thus in the range

$$\frac{[h]l}{d} < l_u < \frac{[h]^{1/2}l}{d^{1/2}}, \tag{3.17}$$

and thus it bears no simple relation to any of the three geometric length scales of the problem but involves them all.

Because $\delta \ll 1$, i.e., $[h] \ll l$, the result in (3.17) suggests that a nonlinear theory for channel formation can be based on the fact that the lateral length scale for growing perturbations is much smaller than the downstream length scale; in other words, we now turn to a direct asymptotic solution of (3.16) when h is large.

4. An evolution equation for channel formation. The discussion above of linear stability when $\delta \ll 1$ suggests that a distinguished lateral length scale of order $\lesssim \delta^{1/2}$ may serve to delineate the unstable growth of rills. Let us now focus on this growth by defining

$$y = \delta^{1/2}Y, \quad \tilde{\eta} = \delta Z, \quad t = \delta\tilde{t}; \tag{4.1}$$

the rescaling of $\tilde{\eta}$ and t is motivated by the linear stability result of Loewenherz-Lawrence (1994), which suggests that when $y \sim 1/k \ll 1$, then $\tilde{\eta} \sim \tilde{q}/k^2$, or more generally $\tilde{\eta} \sim h^{3/2}/k^2$, and $t \sim 1/k^2$. For $k \sim 1/\delta^{1/2}$ and $h \sim O(1)$, we obtain (4.1). Note that if the original time scale $\sim d/U_D$ was 10^6 years, then this new time scale is $[h]/U_D$ (film thickness divided by uplift or erosion rate), of order 10 years.

The equations (3.16) retain their validity based on geometric linearity, and take the form

$$\frac{\partial q}{\partial x} - \frac{\partial}{\partial Y}\left[\frac{q}{S}\frac{\partial Z}{\partial Y}\right] = r,$$

$$\frac{\partial Z}{\partial \tilde{t}} - \frac{\partial h}{\partial \tilde{t}} = U - \frac{\partial V}{\partial x} + \frac{\partial}{\partial Y}\left[\frac{V}{S}\left\{\frac{\partial Z}{\partial Y} - \frac{\beta}{h+\beta}\frac{\partial h}{\partial Y}\right\}\right], \tag{4.2}$$

in which $S(x)$ is the steady slope (i.e., such that $Z = 0$ is a solution of (4.2)), and the water flux q and effective driving stress for sediment transport τ_e are given by

$$\tau_e \approx (h+\beta)S, \quad q = h^{3/2}S^{1/2}. \tag{4.3}$$

To be specific, we pose these equations on a rectangular domain $-L < y < L$ (thus $-L/\delta^{1/2} < Y < L/\delta^{1/2}$) and $0 < x < 1$. In terms of x and y, the no flux and shoreline boundary conditions require

$$\begin{aligned} \frac{\partial h}{\partial y} = \frac{\partial Z}{\partial y} = 0 \quad &\text{on} \quad y = \pm L, \\ q = V = 0 \quad &\text{on} \quad x = 0, \\ Z = 0 \quad &\text{on} \quad x = 1. \end{aligned} \tag{4.4}$$

These equations enclose the linear instability of the steady state (on a lateral space scale $Y = O(1)$, and time scale $\tilde{t} = O(1)$); but they are fully nonlinear equations and may provide a vehicle to understand the nonlinear development of the linear rill instability we have found before.

One possibility is that stable finite amplitude solutions (rills) exist for this model, with $h \sim O(1)$. Such rills have depths of order millimeters or centimeters, and do not correspond to larger river channels, which presumably evolve over longer geological time scales, possibly by coarsening and scale evolution.

We make the supposition that larger channels can evolve in this model, and therefore we seek solutions representing such large channels in which the depth $h \gg 1$, and where it is a function of the short length scale $Y \sim O(1)$. Note that a consequence of $(4.2)_1$ is that

$$\int_{-L/\delta^{1/2}}^{L/\delta^{1/2}} q \, dY = 2Lrx/\delta^{1/2}, \tag{4.5}$$

which serves as a constraint on the channel depth. In particular, (4.3) suggests a distinguished limit $h \sim 1/\delta^{1/3}$ when most of the rainfall finds its way into the channel. Thus we rescale the variables as

$$h = \frac{H}{\delta^{1/3}}, \quad q = \frac{Q}{\delta^{1/2}}, \quad V = \frac{F}{\delta^{1/2}}, \quad \tau_e = \frac{T_e}{\delta^{1/3}}, \quad \tilde{t} = \delta^{1/6} T. \tag{4.6}$$

(This assumes that $V \sim \tau_e^{3/2}$ for large τ_e, as is the case for the Meyer-Peter relation in (2.20).) With $\delta \approx 10^{-5}$, then $1/\delta^{1/3} \approx 46$, and the new depth scale is of the order of a meter, sensible for a developed stream. The choice of time scale (corresponding dimensionally to a year) is so that the time derivative of h in $(4.2)_2$ is balanced. On the other hand, we expect the water surface to remain flat, so that we do not seek to rescale Z: as we will see, this is consistent with the model equations.

Introducing (4.6) into (4.2) and (4.3), we obtain

$$\begin{aligned} \frac{\partial Q}{\partial x} - \frac{\partial}{\partial Y}\left[\frac{Q}{S}\frac{\partial Z}{\partial Y}\right] &= \delta^{1/2} r, \\ \delta^{1/2}\frac{\partial Z}{\partial T} - \frac{\partial H}{\partial T} &= \delta^{1/2} U - \frac{\partial F}{\partial x} + \frac{\partial}{\partial Y}\left[\frac{F}{S}\left\{\frac{\partial Z}{\partial Y} - \frac{\beta}{H + \delta^{1/3}\beta}\frac{\partial H}{\partial Y}\right\}\right], \end{aligned} \tag{4.7}$$

$$T_e \approx (H + \delta^{1/3}\beta)S, \quad Q = H^{3/2} S^{1/2}. \tag{4.8}$$

The rescaled sediment transport function F is only $O(1)$ with this rescaling if $F \sim \tau_e^{3/2}$, which is of course precisely true for the Meyer-Peter–Müller law:

$$F = \left[T_e - \delta^{1/3}\tau_c^*\right]_+^{3/2}. \tag{4.9}$$

Any other choice of transport law would require a more contorted rescaling.

We can use (4.8) to write (4.9) in the form

$$F = QS + \tfrac{3}{2}(\delta QS)^{1/3}(\beta S - \tau_c^*) + \cdots . \tag{4.10}$$

Simplification of $(4.7)_2$ now yields

$$-\delta^{1/2}\frac{\partial Z}{\partial T} + \frac{\partial H}{\partial T} = S'S^{1/2}H^{3/2} + S^{1/2}\frac{\partial}{\partial Y}\left[\beta H^{1/2}\frac{\partial H}{\partial Y}\right] + C\frac{\partial^2 Z}{\partial Y^2}, \tag{4.11}$$

with inessential error terms of $O(\delta^{1/3})$. The instability parameter C is given by

$$C = \frac{Q}{S}\left(F_Q - \frac{F}{Q}\right) \approx -\delta^{1/3}(\beta S - \tau_c^*)\left(\frac{H}{S}\right)^{1/2}. \tag{4.12}$$

It is a peculiarity of the Meyer-Peter–Müller law that $C = 0$ to leading order, so that the steady state is approximately neutrally linearly stable (at these large stresses). This is because at leading order F is linear in Q, and the function is neither mathematically convex nor concave

Equation (4.11) reveals the essence of linear instability and its nonlinear development. Linear instability is associated with the negative diffusion coefficient of Z if $C > 0$, i.e.,

$$S < S_c = \frac{\tau_c^*}{\beta} = \frac{\mu_c l}{d}, \tag{4.13}$$

using (2.21) and (2.22). In dimensional terms, this suggests instability if the slope is less than μ_c, which occurs at the shoreline. If the resulting rills are able to grow to significant depth, then the nonlinear evolution of H is described approximately by

$$\frac{\partial H}{\partial T} = S'S^{1/2}H^{3/2} + S^{1/2}\frac{\partial}{\partial Y}\left[\beta H^{1/2}\frac{\partial H}{\partial Y}\right], \tag{4.14}$$

and Z then follows from (4.7) by quadrature. Equation (4.14) is a degenerate nonlinear diffusion equation, about which a good deal is known. The source term is suggestive (if $S' > 0$, i.e., on the (upper) convex portion of the hillslope) of blow up and the possibility that H could reach ∞ at a finite time. The degenerate diffusion coefficient is suggestive of solutions of compact support.

The integral constraint (4.5) can be written in the limiting form (as $\delta \to 0$)

$$\int_{-\infty}^{\infty} H^{3/2}\, dY = \frac{2Lrx}{S^{1/2}}. \tag{4.15}$$

Note that this constraint is independent of (4.14), which is derived from sediment conservation, whereas (4.15) is a condition of water mass flow.

Suitable boundary conditions for (4.14) follow from matching to an outer film flow, where $Y \sim 1/\delta^{1/2}$ and $H \sim \delta^{1/3}$. Consequently, we require

$$H \to 0 \quad \text{as} \quad Y \to \pm\infty. \tag{4.16}$$

The initial condition is that H is initially small (since we suppose it arises from an instability of the steady state $H \sim \delta^{1/3}$), i.e.,

$$H \to 0 \quad \text{as} \quad T \to 0. \tag{4.17}$$

The precise behavior of H for small T is less easy to describe. The reason for this is that we have omitted an intermediate discussion of the nonlinear stability of the uniform steady state. The long time evolution of an arbitrary (infinitesimal) perturbation to the steady state can be described by consideration of a Fourier integral over normal modes of wave number k. The upshot of such a consideration is that the emerging linear solution is a monochromatic oscillation whose wave number is that with maximum growth rate, and this would serve as a suitable initial condition for the resulting nonlinear equations in (4.2). However, to obtain an appropriate initial condition for (4.14), we really need to know how solutions to (4.2) behave. We suppose that the nonlinear equations (4.2) do not (always) have stable bounded solutions for H and that (for example) they may exhibit some kind of blow up. In that case, one might expect to obtain a suitable form for the initial behavior of H by matching to the large amplitude solution of (4.2). This is similar to the procedure adopted by Stewartson and Stuart (1971).

In directly seeking solutions at larger amplitude, we are motivated by the fact that developed river channels do indeed attain depths on the order of a meter, and this is consistent with the scale of the solutions described by (4.14).

5. Solution properties. The problem (4.14) with the integral constraint, boundary, and initial conditions (4.15)–(4.17) can be written in normalized form by defining new variables u, t, η (note this is unrelated to the use of η for the water surface in sections 2 and 3) via

$$H = \left(\frac{6}{\beta}\right)^{1/3} (Lrx)^{2/3} u, \quad T = \left(\frac{\beta}{6}\right)^{1/6} \frac{S^{1/2} S'}{(Lrx)^{1/3}} t, \quad Y = \left(\frac{2\beta}{3S'}\right)^{1/2} \eta, \tag{5.1}$$

whence we find

$$u_t = u^{3/2} + \left(u^{3/2}\right)_{\eta\eta},$$

$$\int_{-\infty}^{\infty} u^{3/2}\, d\eta = 1, \quad u \to 0 \quad \text{as} \quad \eta \to \pm\infty, \quad t \to 0. \tag{5.2}$$

This equation has been much studied by pure mathematicians, and it features prominently in the book by Samarskii et al. (1995), where numerous results concerning blow up and localization (i.e., attainment of compact support) are proved. The results in this book are, however, concerned with smooth solutions, for which blow up is essentially obvious; that is, for solutions of compact support, it is assumed that the derivative of u is zero at the boundary of the support. The derivation of the same equation here from a real physical model is clearly of some interest, but it is clearly incorrect to suppose that solutions will necessarily have zero derivative at the support margin. In general, the derivatives are finite at the margins, and in fact blow up does not occur (which, physically, is an appropriate behavior).

In our investigation of the solutions of (5.2), we are led to assert the following. A solution of the problem exists, and there is a unique steady state which is globally stable and of compact support. Starting from an initial condition of infinite support, the solution attains finite support immediately (i.e., for all $t > 0$). We have not proved these results, but we show why we think they are true in the following subsections.

Steady state and linear stability. We will limit our attention to symmetric solutions, so that u is even, and $u_\eta = 0$ on $\eta = 0$. It is convenient to define

$$v = u^{3/2}, \tag{5.3}$$

and we note that for symmetric solutions, we have

$$\int_0^\infty v\,d\eta = \tfrac{1}{2}. \tag{5.4}$$

It is trivial to see that there is a unique steady state $v_s(\eta)$, given by

$$v = \tfrac{1}{2}\cos\eta, \quad 0 < \eta < \pi/2,$$
$$v = 0, \quad \eta > \pi/2. \tag{5.5}$$

To examine linear stability, we put

$$v = \tfrac{1}{2}\cos\eta + V, \tag{5.6}$$

and linearize the equations, to obtain

$$\frac{2}{3v_s^{1/3}} V_t = V_{\eta\eta} + V, \tag{5.7}$$

subject to

$$\int_0^{\pi/2} V\,d\eta = 0, \quad V_\eta = 0 \quad \text{at} \quad \eta = 0. \tag{5.8}$$

(The condition on v at the margin determines the motion of the margin.) Separable solutions to this of the form $V = W(\eta)e^{\sigma t}$ exist, and W then satisfies a nonstandard eigenvalue problem. It is convenient to define

$$\phi = W + W_\eta|_{\pi/2}\cos\eta; \tag{5.9}$$

it follows that ϕ satisfies the nonstandard eigenvalue problem

$$\phi'' + \phi = \frac{2\sigma}{3v_s^{1/3}}\left[\phi - v_s\int_0^{\pi/2}\phi\,d\eta\right], \tag{5.10}$$

subject to

$$\phi'(0) = \phi'(\pi/2) = 0. \tag{5.11}$$

Consider for a moment the equation

$$\psi'' + \psi = \lambda\psi, \tag{5.12}$$

subject to

$$\psi'(0) = \psi'(\pi/2) = 0. \tag{5.13}$$

This is a standard eigenvalue problem with eigenfunctions $\cos 2n\eta$ and eigenvalues $\lambda = 1 - 4n^2$, $n \in \mathbf{N}$, and direct integration shows that

$$\lambda = \frac{\displaystyle\int_0^{\pi/2}(\psi^2 - \psi'^2)\,d\eta}{\displaystyle\int_0^{\pi/2}\psi^2\,d\eta}. \tag{5.14}$$

The standard variational formulation for Sturm–Liouville problems then implies that the functional $\lambda(\psi)$ defined by (5.14) is maximized by the principal eigenfunction $\cos 2\eta$, for which $\lambda = -3$. It follows from this that for all functions ϕ satisfying (5.10) and (5.11) (and thus not proportional to this eigenfunction), we have

$$\int_0^{\pi/2} (\phi^2 - \phi'^2)\, d\eta < -3 \int_0^{\pi/2} \phi^2\, d\eta. \tag{5.15}$$

Multiplying (5.10) by ϕ and integrating from 0 to $\pi/2$, we thus have

$$\frac{2\sigma}{3}\left[\int_0^{\pi/2} \frac{\phi^2}{v_s^{1/3}}\, d\eta - \int_0^{\pi/2} v_s^{2/3}\phi\, d\eta \int_0^{\pi/2} \phi\, d\eta\right] = \int_0^{\pi/2} (\phi^2 - \phi'^2)\, d\eta < 0. \tag{5.16}$$

We are assuming for convenience in this exposition that σ is real. The problem (5.10) is not self-adjoint, and so σ may be complex. We leave it as an exercise to show that the proof below that $\sigma < 0$ can be straightforwardly generalized to the result $\operatorname{Re}\sigma < 0$.

From the Cauchy–Schwarz inequality, we have

$$\int_0^{\pi/2} v_s^{2/3}\phi\, d\eta \le \left(\int_0^{\pi/2} v_s^{5/3}\, d\eta\right)^{1/2} \left(\int_0^{\pi/2} \frac{\phi^2}{v_s^{1/3}}\, d\eta\right)^{1/2},$$

$$\int_0^{\pi/2} \phi\, d\eta \le \left(\int_0^{\pi/2} v_s^{1/3}\, d\eta\right)^{1/2} \left(\int_0^{\pi/2} \frac{\phi^2}{v_s^{1/3}}\, d\eta\right)^{1/2}, \tag{5.17}$$

and thus

$$\begin{aligned}\int_0^{\pi/2} v_s^{2/3}\phi\, d\eta \int_0^{\pi/2} \phi\, d\eta &\le \left(\int_0^{\pi/2} v_s^{5/3}\, d\eta \int_0^{\pi/2} v_s^{1/3}\, d\eta\right)^{1/2} \int_0^{\pi/2} \frac{\phi^2}{v_s^{1/3}}\, d\eta \\ &< \frac{\pi}{4} \int_0^{\pi/2} \frac{\phi^2}{v_s^{1/3}}\, d\eta, \end{aligned} \tag{5.18}$$

since $v_s \le \frac{1}{2}$. It follows from this and (5.16) that $\sigma < 0$. More generally, we can prove $\operatorname{Re}\sigma < 0$, so that the steady state is linearly stable as far as the discrete spectrum is concerned.

Front motion. The degeneracy of (5.2) suggests that solutions will be of compact support and that the fronts where $u = 0$ will move at finite speed. The fronts correspond to the location of the margins of the channel. Even if the initial support is unbounded, we suggest below that the solution support instantly becomes finite. It is then of interest to know how the front moves.

We write (5.2) in terms of $v = u^{3/2}$, and thus

$$\frac{2}{3v^{1/3}} v_t = v_{\eta\eta} + v, \tag{5.19}$$

and if the front position is $\eta_m(t)$ (thus $v > 0$ for $\eta < \eta_m$), we assume that near the front,

$$v \sim c(\eta_m - \eta)^\nu + d(\eta_m - \eta)^\mu + \cdots, \tag{5.20}$$

where $\mu > \nu > 0$. Substituting this into (5.19) and balancing the leading-order terms, we obtain $\nu = 3$, $\dot\eta_m = \frac{3}{2}(\nu - 1)c^{1/3}$, and thus

$$v \sim c(\eta_m - \eta)^3, \quad \dot\eta_m \sim 3c^{1/3}. \tag{5.21}$$

In terms of u, this implies

$$u \sim \alpha(\eta_m - \eta)^2, \quad \dot\eta_m \sim 3\sqrt{\alpha}, \tag{5.22}$$

and we see that such solutions are possible only for front advance. In particular, they do not describe the evolution of a channel from the initial conditions in (5.2).

Another balance is possible if $\nu = 1$, when the second-order term in (5.20) comes into play. Balancing of terms then implies $\mu = \frac{5}{3}$, and then

$$v \sim c(\eta_m - \eta) + d(\eta_m - \eta)^{5/3} + \cdots, \quad \dot\eta_m \sim \frac{5d}{3c^{2/3}}. \tag{5.23}$$

In terms of u, this yields

$$u \sim \alpha(\eta_m - \eta)^{2/3} + \beta(\eta_m - \eta)^{4/3} + \cdots, \quad \dot\eta_m \sim \frac{5\beta}{2\sqrt{\alpha}}; \tag{5.24}$$

the slope is infinite at the front, and the direction of motion depends on the coefficient of the higher-order corrective term. Fatter fronts advance, and thinner ones retreat.

Small time solution. We have mentioned above that numerical results are consistent with the idea that the solution immediately becomes of finite support. To examine how this occurs, we study the form of the solution for small t.

It is convenient for the analysis (and also for the numerical solution of the problem) to transform the domain to a fixed interval. A smart way to do this is to define the independent variable

$$s = \int_0^\eta v \, d\eta. \tag{5.25}$$

Changing variables from η, t to s, t leads to the pair of equations for v and η (which now becomes a function of s and t):

$$\begin{gathered} v\eta_s = 1, \\ \frac{2}{3v^{1/3}}[v_t - \eta_t v v_s] = v + v[v v_s]_s, \end{gathered} \tag{5.26}$$

subject to the conditions

$$\begin{aligned} \eta \;=\; & v_s = 0 \quad \text{on} \quad s = 0, \\ & v = 0 \quad \text{on} \quad s = \tfrac{1}{2}, \\ & v = v_0(s) \quad \text{at} \quad t = 0. \end{aligned} \tag{5.27}$$

The front position is then found a posteriori from the equation

$$\eta_m(t) = \eta(\tfrac{1}{2}, t). \tag{5.28}$$

If we take $v_0'(\frac{1}{2})$ to be finite, then the initial support is infinite, $\eta_m(0) = \infty$, and the solution has a singularity at $t = 0$, $s = \frac{1}{2}$. In expanding the solution for small

t, we therefore make use of the method of strained coordinates in order to ensure a uniform expansion. This will enable us to determine the initial position of the front η_m. We define new variables T, ζ via

$$t = \varepsilon T, \quad s = \zeta + \varepsilon s_1(\zeta) + \cdots, \tag{5.29}$$

in terms of which the equations become

$$\begin{aligned} v(1 - \varepsilon s_{1\zeta} \dots)\eta_\zeta &= 1, \\ v_T - \varepsilon s_{1T} v_\zeta \dots - (\eta_T - \varepsilon s_{1T}\eta_\zeta \dots)v(1 - \varepsilon s_{1\zeta} \dots)v_\zeta & \\ = \frac{3}{2}\varepsilon v^{4/3}\left[1 + (1 - \varepsilon s_{1\zeta} \dots)\frac{\partial}{\partial \zeta}\{v(1 - \varepsilon s_{1\zeta} \dots)v_\zeta\}\right]. & \end{aligned} \tag{5.30}$$

Now we seek solutions in the form

$$v \sim v_0 + \varepsilon v_1 \dots, \quad \eta \sim \eta_0 + \varepsilon \eta_1 \dots, \tag{5.31}$$

anticipating that the leading-order solution v_0 is given by the initial function $v_0(\zeta)$. The function s_1 is to be chosen in order to ensure that the expansions in (5.31) are uniformly valid.

Equating powers of ε, we find that at $O(1)$,

$$\begin{aligned} v_0\eta_{0\zeta} &= 1, \\ v_{0T} - \eta_{0T}v_0v_{0\zeta} &= 0. \end{aligned} \tag{5.32}$$

We take the solution of this to be

$$v_0 = v_0(s), \quad \eta_0 = \int_0^\zeta \frac{d\zeta'}{v_0(\zeta')}. \tag{5.33}$$

Then at $O(\varepsilon)$, we find (since $\eta_{0T} = 0$)

$$\begin{aligned} v_0\eta_{1\zeta} + \eta_{0\zeta}v_1 &= s_{1\zeta}, \\ v_{1T} - v_0v_{0\zeta}\eta_{1T} &= \tfrac{3}{2}v_0^{4/3}[1 + (v_0v_{0\zeta})_\zeta]. \end{aligned} \tag{5.34}$$

The conditions we require to be satisfied for the functions η_1, v_1, and s_1 are

$$\begin{aligned} \eta_1 = v_{1\zeta} = s_1 = 0 \quad &\text{on} \quad \zeta = 0, \\ s_1 = v_1 = 0 \quad &\text{at} \quad T = 0. \end{aligned} \tag{5.35}$$

The choice of $s_1 = 0$ ensures that $s = 0$ when $\zeta = 0$ and seems feasible because of the term $s_{1\zeta}$ in $(5.34)_1$; it is less obvious that we will be able to choose $s_1 = 0$ at $T = 0$, but if so, then $s = \zeta$ initially, which allows us to prescribe $v_1 = 0$ initially. Note that there is no boundary condition at the front, as its location in terms of ζ is not known: we do not expect to be able to prescribe $s_1 = 0$ at $\zeta = \frac{1}{2}$.

The solution can be found by eliminating v_1 in (5.34), and we find

$$\begin{aligned} \eta_1 &= \frac{s_1}{v_0} - \frac{3TI(\zeta)}{2v_0}, \\ v_1 &= v_0s_{1\zeta} - v_0^2\eta_{1\zeta}, \end{aligned} \tag{5.36}$$

taking into account the boundary and initial conditions. The function $I(\zeta)$ is defined by

$$I(\zeta) = \int_0^\zeta v_0^{1/3}[1 + (v_0 v_{0\zeta})_\zeta]\, d\zeta'. \tag{5.37}$$

We compute $v_{1\zeta}$ at $\zeta = 0$ and find

$$v_{1\zeta}|_{\zeta=0} = s_1|_{\zeta=0} + \tfrac{3}{2} T v_0^{7/3} v_0'''. \tag{5.38}$$

Because of our assumption of a symmetric solution, v_0 is even, and therefore $v_0'''(0) = 0$. It is because of this that we can consistently choose $s_1 = 0$ at $\zeta = 0$.

Finally, we must specify s_1. This is done by examining the behavior of the solution as $\zeta \to \frac{1}{2}$. We define

$$a = -v_0'(\tfrac{1}{2}). \tag{5.39}$$

Then as $\zeta \to \frac{1}{2}$,

$$v_0 \sim aX, \quad \eta_0 \sim -\frac{1}{a} \ln X + O(1), \tag{5.40}$$

where we write $X = \frac{1}{2} - \zeta$. We thus require s_1 to be such that $v_1 \le O(X)$ and $\eta_1 \le O(\ln X)$. Expanding v_1 and η_1 for small X, we find

$$\begin{aligned} \eta_1 &\sim \frac{s_1}{aX} - \frac{3 I_m T}{2aX} + O(1), \\ v_1 &\sim -a s_1 + \frac{3 I_m a T}{2} \ldots, \end{aligned} \tag{5.41}$$

where

$$I_m = I(\tfrac{1}{2}) = \int_0^{1/2} v_0^{1/3}[1 + (v_0 v_{0\zeta})_\zeta]\, d\zeta'. \tag{5.42}$$

In order to suppress the singular terms, a simple choice of s_1 which also satisfies the requested initial and boundary conditions is

$$s_1 = \tfrac{3}{2} I_m T (1 - 2X). \tag{5.43}$$

Of principal interest is the margin position, which is given implicitly by the pair of equations

$$\begin{aligned} \eta_m &= \eta_0(\zeta) + \varepsilon \eta_1(\zeta, T) + \cdots, \\ \tfrac{1}{2} &= \zeta + \varepsilon s_1(\zeta, T) + \cdots. \end{aligned} \tag{5.44}$$

Using the definitions of η_1 and s_1 and expanding for small ε, we find that the margin position is given in terms of t by the expression

$$\eta_m \approx \frac{1}{a} \ln\left\{\frac{1}{3 I_m t}\right\} + \int_0^{1/2} \left[\frac{1}{v_0(\zeta)} - \frac{1}{a(\frac{1}{2} - \zeta)}\right] d\zeta + O(t). \tag{5.45}$$

This result suggests (but does not prove) that the solution is of compact support for all $t > 0$. The asymptotic form of the front position is consistent with a numerical solution of the problem, as we now describe.

Numerical solution. To solve the system (5.26) and (5.27) numerically, we discretize the equations on uniform meshes. To avoid a sparse mesh in the neighborhood of the endpoint $\eta = \eta_m$, which would occur because of the slow change of s there, we reformulate our problem once again by defining a new positive space variable ξ as

$$(1 - \xi) = \sqrt{1 - 2s}. \tag{5.46}$$

The model can then be written in the form

$$\begin{aligned}
u_t &= \eta_t w u_\xi + w(w v_\xi)_\xi + v \quad \text{for } \xi \in (0,1), \quad t > 0,\\
v &= u^{3/2}, \qquad w = \frac{v}{1-\xi},\\
u_\xi(0,t) &= 0, \qquad u(1,t) = 0,\\
v(\xi,0) &= v_0[s(\xi)],\\
\eta_\xi &= \frac{1-\xi}{v} \quad \text{for } \xi \in (0,1], \quad \eta(0,t) = 0, \quad t \geq 0,
\end{aligned} \tag{5.47}$$

where now $u = u(\xi,t)$, $v = v(\xi,t)$, and $\eta = \eta(\xi,t)$. Furthermore, we now have

$$\eta_m(t) = \eta(1,t).$$

We discretize in time using the first-order explicit Euler method and in space using second-order finite differences on uniform meshes. Hence the time step τ is chosen much smaller than the space mesh size N^{-1}.

Approximations of η are computed at each time level by numerical integration of $(5.47)_5$ and thus will be $O(N^{-2})$-accurate. If we evaluated η_t in $(5.47)_1$ using these computed approximations of η, we would introduce huge errors of order N^{-2}/τ in the discretization of $(5.47)_1$ and fail to get accurate computed solutions. More accurate approximations of η_t are obtained by differentiating $(5.47)_5$ with respect to t, eliminating v_t from the right-hand side by $(5.47)_1$, and solving the resulting differential equation for η_t numerically. This is equivalent to replacing $(5.47)_1$ by

$$\begin{aligned}
\chi_\xi &= -\frac{3(1-\xi)}{2u}\,[w(w v_\xi)_\xi + v], \qquad \chi(0,t) = 0,\\
u_t &= \frac{\chi}{1-\xi}\,u_\xi + w(w v_\xi)_\xi + v,
\end{aligned} \tag{5.48}$$

where χ replaces $v\eta_t$. The convective term u_ξ in $(5.48)_2$ was discretized using second-order upwinding that depends on the sign of χ; for details see, e.g., Kopteva (1996).

In our computations, we used the initial condition

$$v \propto \exp\left\{-\frac{a\eta^2}{\eta+1}\right\} \quad \text{at} \quad t = 0, \tag{5.49}$$

since it follows from (5.26) that if $v_0'(s = \frac{1}{2}) = -a$, then $v \sim \exp(-a\eta)$ as $\eta \to \infty$. Figure 6 shows snapshots of the relaxation of the solution towards the steady state, while Figure 7 shows the margin evolution. We have checked the initial evolution of the margin against the asymptotic formula (5.45) and found excellent agreement. The results support the conjecture that the steady state solution is globally stable.

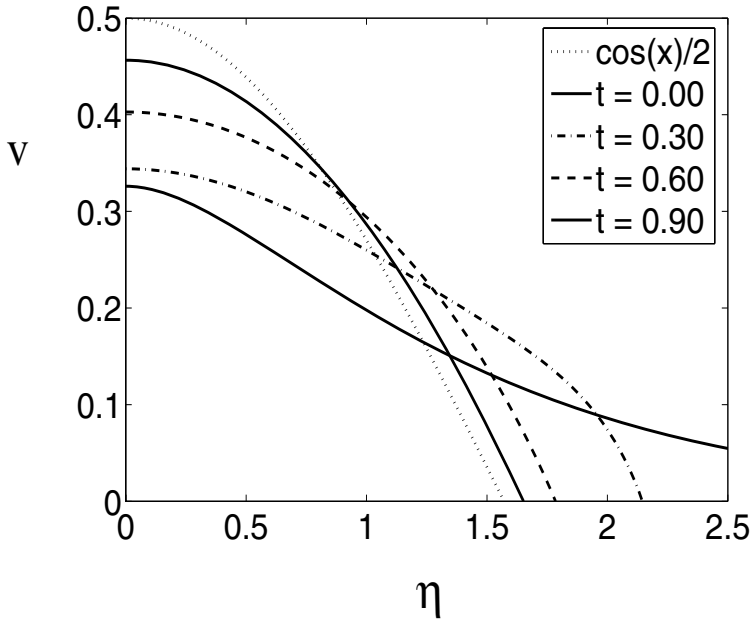


FIG. 6. *Relaxation of the solution of* (5.19) *to the steady state. The initial condition* $v_0(\eta)$ *(using the formulation in* (5.25)–(5.27)*) is given by* $v_0 \propto \exp\left\{-\frac{\eta^2}{\eta+1}\right\}$.

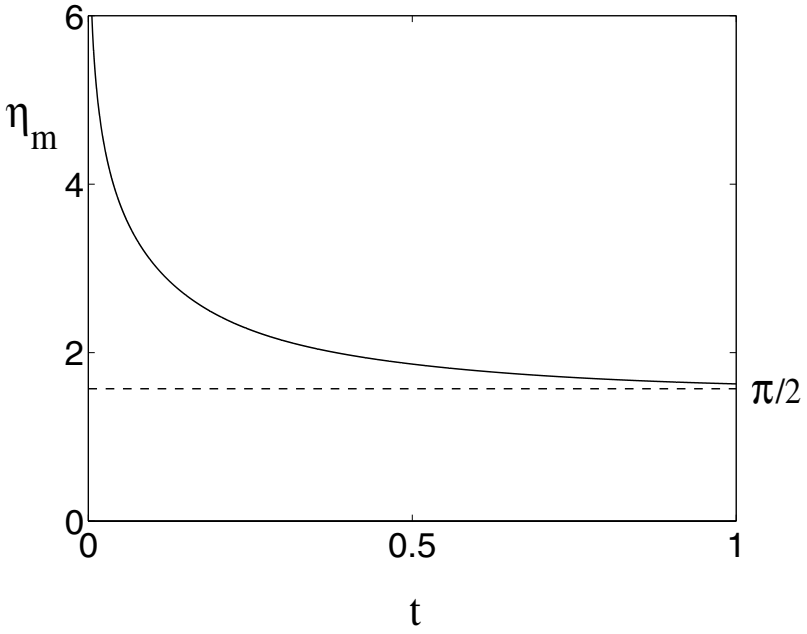


FIG. 7. *Evolution of the front position* η_m *as a function of* t *for the solution in Figure* 6. *The singularity at* $t = 0$ *is approximately (numerically) logarithmic.*

6. Conclusions. Beginning with a physics-based model of hillslope evolution and hydraulic drainage, we have shown how one can obtain a rational model for the local evolution of a stream or drainage channel. This model takes the form of a nonlinear diffusion equation with a nonlinear source term, similar to equations which have been much studied by analysts, but with the novelty of an additional integral constraint. The evidence we have gathered appears to indicate that this model is well-posed, and that its solution evolves to a unique steady state, with a width which is self-determining. This observation is interesting in view of the continuing difficulty in finding models of stream flow which can describe the stream width (see, for example, Parker (1978)).

A question of concern (but which is not addressed here) is that of putting our channel model within the context of the large scale evolution of hillslope topography. The way this can be done is as follows. As a river channel evolves, sediment is transported from the adjoining hillslope which is thus lowered. In a maturing hillslope, the channel thus eats its way down into the valley. In terms of the mathematical model, the channel will act as a thin, "shock-like" transition region between regions of hillslope with different gradients; it is a boundary layer connecting the different parts of the outer hillslope solution. Thus the results of the present paper can be used to provide a parameterization of the local channel dynamics in terms of the fluxes of sediment and water delivered from the surrounding hillslope, which evolves essentially via the Smith–Bretherton model. In this description, the hillslope evolves smoothly until it becomes concave, at which point a new channel will form. Specifically, this occurs where the characteristics of the water flow equation intersect, and the evolution of the head of the channel up the hillslope is determined by the point of shock formation. This is similar in tone but not in application to the discussion by Birnir, Smith, and Merchant (2001).

There are a number of interesting mathematical questions which deserve further study: the nonstandard eigenvalue problem (5.10) and (5.11), and the selection of front advance rate between (5.22) and (5.24), are two obvious ones. Of most concern in the application of the model to river system development is the fact that these channels grow (see (4.14)) when $S' > 0$; i.e., the hillslope is convex (upwards, in the sense of Figure 3). This is precisely the Smith–Bretherton condition which ensures that a uniform overland flow is stable. We thus have the paradoxical result that finite amplitude channels exist and are stable when the uniform steady state is also stable.

This observation is suggestive of bistability. We have not yet performed a study of the "rill" scaled model (4.2), but it is reasonable to expect it to have finite amplitude steady solutions, and these might plausibly connect to the uniform state branch at the linear stability, and "become" the channel branch as S' increases. It has to be said that it is not at all obvious how such a bifurcation diagram should be constructed. As with the other problems described above, this problem also awaits study.

Two other practical considerations deserve mention on this point. One is that our model assumes an unlimited sediment supply. In mature landscapes, erosion may become detachment limited (Howard 1994), and the form of the channel equation is somewhat changed. In essence, it appears that a similar equation may be appropriate in that case also but with a source term $H^{3/2}S^{3/2}$ which is independent of hillslope curvature.

The other comment is that in mature landscapes, such as that of Figure 1, it is evident that there will be flux of water and sediment to the channel; the hillslope is essentially three-dimensional, and it is possible that in such an altered geometry, the

conditions for channel formation are simply slope (and not curvature) dependent.

Acknowledgment. We thank Bruce Malamud for assistance with computer graphics.

REFERENCES

B. Birnir, T. R. Smith, and G. E. Merchant (2001), *The scaling of fluvial landscapes*, Comput. Geosci., 27, pp. 1189–1216.

G. F. Carrier, M. Krook, and C. E. Pearson (1966), *Functions of a Complex Variable*, McGraw–Hill, New York.

A. D. Howard (1994), *A detachment-limited model of drainage basin evolution*, Water Resour. Res., 30, pp. 2261–2285.

N. Izumi and G. Parker (1995), *Inception and channelization and drainage basin formation: Upstream-driven theory*, J. Fluid Mech., 283, pp. 341–363.

N. Izumi and G. Parker (2000), *Linear stability analysis of channel inception: Downstream-driven theory*, J. Fluid Mech., 419, pp. 239–262.

N. V. Kopteva (1996), *On the convergence, uniform with respect to a small parameter, of a four-point scheme for a one-dimensional stationary convection-diffusion equation*, Differ. Uravn., 32, pp. 951–957 (in Russian); Differential Equations, 32 (1997), pp. 958–964 (in English).

S. Kramer and M. Marder (1992), *Evolution of river networks*, Phys. Rev. Lett., 68, pp. 205–208.

D. S. Loewenherz (1991), *Stability and the initiation of channelized surface drainage: A reassessment of the short wavelength limit*, J. Geophys. Res., 96, pp. 8453–8464.

D. S. Loewenherz-Lawrence (1994), *Hydrodynamic description for advective sediment transport processes and rill initiation*, Water Resour. Res., 30, pp. 3203–3212.

E. Meyer-Peter and R. Müller (1948), *Formulas for bed-load transport*, in Proceedings of the International Association for Hydraulic Structures Research, 3rd Annual Conference, Stockholm, Sweden, pp. 39–64.

G. Parker (1978), *Self-formed straight rivers with equilibrium banks and mobile bed. Part* 1. *The sand-silt river*, J. Fluid Mech., 89, pp. 109–125.

A. A. Samarskii, V. A. Galaktionov, S. P. Kurdyumov, and A. P. Mikhailov (1995), *Blow-Up in Quasilinear Parabolic Equations*, de Gruyter Exp. Math. 19, de Gruyter, Berlin.

R. Slingerland, J. W. Harbaugh, and K. P. Furlong (1994), *Simulating Clastic Sedimentary Basins: Physical Fundamentals and Computer Programs for Creating Dynamic Systems*, Prentice–Hall, Englewood Cliffs, NJ.

T. R. Smith and F. P. Bretherton (1972), *Stability and the conservation of mass in drainage basin evolution*, Water Resour. Res., 8, pp. 1506–1529.

T. R. Smith, B. Birnir, and G. E. Merchant (1997a), *Towards an elementary theory of drainage basin evolution:* I. *The theoretical basis*, Comput. Geosci., 23, pp. 811–822.

T. R. Smith, B. Birnir, and G. E. Merchant (1997b), *Towards an elementary theory of drainage basin evolution:* II. *A computational evaluation*, Comput. Geosci., 23, pp. 823–849.

T. R. Smith, G. E. Merchant, and B. Birnir (2000), *Transient attractors: Towards a theory of the graded stream for alluvial and bedrock channels*, Comput. Geosci., 26, pp. 541–580.

K. Stewartson and J. T. Stuart (1971), *A non-linear instability theory for a wave system in plane Poiseuille flow*, J. Fluid Mech., 48, pp. 529–545.

G. E. Tucker and R. L. Slingerland (1994), *Erosional dynamics, flexural isostasy, and long-lived escarpments: A numerical modeling study*, J. Geophys. Res., 99, pp. 12229–12243.

E. W. Welsh, B. Birnir, and A. L. Bertozzi (2006), *Shocks in the evolution of an eroding channel*, Appl. Math. Res. Express, 71638.

G. Willgoose, R. L. Bras, and I. Rodríguez-Iturbe (1991), *A coupled channel network growth and hillslope evolution model:* I. *Theory*, Water Resour. Res., 27, pp. 1671–1684.

SIAM J. APPL. MATH.
Vol. 67, No. 4, pp. 1041–1063

CHANNEL FLOW OF A BINARY MIXTURE OF RIGID SPHERES DESCRIBED BY THE LINEARIZED BOLTZMANN EQUATION AND DRIVEN BY TEMPERATURE, PRESSURE, AND CONCENTRATION GRADIENTS*

R. D. M. GARCIA† AND C. E. SIEWERT‡

Abstract. An analytical version of the discrete-ordinates method (the ADO method) is used with recently established analytical expressions for the rigid-sphere scattering kernels in a study devoted to the flow of a binary gas mixture in a plane channel. In particular, concise and accurate solutions to basic flow problems in a plane channel driven by temperature, pressure, and concentration gradients and described by the linearized Boltzmann equation are established for the case of Maxwell boundary conditions for each of the two species. The velocity, heat-flow, and shear-stress profiles, as well as the mass- and heat-flow rates, are established for each species of particles, and numerical results are reported for two binary mixtures (Ne-Ar and He-Xe).

Key words. rarefied gas dynamics, binary mixtures, rigid spheres, channel flow, linearized Boltzmann equation

AMS subject classifications. 76P05, 65N35

DOI. 10.1137/060673606

1. Introduction.

While the classical problems of Poiseuille flow and thermal-creep flow in a plane channel in the general field of rarefied gas dynamics [24, 3, 5, 4] have been extensively studied for the case of a single-species gas (see, for example, [1, 25, 20, 22, 21, 14, 17] and the references therein), there are relatively few works (for example, [23, 16, 13]) devoted to these problems for gas mixtures. While [23] and [16] are based on the McCormack kinetic model [15], the work of Kosuge et al. [13] is carried out in terms of the linearized Boltzmann equation (LBE). It can be noted that the paper by Siewert and Valougeorgis [23] reports (in terms of the McCormack model) concise and accurate solutions to the problems of channel flow driven by pressure, temperature, and concentration gradients. While the approach used in [16], also based on the McCormack model, is purely numerical, that work does investigate flow in a two-dimensional channel. Most closely related to this work is [13], where purely numerical methods are used to establish some results for channel-flow problems based on the LBE.

In this work, we develop and evaluate concise and accurate solutions for flow problems in a plane-parallel channel driven by pressure, temperature, and concentration gradients. We make use of an analytical discrete-ordinates method (ADO method, [2]), and we use (in the LBE) explicit forms of the rigid-sphere collision kernels for binary gas mixtures [12, 6, 8]. The developed solutions depend (aside from some normalizations) only on the mass and diameter ratios and the relative equilibrium concentration of the two species of particles. We allow a free choice of the accommodation coefficients for each species at the confining surfaces of the channel. Our

*Received by the editors October 28, 2006; accepted for publication (in revised form) January 16, 2007; published electronically May 10, 2007.
http://www.siam.org/journals/siap/67-4/67360.html
†HSH Scientific Computing, Rua Carlos de Campos 286, São José dos Campos, SP 12242-540, Brazil (rdmgarcia@uol.com.br).
‡Mathematics Department, North Carolina State University, Raleigh, NC 27695-8205 (siewert@ncsu.edu).

approach relies on a continuous treatment of both the space and speed variables that has proved to be particularly efficient and accurate for other classical problems for binary gas mixtures [7, 9, 10].

2. Basic formulation. The flow problems considered in this work are driven by a temperature gradient, a pressure gradient, or concentration gradients (or any linear combination of these effects), and so we base our linearizations of the particle distribution functions about local rather than absolute conditions, as was done in [9], for example. We use x to measure distance in the direction (parallel to the confining walls of the plane-parallel channel) of the mentioned gradients, and so we write the local Maxwellians (for the two species of particles identified by the subscripts $\alpha = 1$ and 2) as

$$f_{\alpha,0}(x,v) = n_\alpha(x)\left[\frac{m_\alpha}{2\pi kT(x)}\right]^{3/2}\exp\left\{-\frac{m_\alpha v^2}{2kT(x)}\right\}, \quad \alpha = 1,2, \tag{2.1}$$

where v is the magnitude of the velocity $\boldsymbol{v}$. If we now express the considered linear variations in the number densities and the temperature as

$$n_\alpha(x) = n_\alpha(1 + R_\alpha x), \quad \alpha = 1,2, \tag{2.2}$$

and

$$T(x) = T_0(1 + K_T x), \tag{2.3}$$

where R_α and K_T are considered to be given (small) constants, we can linearize (2.1) to obtain the approximations

$$f^*_{\alpha,0}(x,v) = f_{\alpha,0}(v)[1 + f_\alpha(v)x], \quad \alpha = 1,2, \tag{2.4}$$

where

$$f_{\alpha,0}(v) = n_\alpha(\lambda_\alpha/\pi)^{3/2}\mathrm{e}^{-\lambda_\alpha v^2}, \quad \lambda_\alpha = m_\alpha/(2kT_0), \tag{2.5}$$

is the absolute Maxwellian distribution for n_α particles of mass m_α in equilibrium at temperature T_0. Here k is the Boltzmann constant, and the $f_\alpha(v)$ are to be determined. If we express the pressure distribution as

$$p(x) = p_0(1 + K_P x), \tag{2.6}$$

where $p_0 = nkT_0$, $n = n_1 + n_2$, and K_P is a given (small) constant, then using the perfect gas law

$$p(x) = n(x)kT(x), \tag{2.7}$$

where

$$n(x) = n_1(x) + n_2(x), \tag{2.8}$$

we find, after neglecting second-order effects,

$$c_1R_1 + c_2R_2 = K_P - K_T, \tag{2.9}$$

where $c_\alpha = n_\alpha/n$, $\alpha = 1,2$. And so, making use of (2.9), we find that we can use

$$f_1(v) = [m_1v^2/(2kT_0) - 5/2]K_T + K_P + c_2K_C \tag{2.10a}$$

and

$$f_2(v) = [m_2 v^2/(2kT_0) - 5/2]K_T + K_P - c_1 K_C, \tag{2.10b}$$

with $K_C = R_1 - R_2$, to complete (2.4). Using the variable $z \in [-z_0, z_0]$ to measure the transverse or cross-channel direction, we now write the true velocity distributions as

$$f_\alpha(x, z, \boldsymbol{v}) = f_{\alpha,0}(v)\{1 + f_\alpha(v)x + h_\alpha(z, \lambda_\alpha^{1/2}\boldsymbol{v})\}, \tag{2.11}$$

where the perturbations $h_\alpha(z, \lambda_\alpha^{1/2}\boldsymbol{v})$ are to be determined from a form of the LBE used in [12, 6, 8, 7, 10] that has an added inhomogeneous driving term due to the x variation in (2.11).

And so we proceed with an inhomogeneous form of the LBE, for a binary mixture of rigid spheres, written as

$$\boldsymbol{S}(\boldsymbol{c}) + c\mu\frac{\partial}{\partial z}\boldsymbol{H}(z, \boldsymbol{c}) + \varepsilon_0 \boldsymbol{V}(c)\boldsymbol{H}(z, \boldsymbol{c}) = \varepsilon_0 \int \mathrm{e}^{-c'^2}\boldsymbol{\mathcal{K}}(\boldsymbol{c}' : \boldsymbol{c})\boldsymbol{H}(z, \boldsymbol{c}')\mathrm{d}^3 c', \tag{2.12}$$

where ε_0 is, at this point, an arbitrary parameter that we will soon use to define a dimensionless spatial variable,

$$\boldsymbol{H}(z, \boldsymbol{c}) = \begin{bmatrix} h_1(z, \boldsymbol{c}) \\ h_2(z, \boldsymbol{c}) \end{bmatrix}, \tag{2.13}$$

and

$$\boldsymbol{S}(\boldsymbol{c}) = c(1-\mu^2)^{1/2}\cos\phi \left\{ (c^2 - 5/2)K_T \begin{bmatrix} 1 \\ 1 \end{bmatrix} + K_P \begin{bmatrix} 1 \\ 1 \end{bmatrix} + K_C \begin{bmatrix} c_2 \\ -c_1 \end{bmatrix} \right\}. \tag{2.14}$$

Considering that the driving term in (2.12) is given by (2.14), we note that (i) the case of flow driven by a temperature gradient corresponds to $K_P = 0$, $K_C = 0$, and $K_T \neq 0$, (ii) the case of flow driven by a pressure gradient corresponds to $K_T = 0$, $K_C = 0$, and $K_P \neq 0$, and (iii) the case of flow driven by concentration gradients corresponds to $K_P = 0$, $K_T = 0$, and $K_C \neq 0$. Furthermore, we note that in writing (2.12), we have introduced the variable changes

$$h_\alpha(z, \boldsymbol{c}) = h_\alpha(z, \lambda_\alpha^{1/2}\boldsymbol{v}), \quad \alpha = 1, 2, \tag{2.15}$$

in order to work with the dimensionless velocity variable $\boldsymbol{c}$. Continuing, we note that we use spherical coordinates $\{c, \theta, \phi\}$, with $\mu = \cos\theta$, to describe the dimensionless velocity vector, so that

$$\boldsymbol{H}(z, \boldsymbol{c}) \Leftrightarrow \boldsymbol{H}(z, c, \mu, \phi).$$

In our notation, $c\mu$ is the component of the (dimensionless) velocity vector in the positive z direction, and

$$c_x = c(1 - \mu^2)^{1/2}\cos\phi \tag{2.16}$$

is the component of velocity in the direction x (parallel to the confining surfaces) of the flow.

In regard to the homogeneous version of (2.12), we note that all of the defining elements have been developed in a recent series of papers [12, 6, 8]. We consider

that these works [12, 6, 8] can be consulted if a complete understanding of all of the required elements is desired. And so at this point we simply quote from our previous work [12, 6, 8] and list without additional comments the required definitions. First,

$$\boldsymbol{V}(c) = (1/\varepsilon_0)\boldsymbol{\Sigma}(c) \tag{2.17}$$

and

$$\boldsymbol{\mathcal{K}}(\boldsymbol{c}' : \boldsymbol{c}) = (1/\varepsilon_0)\boldsymbol{K}(\boldsymbol{c}' : \boldsymbol{c}), \tag{2.18}$$

where

$$\boldsymbol{\Sigma}(c) = \begin{bmatrix} \varpi_1(c) & 0 \\ 0 & \varpi_2(c) \end{bmatrix}, \tag{2.19}$$

with

$$\varpi_\alpha(c) = \varpi_\alpha^{(1)}(c) + \varpi_\alpha^{(2)}(c) \tag{2.20}$$

and

$$\varpi_\alpha^{(\beta)}(c) = 4\pi^{1/2} n_\beta \sigma_{\alpha,\beta} a_{\beta,\alpha} \nu(a_{\alpha,\beta} c). \tag{2.21}$$

Here

$$\nu(c) = \frac{2c^2 + 1}{c} \int_0^c \mathrm{e}^{-x^2} \mathrm{d}x + \mathrm{e}^{-c^2} \tag{2.22}$$

and

$$a_{\alpha,\beta} = (m_\beta/m_\alpha)^{1/2}, \quad \alpha, \beta = 1, 2. \tag{2.23}$$

We use $\sigma_{\alpha,\beta}$ to denote the differential-scattering cross section, which (for the case of rigid-sphere scattering that is isotropic in the center-of-mass system) we write as [4]

$$\sigma_{\alpha,\beta} = \frac{1}{4}\left(\frac{d_\alpha + d_\beta}{2}\right)^2, \quad \alpha, \beta = 1, 2, \tag{2.24}$$

where d_1 and d_2 are the atomic diameters of the two types of gas particles. We continue to follow [12, 6, 8] and write

$$\boldsymbol{K}(\boldsymbol{c}' : \boldsymbol{c}) = \begin{bmatrix} K_{1,1}(\boldsymbol{c}' : \boldsymbol{c}) & K_{1,2}(\boldsymbol{c}' : \boldsymbol{c}) \\ K_{2,1}(\boldsymbol{c}' : \boldsymbol{c}) & K_{2,2}(\boldsymbol{c}' : \boldsymbol{c}) \end{bmatrix}, \tag{2.25}$$

where

$$K_{1,1}(\boldsymbol{c}' : \boldsymbol{c}) = 4n_1\sigma_{1,1}\pi^{1/2}\mathcal{P}(\boldsymbol{c}' : \boldsymbol{c}) + n_2\sigma_{1,2}\pi^{1/2}\mathcal{F}_{1,2}(\boldsymbol{c}' : \boldsymbol{c}), \tag{2.26}$$

$$K_{1,2}(\boldsymbol{c}' : \boldsymbol{c}) = 4n_2\sigma_{1,2}\pi^{1/2}\mathcal{G}_{1,2}(\boldsymbol{c}' : \boldsymbol{c}), \tag{2.27}$$

$$K_{2,1}(\boldsymbol{c}' : \boldsymbol{c}) = 4n_1\sigma_{2,1}\pi^{1/2}\mathcal{G}_{2,1}(\boldsymbol{c}' : \boldsymbol{c}), \tag{2.28}$$

and

$$K_{2,2}(\boldsymbol{c}' : \boldsymbol{c}) = 4n_2\sigma_{2,2}\pi^{1/2}\mathcal{P}(\boldsymbol{c}' : \boldsymbol{c}) + n_1\sigma_{2,1}\pi^{1/2}\mathcal{F}_{2,1}(\boldsymbol{c}' : \boldsymbol{c}). \tag{2.29}$$

Here

$$\mathcal{P}(\boldsymbol{c}' : \boldsymbol{c}) = \frac{1}{\pi}\left(\frac{2}{|\boldsymbol{c}' - \boldsymbol{c}|}\exp\left\{\frac{|\boldsymbol{c}' \times \boldsymbol{c}|^2}{|\boldsymbol{c}' - \boldsymbol{c}|^2}\right\} - |\boldsymbol{c}' - \boldsymbol{c}|\right) \tag{2.30}$$

is the basic kernel for a single-species gas used by Pekeris [18]. In addition,

$$\mathcal{F}_{\alpha,\beta}(\boldsymbol{c}' : \boldsymbol{c}) = \mathcal{F}(a_{\alpha,\beta}; \boldsymbol{c}' : \boldsymbol{c}) \tag{2.31}$$

and

$$\mathcal{G}_{\alpha,\beta}(\boldsymbol{c}' : \boldsymbol{c}) = \mathcal{G}(a_{\alpha,\beta}; \boldsymbol{c}' : \boldsymbol{c}), \tag{2.32}$$

where

$$\mathcal{F}(a; \boldsymbol{c}' : \boldsymbol{c}) = \frac{(a^2+1)^2}{a^3\pi|\boldsymbol{c}' - \boldsymbol{c}|}\exp\left\{a^2\frac{|\boldsymbol{c}' \times \boldsymbol{c}|^2}{|\boldsymbol{c}' - \boldsymbol{c}|^2} - \frac{(1-a^2)^2({c'}^2 + c^2)}{4a^2} - \frac{(a^4-1)\boldsymbol{c}' \cdot \boldsymbol{c}}{2a^2}\right\} \tag{2.33}$$

and

$$\mathcal{G}(a; \boldsymbol{c}' : \boldsymbol{c}) = \frac{1}{a\pi}\big|\boldsymbol{c}' - a\boldsymbol{c}\big|[J(a; \boldsymbol{c}' : \boldsymbol{c}) - 1], \tag{2.34}$$

with

$$J(a; \boldsymbol{c}' : \boldsymbol{c}) = \frac{(a + 1/a)^2}{2\Delta(a; \boldsymbol{c}' : \boldsymbol{c})}\exp\left\{\frac{-2C(a; \boldsymbol{c}' : \boldsymbol{c})}{(a - 1/a)^2}\right\}\sinh\left\{\frac{2\Delta(a; \boldsymbol{c}' : \boldsymbol{c})}{(a - 1/a)^2}\right\}, \quad a \neq 1, \tag{2.35a}$$

or

$$J(a; \boldsymbol{c}' : \boldsymbol{c}) = \frac{1}{|\boldsymbol{c}' - \boldsymbol{c}|^2}\exp\left\{\frac{|\boldsymbol{c}' \times \boldsymbol{c}|^2}{|\boldsymbol{c}' - \boldsymbol{c}|^2}\right\}, \quad a = 1. \tag{2.35b}$$

We have used the definitions [12, 6, 8]

$$\Delta(a; \boldsymbol{c}' : \boldsymbol{c}) = \left\{C^2(a; \boldsymbol{c}' : \boldsymbol{c}) + (a - 1/a)^2|\boldsymbol{c}' \times \boldsymbol{c}|^2\right\}^{1/2} \tag{2.36}$$

and

$$C(a; \boldsymbol{c}' : \boldsymbol{c}) = {c'}^2 + c^2 - (a + 1/a)\boldsymbol{c}' \cdot \boldsymbol{c}. \tag{2.37}$$

In this work, we intend to compute the velocity, the shear-stress, and the heat-flow profiles which we express as

$$\boldsymbol{U}(z) = \frac{1}{\pi^{3/2}}\int_0^\infty\int_{-1}^1\int_0^{2\pi} \mathrm{e}^{-c^2}\boldsymbol{H}(z, \boldsymbol{c})c^3(1-\mu^2)^{1/2}\cos\phi \mathrm{d}c\mathrm{d}\phi\mathrm{d}\mu\mathrm{d}c, \tag{2.38}$$

$$\boldsymbol{P}(z) = \frac{2}{\pi^{3/2}}\int_0^\infty\int_{-1}^1\int_0^{2\pi} \mathrm{e}^{-c^2}\boldsymbol{H}(z, \boldsymbol{c})c^4\mu(1-\mu^2)^{1/2}\cos\phi \mathrm{d}c\mathrm{d}\phi\mathrm{d}\mu\mathrm{d}c, \tag{2.39}$$

and

$$\boldsymbol{Q}(z) = \frac{1}{\pi^{3/2}}\int_0^\infty\int_{-1}^1\int_0^{2\pi} \mathrm{e}^{-c^2}\boldsymbol{H}(z, \boldsymbol{c})\left(c^2 - \frac{5}{2}\right)c^3(1-\mu^2)^{1/2}\cos\phi \mathrm{d}c\mathrm{d}\phi\mathrm{d}\mu\mathrm{d}c, \tag{2.40}$$

where the components of $\boldsymbol{U}(z)$, $\boldsymbol{P}(z)$, and $\boldsymbol{Q}(z)$ are the functions $U_\alpha(z)$, $P_\alpha(z)$, and $Q_\alpha(z)$, for $\alpha = 1, 2$, that can be used, as mentioned in Appendix A of [9], to define the macroscopic quantities for a binary mixture.

As in [9], it is clear (for the specific flow problems considered here) that an expansion of $\boldsymbol{H}(z, \boldsymbol{c})$ in a Fourier series (in the angle ϕ) requires only one term—that is, one proportional to $\cos\phi$. And so we follow [8] and introduce the dimensionless spatial variable

$$\tau = z\varepsilon_0, \tag{2.41}$$

where

$$\varepsilon_0 = (n_1 + n_2)\pi^{1/2}\left(\frac{n_1 d_1 + n_2 d_2}{n_1 + n_2}\right)^2, \tag{2.42}$$

and write

$$\boldsymbol{H}(\tau/\varepsilon_0, \boldsymbol{c}) = \boldsymbol{\Psi}(\tau, c, \mu)(1 - \mu^2)^{1/2}\cos\phi, \tag{2.43}$$

where $\boldsymbol{\Psi}(\tau, c, \mu)$ is the (vector-valued) function to be determined. We now let $z = \tau/\varepsilon_0$ in (2.38)–(2.40) and consider that

$$\boldsymbol{U}(\tau) = \frac{1}{\pi^{1/2}}\int_0^\infty\int_{-1}^1 \mathrm{e}^{-c^2}\boldsymbol{\Psi}(\tau, c, \mu)c^3(1 - \mu^2)\mathrm{d}\mu\mathrm{d}c, \tag{2.44}$$

$$\boldsymbol{P}(\tau) = \frac{2}{\pi^{1/2}}\int_0^\infty\int_{-1}^1 \mathrm{e}^{-c^2}\boldsymbol{\Psi}(\tau, c, \mu)c^4(1 - \mu^2)\mu\mathrm{d}\mu\mathrm{d}c, \tag{2.45}$$

and

$$\boldsymbol{Q}(\tau) = \frac{1}{\pi^{1/2}}\int_0^\infty\int_{-1}^1 \mathrm{e}^{-c^2}\boldsymbol{\Psi}(\tau, c, \mu)\left(c^2 - \frac{5}{2}\right)c^3(1 - \mu^2)\mathrm{d}\mu\mathrm{d}c \tag{2.46}$$

are the quantities to be computed. It should be noted that to avoid excessive notation, we have, in writing (2.44)–(2.46), followed the (often-used) procedure of not always introducing new labels for dependent quantities (in this case $\boldsymbol{U}$, $\boldsymbol{P}$, and $\boldsymbol{Q}$) when the independent variable is changed.

We can now use (2.43) in (2.12), multiply the resulting equation by $\cos\phi$, integrate over all ϕ, and use the Legendre expansion of the scattering kernel $\boldsymbol{\mathcal{K}}(\boldsymbol{c}' : \boldsymbol{c})$ that was introduced in a previous work—see equations (26) and (65) of [8]—to find

$$\begin{aligned}\boldsymbol{\Upsilon}(c) + c\mu\frac{\partial}{\partial\tau}\boldsymbol{\Psi}(\tau, c, \mu) + \boldsymbol{V}(c)\boldsymbol{\Psi}(\tau, c, \mu)\\ = \int_0^\infty\int_{-1}^1 \mathrm{e}^{-c'^2} f(\mu', \mu)\boldsymbol{\mathcal{K}}(c', \mu' : c, \mu)\boldsymbol{\Psi}(\tau, c', \mu')c'^2\mathrm{d}\mu'\mathrm{d}c',\end{aligned} \tag{2.47}$$

where

$$f(\mu', \mu) = \left(\frac{1 - \mu'^2}{1 - \mu^2}\right)^{1/2}. \tag{2.48}$$

In addition,

$$\boldsymbol{\mathcal{K}}(c', \mu' : c, \mu)\cos\phi' = \int_0^{2\pi}\boldsymbol{\mathcal{K}}(\boldsymbol{c}' : \boldsymbol{c})\cos\phi\mathrm{d}\phi, \tag{2.49}$$

which we can express, in the notation of [8], as

$$\boldsymbol{\mathcal{K}}(c',\mu':c,\mu)=(1/2)\sum_{n=1}^{\infty}(2n+1)P_n^1(\mu')P_n^1(\mu)\boldsymbol{\mathcal{K}}_n(c',c), \tag{2.50}$$

where $P_n^1(x)$ is used to denote one of the normalized associated Legendre functions. More explicitly,

$$P_l^m(\mu)=\left[\frac{(l-m)!}{(l+m)!}\right]^{1/2}(1-\mu^2)^{m/2}\frac{\mathrm{d}^m}{\mathrm{d}\mu^m}P_l(\mu), \tag{2.51}$$

where $P_l(\mu)$ is the Legendre polynomial. In addition,

$$\boldsymbol{\mathcal{K}}_n(c',c)=\begin{bmatrix}\mathcal{K}_n^{(1,1)}(c',c) & \mathcal{K}_n^{(1,2)}(c',c)\\ \mathcal{K}_n^{(2,1)}(c',c) & \mathcal{K}_n^{(2,2)}(c',c)\end{bmatrix}, \tag{2.52}$$

with

$$\mathcal{K}_n^{(1,1)}(c',c)=p_1\mathcal{P}^{(n)}(c',c)+(g_2/4)\mathcal{F}^{(n)}(a_{1,2};c',c), \tag{2.53a}$$

$$\mathcal{K}_n^{(1,2)}(c',c)=g_2\mathcal{G}^{(n)}(a_{1,2};c',c), \tag{2.53b}$$

$$\mathcal{K}_n^{(2,1)}(c',c)=g_1\mathcal{G}^{(n)}(a_{2,1};c',c), \tag{2.53c}$$

and

$$\mathcal{K}_n^{(2,2)}(c',c)=p_2\mathcal{P}^{(n)}(c',c)+(g_1/4)\mathcal{F}^{(n)}(a_{2,1};c',c). \tag{2.53d}$$

We also can write

$$\boldsymbol{V}(c)=\begin{bmatrix}v_1(c) & 0\\ 0 & v_2(c)\end{bmatrix}, \tag{2.54}$$

where now

$$v_1(c)=p_1\nu(c)+g_2a_{2,1}\nu(a_{1,2}c) \tag{2.55a}$$

and

$$v_2(c)=p_2\nu(c)+g_1a_{1,2}\nu(a_{2,1}c). \tag{2.55b}$$

In writing (2.53) and (2.55), we have used

$$p_\alpha=c_\alpha\left(\frac{nd_\alpha}{n_1d_1+n_2d_2}\right)^2,\quad \alpha=1,2, \tag{2.56a}$$

and

$$g_\alpha=c_\alpha\left(\frac{nd_{\mathrm{avg}}}{n_1d_1+n_2d_2}\right)^2,\quad \alpha=1,2, \tag{2.56b}$$

where

$$d_{\mathrm{avg}}=(d_1+d_2)/2. \tag{2.57}$$

In order to avoid too much repetition, we do not list here our expressions for the Legendre moments

$$\mathcal{P}^{(n)}(c',c), \quad \mathcal{F}^{(n)}(a;c',c), \quad \text{and} \quad \mathcal{G}^{(n)}(a;c',c),$$

since they are explicitly given in Appendix A of [8]. To complete (2.47), we note that the inhomogeneous driving term is

$$\boldsymbol{\Upsilon}(c) = (c/\varepsilon_0) \left[\begin{array}{c} (c^2 - 5/2)K_T + K_P + c_2 K_C \\ (c^2 - 5/2)K_T + K_P - c_1 K_C \end{array} \right]. \tag{2.58}$$

At the walls located at $\tau = -a$ and $\tau = a$, we use a combination of specular and diffuse reflection, and so, in regard to (2.12), we write the boundary conditions as

$$\boldsymbol{H}(-a,c,\mu,\phi) - (\boldsymbol{I} - \boldsymbol{\alpha})\boldsymbol{H}(-a,c,-\mu,\phi) - \boldsymbol{\alpha}\boldsymbol{\mathcal{I}}_-\{\boldsymbol{H}\}(-a) = \boldsymbol{0} \tag{2.59a}$$

and

$$\boldsymbol{H}(a,c,-\mu,\phi) - (\boldsymbol{I} - \boldsymbol{\beta})\boldsymbol{H}(a,c,\mu,\phi) - \boldsymbol{\beta}\boldsymbol{\mathcal{I}}_+\{\boldsymbol{H}\}(a) = \boldsymbol{0} \tag{2.59b}$$

for $\mu \in (0,1]$ and all c and all ϕ. Here

$$\boldsymbol{\mathcal{I}}_\mp\{\boldsymbol{H}\}(z) = \frac{2}{\pi}\int_0^\infty \int_0^1 \int_0^{2\pi} \mathrm{e}^{-c'^2} \boldsymbol{H}(z,c',\mp\mu',\phi')\mu' c'^3 \mathrm{d}\phi' \mathrm{d}\mu' \mathrm{d}c', \tag{2.60}$$

$$\boldsymbol{\alpha} = \mathrm{diag}\{\alpha_1, \alpha_2\}, \tag{2.61a}$$

and

$$\boldsymbol{\beta} = \mathrm{diag}\{\beta_1, \beta_2\}, \tag{2.61b}$$

where α_1, α_2, β_1, and β_2 are the accommodation coefficients to be used for the two species of gas particles at the confining surfaces. Taking note of (2.43), we find from (2.59) the boundary conditions subject to which we must solve (2.47), that is,

$$\boldsymbol{\Psi}(-a,c,\mu) - (\boldsymbol{I} - \boldsymbol{\alpha})\boldsymbol{\Psi}(-a,c,-\mu) = \boldsymbol{0} \tag{2.62a}$$

and

$$\boldsymbol{\Psi}(a,c,-\mu) - (\boldsymbol{I} - \boldsymbol{\beta})\boldsymbol{\Psi}(a,c,\mu) = \boldsymbol{0}, \tag{2.62b}$$

for $\mu \in (0,1]$ and all c. We use $\boldsymbol{I}$ to denote the 2×2 identity matrix.

3. Solutions. Following our previous work as reported in [9, 11], we express our solution (evaluated at the N pairs of discrete ordinates $\pm\mu_i$) of (2.47) in the form

$$\boldsymbol{\Psi}(\tau,c,\pm\mu_i) = \boldsymbol{\Psi}_{ps}(\tau,c,\pm\mu_i) + \boldsymbol{\Psi}_{*}(\tau,c,\pm\mu_i) + \boldsymbol{\Psi}_{app}(\tau,c,\pm\mu_i) \tag{3.1}$$

for $i = 1,2,\ldots,N$. We note that $\boldsymbol{\Psi}_*(\tau,c,\mu)$ is defined in terms of two of the exact elementary solutions we reported in a previous work [8], that is,

$$\boldsymbol{\Psi}_*(\tau,c,\mu) = A_1 c\boldsymbol{\Phi} + B_1\big[c\tau\boldsymbol{\Phi} - \mu\boldsymbol{B}(c)\big], \tag{3.2}$$

where

$$\boldsymbol{\Phi} = \left[\begin{array}{c} 1 \\ a_{1,2} \end{array} \right], \tag{3.3}$$

and where $\boldsymbol{B}(c)$ is one of the generalized Chapman–Enskog (vector-valued) functions discussed in [8]. In addition,

(3.4)
$$\boldsymbol{\Psi}_{app}(\tau, c, \pm\mu_i) = \boldsymbol{\Pi}(c) \sum_{j=2}^{J} \left[A_j \boldsymbol{\Phi}(\nu_j, \pm\mu_i) \mathrm{e}^{-(a+\tau)/\nu_j} + B_j \boldsymbol{\Phi}(\nu_j, \mp\mu_i) \mathrm{e}^{-(a-\tau)/\nu_j} \right].$$

For our computations, we use the $2 \times 2(K+1)$ matrix

$$\boldsymbol{\Pi}(c) = \left[P_0(2\mathrm{e}^{-c} - 1)\boldsymbol{I} \;\; P_1(2\mathrm{e}^{-c} - 1)\boldsymbol{I} \cdots P_K(2\mathrm{e}^{-c} - 1)\boldsymbol{I} \right], \tag{3.5}$$

where $K+1$ is the number of basis functions used to represent the speed dependence of the approximate part of our solution. We note that [9] can be consulted if a complete understanding of the eigenvalue spectrum $\{\nu_j\}$ and the elementary solutions $\{\boldsymbol{\Phi}(\nu_j, \pm\mu_i)\}$ is desired. Since (2.47) has the inhomogeneous driving term $\boldsymbol{\Upsilon}(c)$, we have included in (3.1) the particular solution

$$\boldsymbol{\Psi}_{ps}(\tau, c, \mu) = \boldsymbol{\Psi}_P(\tau, c, \mu) + \boldsymbol{\Psi}_T(\tau, c, \mu) + \boldsymbol{\Psi}_C(\tau, c, \mu), \tag{3.6}$$

the elements of which were developed and reported in [11]. We repeat from [11]:

(3.7)
$$\boldsymbol{\Psi}_P(\tau, c, \mu) = [1/(\varepsilon_0 \varepsilon_p)] \{ c\tau^2 \boldsymbol{\Phi} - 2\mu\tau \boldsymbol{B}(c) + (1/5)\boldsymbol{D}(c) + [(5\mu^2 - 1)/5]\boldsymbol{E}(c) \} K_P,$$

$$\boldsymbol{\Psi}_T(\tau, c, \mu) = -(1/\varepsilon_0)[\boldsymbol{A}^{(1)}(c) + \boldsymbol{A}^{(2)}(c)] K_T, \tag{3.8}$$

and

$$\boldsymbol{\Psi}_C(\tau, c, \mu) = (1/\varepsilon_0)[c_2 \boldsymbol{A}^{(1)}(c) - c_1 \boldsymbol{A}^{(2)}(c)] K_C. \tag{3.9}$$

In [8] and [11], we have defined and computed, for selected cases, the generalized Chapman–Enskog and Burnett (vector-valued) functions $\boldsymbol{A}^{(1)}(c)$, $\boldsymbol{A}^{(2)}(c)$, $\boldsymbol{B}(c)$, $\boldsymbol{D}(c)$, and $\boldsymbol{E}(c)$ that appear in (3.7)–(3.9). In addition, the constant ε_p is expressed in [11] as

$$\varepsilon_p = \begin{bmatrix} c_1 & c_2 \end{bmatrix} \boldsymbol{\varepsilon}_p, \tag{3.10}$$

where

$$\boldsymbol{\varepsilon}_p = \frac{16}{15\pi^{1/2}} \int_0^\infty \mathrm{e}^{-c^2} \boldsymbol{B}(c) c^4 \mathrm{d}c. \tag{3.11}$$

We note that the components $\varepsilon_{p,1}$ and $\varepsilon_{p,2}$ of $\boldsymbol{\varepsilon}_p$ have been evaluated (for several data sets) in [8].

Finally, to complete our discussion of (3.1), we note that the arbitrary constants $\{A_j, B_j\}$ are to be determined from boundary conditions to be applied at $\tau = \pm a$. For this purpose, we substitute (3.1) into discrete-ordinates versions of (2.62), multiply the resulting equations by

$$c^2 \exp\{-c^2\} \boldsymbol{\Pi}^T(c),$$

where the superscript T is used to denote the transpose operation, and integrate over all c to define a system of $2J$ linear algebraic equations for the $2J$ unspecified constants. We note that only the right-hand-side vector of such system is problem-dependent.

4. Quantities of interest. Considering that we have solved the system of linear algebraic equations to establish the arbitrary constants $\{A_j, B_j\}$, we can use (3.1) to find our final expressions for the quantities of interest here, that is, the velocity, heat-flow, and shear-stress profiles. And so, using (3.1) in discrete-ordinates versions of (2.44)–(2.46), we find

(4.1a)

$$\boldsymbol{U}(\tau) = \boldsymbol{U}_{ps}(\tau) + (1/2)(A_1 + B_1\tau)\boldsymbol{\Phi} + \sum_{j=2}^{J} \left[A_j \mathrm{e}^{-(a+\tau)/\nu_j} + B_j \mathrm{e}^{-(a-\tau)/\nu_j}\right]\boldsymbol{\mathcal{U}}_j,$$

(4.1b)

$$\boldsymbol{Q}(\tau) = \boldsymbol{Q}_{ps}(\tau) + \sum_{j=2}^{J} \left[A_j \mathrm{e}^{-(a+\tau)/\nu_j} + B_j \mathrm{e}^{-(a-\tau)/\nu_j}\right]\boldsymbol{\mathcal{Q}}_j,$$

and

(4.1c)

$$\boldsymbol{P}(\tau) = \boldsymbol{P}_{ps}(\tau) - (1/2)B_1\boldsymbol{\varepsilon}_p + \sum_{j=2}^{J} \left[A_j \mathrm{e}^{-(a+\tau)/\nu_j} - B_j \mathrm{e}^{-(a-\tau)/\nu_j}\right]\boldsymbol{\mathcal{P}}_j.$$

In writing (4.1), we have used the definitions

(4.2a)

$$\boldsymbol{\mathcal{U}}_j = \boldsymbol{\Pi}_1 \boldsymbol{X}_j,$$

(4.2b)

$$\boldsymbol{\mathcal{P}}_j = 2\boldsymbol{\Pi}_2 \boldsymbol{Y}_j,$$

and

(4.2c)

$$\boldsymbol{\mathcal{Q}}_j = [\boldsymbol{\Pi}_3 - (5/2)\boldsymbol{\Pi}_1]\boldsymbol{X}_j,$$

where

(4.3a)

$$\boldsymbol{X}_j = \frac{1}{\pi^{1/2}} \sum_{k=1}^{N} w_k(1-\mu_k^2)[\boldsymbol{\Phi}(\nu_j, \mu_k) + \boldsymbol{\Phi}(\nu_j, -\mu_k)],$$

(4.3b)

$$\boldsymbol{Y}_j = \frac{1}{\pi^{1/2}} \sum_{k=1}^{N} w_k\mu_k(1-\mu_k^2)[\boldsymbol{\Phi}(\nu_j, \mu_k) - \boldsymbol{\Phi}(\nu_j, -\mu_k)],$$

and

(4.4)

$$\boldsymbol{\Pi}_n = \int_0^\infty \mathrm{e}^{-c^2} \boldsymbol{\Pi}(c) c^{n+2} \mathrm{d}c.$$

In (4.3), we use the weights $\{w_k\}$, along with the nodes $\{\mu_k\}$, to complete the definition of our N-point, half-range quadrature scheme. Finally, to complete (4.1), we must compute

(4.5)

$$\boldsymbol{U}_{ps}(\tau) = \frac{1}{\pi^{1/2}} \int_0^\infty \int_{-1}^1 \mathrm{e}^{-c^2} \boldsymbol{\Psi}_{ps}(\tau, c, \mu) c^3 (1-\mu^2) \mathrm{d}\mu \mathrm{d}c,$$

(4.6)

$$\boldsymbol{Q}_{ps}(\tau) = \frac{1}{\pi^{1/2}} \int_0^\infty \int_{-1}^1 \mathrm{e}^{-c^2} \boldsymbol{\Psi}_{ps}(\tau, c, \mu) \left(c^2 - \frac{5}{2}\right) c^3 (1-\mu^2) \mathrm{d}\mu \mathrm{d}c,$$

and

$$\boldsymbol{P}_{ps}(\tau) = \frac{2}{\pi^{1/2}} \int_0^\infty \int_{-1}^1 \mathrm{e}^{-c^2} \boldsymbol{\Psi}_{ps}(\tau, c, \mu) c^4 (1-\mu^2) \mu \mathrm{d}\mu \mathrm{d}c. \tag{4.7}$$

Using (3.6)–(3.9), we find

$$\boldsymbol{U}_{ps}(\tau) = (1/\varepsilon_0)\{(1/\varepsilon_p)[(1/2)\tau^2 \boldsymbol{\Phi} + \boldsymbol{D}_U] K_p - [\boldsymbol{A}_U^{(1)} + \boldsymbol{A}_U^{(2)}] K_T + [c_2 \boldsymbol{A}_U^{(1)} - c_1 \boldsymbol{A}_U^{(2)}] K_C\}, \tag{4.8}$$

$$\boldsymbol{Q}_{ps}(\tau) = (1/\varepsilon_0)\{(1/\varepsilon_p) \boldsymbol{D}_Q K_p - [\boldsymbol{A}_Q^{(1)} + \boldsymbol{A}_Q^{(2)}] K_T + [c_2 \boldsymbol{A}_Q^{(1)} - c_1 \boldsymbol{A}_Q^{(2)}] K_C\}, \tag{4.9}$$

and

$$\boldsymbol{P}_{ps}(\tau) = -[\tau/(\varepsilon_0 \varepsilon_p)] \boldsymbol{\varepsilon}_p K_p, \tag{4.10}$$

where

$$\boldsymbol{D}_U = \frac{4}{15\pi^{1/2}} \int_0^\infty \mathrm{e}^{-c^2} \boldsymbol{D}(c) c^3 \mathrm{d}c, \tag{4.11a}$$

$$\boldsymbol{A}_U^{(\alpha)} = \frac{4}{3\pi^{1/2}} \int_0^\infty \mathrm{e}^{-c^2} \boldsymbol{A}^{(\alpha)}(c) c^3 \mathrm{d}c, \quad \alpha = 1, 2, \tag{4.11b}$$

$$\boldsymbol{D}_Q = \frac{4}{15\pi^{1/2}} \int_0^\infty \mathrm{e}^{-c^2} \boldsymbol{D}(c) \left(c^2 - \frac{5}{2}\right) c^3 \mathrm{d}c, \tag{4.11c}$$

and

$$\boldsymbol{A}_Q^{(\alpha)} = \frac{4}{3\pi^{1/2}} \int_0^\infty \mathrm{e}^{-c^2} \boldsymbol{A}^{(\alpha)}(c) \left(c^2 - \frac{5}{2}\right) c^3 \mathrm{d}c, \quad \alpha = 1, 2. \tag{4.11d}$$

Since the expressions listed as (4.1a), (4.1b), (4.8), and (4.9) are analytical and continuous in the space variable, we can immediately find results for the normalized mass- and heat-flow rates

$$\boldsymbol{U} = \frac{1}{2a^2} \int_{-a}^a \boldsymbol{U}(\tau) \mathrm{d}\tau \tag{4.12}$$

and

$$\boldsymbol{Q} = \frac{1}{2a^2} \int_{-a}^a \boldsymbol{Q}(\tau) \mathrm{d}\tau, \tag{4.13}$$

where the factor $1/(2a^2)$ is included in order to be consistent with definitions adopted in other works and to facilitate comparisons with numerical results reported in these works. We find

$$\boldsymbol{U} = \frac{1}{2a^2} \left[\boldsymbol{U}_{ps} + a A_1 \boldsymbol{\Phi} + \sum_{j=2}^J \nu_j (A_j + B_j)(1 - \mathrm{e}^{-2a/\nu_j}) \boldsymbol{\mathcal{U}}_j \right] \tag{4.14}$$

and

$$\boldsymbol{Q} = \frac{1}{2a^2} \left[\boldsymbol{Q}_{ps} + \sum_{j=2}^J \nu_j (A_j + B_j)(1 - \mathrm{e}^{-2a/\nu_j}) \boldsymbol{\mathcal{Q}}_j \right], \tag{4.15}$$

TABLE 1
Pressure-driven flow: species-specific velocity, heat-flow, and shear-stress profiles for the Ne-Ar mixture with $2a = 0.1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, *and* $n_1/n_2 = 2/3$.

η	$-U_1(-a+2\eta a)$	$-U_2(-a+2\eta a)$	$Q_1(-a+2\eta a)$	$Q_2(-a+2\eta a)$	$P_1(-a+2\eta a)$	$P_2(-a+2\eta a)$
0.0	2.0012(−1)	1.6301(−1)	6.6293(−2)	4.8754(−2)	1.7779(−2)	3.2742(−2)
0.1	2.0308(−1)	1.7049(−1)	6.7557(−2)	5.1904(−2)	9.6303(−3)	2.1508(−2)
0.2	2.0430(−1)	1.7393(−1)	6.8083(−2)	5.3332(−2)	1.4419(−3)	1.0300(−2)
0.3	2.0455(−1)	1.7547(−1)	6.8208(−2)	5.3981(−2)	−6.7713(−3)	−8.9089(−4)
0.4	2.0395(−1)	1.7549(−1)	6.7987(−2)	5.4019(−2)	−1.5002(−2)	−1.2071(−2)
0.5	2.0254(−1)	1.7409(−1)	6.7431(−2)	5.3493(−2)	−2.3243(−2)	−2.3243(−2)
0.6	2.0028(−1)	1.7128(−1)	6.6527(−2)	5.2399(−2)	−3.1491(−2)	−3.4411(−2)
0.7	1.9708(−1)	1.6691(−1)	6.5233(−2)	5.0678(−2)	−3.9742(−2)	−4.5577(−2)
0.8	1.9274(−1)	1.6068(−1)	6.3464(−2)	4.8191(−2)	−4.7990(−2)	−5.6745(−2)
0.9	1.8682(−1)	1.5183(−1)	6.1023(−2)	4.4605(−2)	−5.6229(−2)	−6.7919(−2)
1.0	1.7702(−1)	1.3641(−1)	5.6912(−2)	3.8171(−2)	−6.4447(−2)	−7.9107(−2)

where

$$\boldsymbol{U}_{ps} = (2a/\varepsilon_0)\{(1/\varepsilon_p)[(1/6)a^2\boldsymbol{\Phi} + \boldsymbol{D}_U]K_p - [\boldsymbol{A}_U^{(1)} + \boldsymbol{A}_U^{(2)}]K_T + [c_2\boldsymbol{A}_U^{(1)} - c_1\boldsymbol{A}_U^{(2)}]K_C\} \tag{4.16}$$

and

$$\boldsymbol{Q}_{ps} = (2a/\varepsilon_0)\{(1/\varepsilon_p)\boldsymbol{D}_Q K_p - [\boldsymbol{A}_Q^{(1)} + \boldsymbol{A}_Q^{(2)}]K_T + [c_2\boldsymbol{A}_Q^{(1)} - c_1\boldsymbol{A}_Q^{(2)}]K_C\}. \tag{4.17}$$

As our solutions are now complete, we are ready for some numerical results.

5. Numerical results. The sample cases for which we report numerical results in this work are defined in terms of two binary mixtures: Ne-Ar and He-Xe. We note that only the mass ratio m_1/m_2, the diameter ratio d_1/d_2, and the density ratio n_1/n_2 are needed to define the LBE for rigid-sphere interactions, and so we use the basic data:

$$m_2 = 39.948, \quad m_1 = 20.183, \quad d_2/d_1 = 1.406, \quad n_1/n_2 = 2/3$$

for the Ne-Ar mixture and

$$m_2 = 131.30, \quad m_1 = 4.0026, \quad d_2/d_1 = 2.226, \quad n_1/n_2 = 2/3$$

for the He-Xe mixture. It should be noted here that the values of the masses of these gas species were taken from [23] and those of the diameter ratios from [19].

We report in Tables 1–12 the velocity, heat-flow, and shear-stress profiles computed for the three considered problems of pressure-driven, temperature-driven, and concentration-driven flow, using as additional input data the accommodation coefficients $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, and $\beta_2 = 0.8$ and two different values of the channel width ($2a = 0.1$ and 1.0). The numerical results reported in Tables 1–12 are thought to be correct to within ± 1 in the last reported digit and were obtained by increasing the values of the approximation parameters $\{L, M, K, N, K_s\}$ of our method in steps, until numerical convergence was observed. Here L is the kernel truncation parameter (the maximum value of n considered in the summation of (2.50)), M is the order of the Gaussian quadrature used to evaluate numerically integrals over the

TABLE 2

Pressure-driven flow: species-specific velocity, heat-flow, and shear-stress profiles for the He-Xe mixture with $2a = 0.1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, *and* $n_1/n_2 = 2/3$.

η	$-U_1(-a+2\eta a)$	$-U_2(-a+2\eta a)$	$Q_1(-a+2\eta a)$	$Q_2(-a+2\eta a)$	$P_1(-a+2\eta a)$	$P_2(-a+2\eta a)$
0.0	1.7251(−1)	1.7483(−1)	7.3532(−2)	4.9529(−2)	1.4254(−2)	3.5944(−2)
0.1	1.7469(−1)	1.8461(−1)	7.4540(−2)	5.4100(−2)	7.6815(−3)	2.3659(−2)
0.2	1.7559(−1)	1.8881(−1)	7.4952(−2)	5.6008(−2)	1.1142(−3)	1.1370(−2)
0.3	1.7577(−1)	1.9069(−1)	7.5037(−2)	5.6886(−2)	−5.4543(−3)	−9.1737(−4)
0.4	1.7532(−1)	1.9079(−1)	7.4839(−2)	5.6994(−2)	−1.2031(−2)	−1.3200(−2)
0.5	1.7426(−1)	1.8925(−1)	7.4367(−2)	5.6408(−2)	−1.8622(−2)	−2.5472(−2)
0.6	1.7257(−1)	1.8606(−1)	7.3610(−2)	5.5124(−2)	−2.5236(−2)	−3.7729(−2)
0.7	1.7017(−1)	1.8106(−1)	7.2531(−2)	5.3067(−2)	−3.1881(−2)	−4.9966(−2)
0.8	1.6692(−1)	1.7388(−1)	7.1059(−2)	5.0049(−2)	−3.8566(−2)	−6.2176(−2)
0.9	1.6246(−1)	1.6353(−1)	6.9031(−2)	4.5589(−2)	−4.5303(−2)	−7.4352(−2)
1.0	1.5509(−1)	1.4439(−1)	6.5634(−2)	3.6907(−2)	−5.2111(−2)	−8.6479(−2)

TABLE 3

Pressure-driven flow: species-specific velocity, heat-flow, and shear-stress profiles for the Ne-Ar mixture with $2a = 1.0$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, *and* $n_1/n_2 = 2/3$.

η	$-U_1(-a+2\eta a)$	$-U_2(-a+2\eta a)$	$Q_1(-a+2\eta a)$	$Q_2(-a+2\eta a)$	$P_1(-a+2\eta a)$	$P_2(-a+2\eta a)$
0.0	1.3905	1.5596	1.3575(−1)	9.4647(−2)	1.6828(−1)	4.3756(−1)
0.1	1.4709	1.8014	1.5711(−1)	1.5084(−1)	1.1953(−1)	3.0339(−1)
0.2	1.5097	1.9042	1.6796(−1)	1.7100(−1)	5.7794(−2)	1.7788(−1)
0.3	1.5257	1.9529	1.7378(−1)	1.8082(−1)	−1.0287(−2)	5.6604(−2)
0.4	1.5209	1.9600	1.7546(−1)	1.8426(−1)	−8.1812(−2)	−6.2380(−2)
0.5	1.4958	1.9292	1.7326(−1)	1.8256(−1)	−1.5489(−1)	−1.8033(−1)
0.6	1.4498	1.8612	1.6702(−1)	1.7579(−1)	−2.2800(−1)	−2.9825(−1)
0.7	1.3818	1.7538	1.5622(−1)	1.6306(−1)	−2.9965(−1)	−4.1716(−1)
0.8	1.2893	1.6006	1.3972(−1)	1.4198(−1)	−3.6806(−1)	−5.3821(−1)
0.9	1.1651	1.3840	1.1491(−1)	1.0648(−1)	−4.3065(−1)	−6.6315(−1)
1.0	9.6070(−1)	9.8737(−1)	6.7860(−2)	2.0640(−2)	−4.8170(−1)	−7.9579(−1)

TABLE 4

Pressure-driven flow: species-specific velocity, heat-flow, and shear-stress profiles for the He-Xe mixture with $2a = 1.0$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, *and* $n_1/n_2 = 2/3$.

η	$-U_1(-a+2\eta a)$	$-U_2(-a+2\eta a)$	$Q_1(-a+2\eta a)$	$Q_2(-a+2\eta a)$	$P_1(-a+2\eta a)$	$P_2(-a+2\eta a)$
0.0	6.4867(−1)	1.8406	1.4476(−1)	9.1086(−2)	7.4981(−2)	5.2277(−1)
0.1	6.8104(−1)	2.1498	1.5355(−1)	1.7507(−1)	5.1223(−2)	3.7194(−1)
0.2	6.9653(−1)	2.2785	1.5721(−1)	2.0267(−1)	2.4189(−2)	2.2330(−1)
0.3	7.0306(−1)	2.3412	1.5884(−1)	2.1651(−1)	−4.4918(−3)	7.5753(−2)
0.4	7.0167(−1)	2.3519	1.5896(−1)	2.2175(−1)	−3.4250(−2)	−7.1074(−2)
0.5	6.9255(−1)	2.3149	1.5770(−1)	2.1998(−1)	−6.4771(−2)	−2.1739(−1)
0.6	6.7546(−1)	2.2308	1.5497(−1)	2.1132(−1)	−9.5859(−2)	−3.6334(−1)
0.7	6.4971(−1)	2.0973	1.5044(−1)	1.9468(−1)	−1.2739(−1)	−5.0898(−1)
0.8	6.1378(−1)	1.9075	1.4339(−1)	1.6716(−1)	−1.5929(−1)	−6.5438(−1)
0.9	5.6382(−1)	1.6418	1.3211(−1)	1.2119(−1)	−1.9153(−1)	−7.9956(−1)
1.0	4.7729(−1)	1.1502	1.0784(−1)	1.7932(−3)	−2.2385(−1)	−9.4468(−1)

speed variable, K is the order of the basis-function approximation introduced in (3.4) and (3.5) to take care of the speed dependence of the solution, N is the number of discrete ordinates used to represent the μ variable in $(0, 1)$, and K_s is the number

TABLE 5

Temperature-driven flow: species-specific velocity, heat-flow, and shear-stress profiles for the Ne-Ar mixture with $2a = 0.1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, *and* $n_1/n_2 = 2/3$.

η	$U_1(-a+2\eta a)$	$U_2(-a+2\eta a)$	$-Q_1(-a+2\eta a)$	$-Q_2(-a+2\eta a)$	$-P_1(-a+2\eta a)$	$-P_2(-a+2\eta a)$
0.0	6.9591(−2)	4.6009(−2)	3.3654(−1)	2.2923(−1)	6.9271(−5)	3.4849(−4)
0.1	7.0821(−2)	4.8693(−2)	3.4162(−1)	2.3995(−1)	9.1318(−5)	3.3379(−4)
0.2	7.1328(−2)	4.9914(−2)	3.4369(−1)	2.4483(−1)	9.9311(−5)	3.2847(−4)
0.3	7.1444(−2)	5.0468(−2)	3.4414(−1)	2.4704(−1)	9.9592(−5)	3.2828(−4)
0.4	7.1223(−2)	5.0495(−2)	3.4320(−1)	2.4716(−1)	9.5821(−5)	3.3079(−4)
0.5	7.0675(−2)	5.0036(−2)	3.4092(−1)	2.4536(−1)	9.0902(−5)	3.3407(−4)
0.6	6.9786(−2)	4.9086(−2)	3.3723(−1)	2.4160(−1)	8.7513(−5)	3.3633(−4)
0.7	6.8517(−2)	4.7594(−2)	3.3197(−1)	2.3566(−1)	8.8411(−5)	3.3573(−4)
0.8	6.6783(−2)	4.5443(−2)	3.2477(−1)	2.2707(−1)	9.6781(−5)	3.3015(−4)
0.9	6.4392(−2)	4.2344(−2)	3.1481(−1)	2.1462(−1)	1.1689(−4)	3.1675(−4)
1.0	6.0363(−2)	3.6811(−2)	2.9793(−1)	1.9222(−1)	1.5656(−4)	2.9030(−4)

TABLE 6

Temperature-driven flow: species-specific velocity, heat-flow, and shear-stress profiles for the He-Xe mixture with $2a = 0.1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, *and* $n_1/n_2 = 2/3$.

η	$U_1(-a+2\eta a)$	$U_2(-a+2\eta a)$	$-Q_1(-a+2\eta a)$	$-Q_2(-a+2\eta a)$	$P_1(-a+2\eta a)$	$-P_2(-a+2\eta a)$
0.0	7.6247(−2)	4.0307(−2)	3.4185(−1)	2.1217(−1)	−2.3469(−4)	1.8717(−4)
0.1	7.7279(−2)	4.3532(−2)	3.4621(−1)	2.2424(−1)	−1.4265(−4)	2.4853(−4)
0.2	7.7703(−2)	4.4892(−2)	3.4798(−1)	2.2939(−1)	−4.3015(−5)	3.1495(−4)
0.3	7.7789(−2)	4.5519(−2)	3.4834(−1)	2.3174(−1)	6.0403(−5)	3.8389(−4)
0.4	7.7585(−2)	4.5592(−2)	3.4747(−1)	2.3196(−1)	1.6554(−4)	4.5398(−4)
0.5	7.7098(−2)	4.5163(−2)	3.4541(−1)	2.3024(−1)	2.7077(−4)	5.2414(−4)
0.6	7.6316(−2)	4.4230(−2)	3.4212(−1)	2.2658(−1)	3.7457(−4)	5.9334(−4)
0.7	7.5204(−2)	4.2737(−2)	3.3743(−1)	2.2073(−1)	4.7534(−4)	6.6052(−4)
0.8	7.3690(−2)	4.0546(−2)	3.3104(−1)	2.1217(−1)	5.7114(−4)	7.2438(−4)
0.9	7.1607(−2)	3.7317(−2)	3.2222(−1)	1.9955(−1)	6.5924(−4)	7.8312(−4)
1.0	6.8134(−2)	3.1091(−2)	3.0745(−1)	1.7544(−1)	7.3409(−4)	8.3302(−4)

TABLE 7

Temperature-driven flow: species-specific velocity, heat-flow, and shear-stress profiles for the Ne-Ar mixture with $2a = 1.0$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, *and* $n_1/n_2 = 2/3$.

η	$U_1(-a+2\eta a)$	$U_2(-a+2\eta a)$	$-Q_1(-a+2\eta a)$	$-Q_2(-a+2\eta a)$	$P_1(-a+2\eta a)$	$-P_2(-a+2\eta a)$
0.0	1.6352(−1)	1.1439(−1)	7.3626(−1)	5.3931(−1)	−1.5921(−3)	1.8955(−3)
0.1	1.7493(−1)	1.3808(−1)	7.7567(−1)	6.1916(−1)	−2.3711(−3)	1.3762(−3)
0.2	1.8001(−1)	1.4709(−1)	7.9100(−1)	6.4734(−1)	−2.2690(−3)	1.4442(−3)
0.3	1.8245(−1)	1.5156(−1)	7.9742(−1)	6.6053(−1)	−1.8145(−3)	1.7473(−3)
0.4	1.8282(−1)	1.5304(−1)	7.9733(−1)	6.6445(−1)	−1.1850(−3)	2.1669(−3)
0.5	1.8121(−1)	1.5202(−1)	7.9125(−1)	6.6080(−1)	−4.6766(−4)	2.6451(−3)
0.6	1.7752(−1)	1.4845(−1)	7.7869(−1)	6.4946(−1)	2.7989(−4)	3.1435(−3)
0.7	1.7133(−1)	1.4188(−1)	7.5812(−1)	6.2872(−1)	9.9769(−4)	3.6220(−3)
0.8	1.6181(−1)	1.3114(−1)	7.2619(−1)	5.9426(−1)	1.5901(−3)	4.0170(−3)
0.9	1.4703(−1)	1.1335(−1)	6.7487(−1)	5.3501(−1)	1.8583(−3)	4.1958(−3)
1.0	1.1654(−1)	7.1981(−2)	5.6056(−1)	3.8620(−1)	1.2125(−3)	3.7653(−3)

of spline functions used to compute (without postprocessing) the Chapman–Enskog (vector-valued) functions $\boldsymbol{A}^{(1)}(c)$, $\boldsymbol{A}^{(2)}(c)$, $\boldsymbol{B}(c)$, $\boldsymbol{D}(c)$, and $\boldsymbol{E}(c)$, as explained in de-

TABLE 8

Temperature-driven flow: species-specific velocity, heat-flow, and shear-stress profiles for the He-Xe mixture with $2a = 1.0$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, *and* $n_1/n_2 = 2/3$.

η	$U_1(-a+2\eta a)$	$U_2(-a+2\eta a)$	$-Q_1(-a+2\eta a)$	$-Q_2(-a+2\eta a)$	$P_1(-a+2\eta a)$	$-P_2(-a+2\eta a)$
0.0	1.6158(−1)	9.7015(−2)	6.8378(−1)	4.7673(−1)	−1.6983(−3)	1.0746(−3)
0.1	1.6960(−1)	1.2111(−1)	7.1404(−1)	5.5454(−1)	−1.5536(−3)	1.1711(−3)
0.2	1.7288(−1)	1.2977(−1)	7.2547(−1)	5.8063(−1)	−1.0929(−3)	1.4782(−3)
0.3	1.7427(−1)	1.3418(−1)	7.2994(−1)	5.9301(−1)	−5.1014(−4)	1.8667(−3)
0.4	1.7425(−1)	1.3582(−1)	7.2943(−1)	5.9702(−1)	1.4055(−4)	2.3005(−3)
0.5	1.7293(−1)	1.3512(−1)	7.2436(−1)	5.9424(−1)	8.3761(−4)	2.7652(−3)
0.6	1.7019(−1)	1.3205(−1)	7.1433(−1)	5.8455(−1)	1.5696(−3)	3.2532(−3)
0.7	1.6572(−1)	1.2617(−1)	6.9803(−1)	5.6636(−1)	2.3234(−3)	3.7557(−3)
0.8	1.5882(−1)	1.1639(−1)	6.7267(−1)	5.3564(−1)	3.0706(−3)	4.2538(−3)
0.9	1.4786(−1)	9.9969(−2)	6.3148(−1)	4.8193(−1)	3.7379(−3)	4.6988(−3)
1.0	1.2437(−1)	5.9234(−2)	5.3915(−1)	3.3921(−1)	4.0610(−3)	4.9141(−3)

TABLE 9

Concentration-driven flow: species-specific velocity, heat-flow, and shear-stress profiles for the Ne-Ar mixture with $2a = 0.1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, *and* $n_1/n_2 = 2/3$.

η	$-U_1(-a+2\eta a)$	$U_2(-a+2\eta a)$	$Q_1(-a+2\eta a)$	$-Q_2(-a+2\eta a)$	$P_1(-a+2\eta a)$	$P_2(-a+2\eta a)$
0.0	8.5004(−2)	3.3436(−2)	3.2208(−2)	9.7700(−3)	7.5261(−3)	−6.8152(−3)
0.1	8.6279(−2)	3.4967(−2)	3.2792(−2)	1.0403(−2)	4.0901(−3)	−4.5245(−3)
0.2	8.6795(−2)	3.5673(−2)	3.3027(−2)	1.0691(−2)	6.8986(−4)	−2.2577(−3)
0.3	8.6903(−2)	3.5997(−2)	3.3081(−2)	1.0828(−2)	−2.6950(−3)	−1.1075(−6)
0.4	8.6666(−2)	3.6015(−2)	3.2985(−2)	1.0847(−2)	−6.0786(−3)	2.2546(−3)
0.5	8.6094(−2)	3.5752(−2)	3.2742(−2)	1.0758(−2)	−9.4735(−3)	4.5179(−3)
0.6	8.5173(−2)	3.5201(−2)	3.2347(−2)	1.0559(−2)	−1.2892(−2)	6.7972(−3)
0.7	8.3857(−2)	3.4331(−2)	3.1777(−2)	1.0236(−2)	−1.6349(−2)	9.1014(−3)
0.8	8.2054(−2)	3.3072(−2)	3.0988(−2)	9.7578(−3)	−1.9858(−2)	1.1441(−2)
0.9	7.9550(−2)	3.1254(−2)	2.9878(−2)	9.0476(−3)	−2.3442(−2)	1.3830(−2)
1.0	7.5288(−2)	2.8012(−2)	2.7939(−2)	7.7216(−3)	−2.7137(−2)	1.6294(−2)

TABLE 10

Concentration-driven flow: species-specific velocity, heat-flow, and shear-stress profiles for the He-Xe mixture with $2a = 0.1$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, *and* $n_1/n_2 = 2/3$.

η	$-U_1(-a+2\eta a)$	$U_2(-a+2\eta a)$	$Q_1(-a+2\eta a)$	$-Q_2(-a+2\eta a)$	$P_1(-a+2\eta a)$	$P_2(-a+2\eta a)$
0.0	9.3618(−2)	3.3805(−2)	4.3904(−2)	9.3958(−3)	7.6577(−3)	−6.9794(−3)
0.1	9.4798(−2)	3.5686(−2)	4.4496(−2)	1.0264(−2)	4.1248(−3)	−4.6241(−3)
0.2	9.5276(−2)	3.6496(−2)	4.4737(−2)	1.0627(−2)	6.1580(−4)	−2.2848(−3)
0.3	9.5370(−2)	3.6870(−2)	4.4785(−2)	1.0800(−2)	−2.8838(−3)	4.8252(−5)
0.4	9.5131(−2)	3.6911(−2)	4.4669(−2)	1.0833(−2)	−6.3842(−3)	2.3819(−3)
0.5	9.4572(−2)	3.6648(−2)	4.4392(−2)	1.0741(−2)	−9.8947(−3)	4.7222(−3)
0.6	9.3677(−2)	3.6077(−2)	4.3949(−2)	1.0523(−2)	−1.3425(−2)	7.0757(−3)
0.7	9.2405(−2)	3.5164(−2)	4.3317(−2)	1.0161(−2)	−1.6985(−2)	9.4488(−3)
0.8	9.0671(−2)	3.3826(−2)	4.2455(−2)	9.6159(−3)	−2.0586(−2)	1.1850(−2)
0.9	8.8280(−2)	3.1859(−2)	4.1263(−2)	8.7885(−3)	−2.4244(−2)	1.4288(−2)
1.0	8.4274(−2)	2.8114(−2)	3.9260(−2)	7.1141(−3)	−2.7987(−2)	1.6783(−2)

tail in [8] and [11]. To be more specific, we note that we have used $20 \le L \le 95$, $100 \le M \le 400$, $20 \le K \le 35$, $20 \le N \le 50$, and $80 \le K_s - 2 \le 1280$. In addition to the profiles reported in Tables 1–12, we report in Tables 13–15 mass- and heat-flow

TABLE 11

Concentration-driven flow: species-specific velocity, heat-flow, and shear-stress profiles for the Ne-Ar mixture with $2a = 1.0$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, *and* $n_1/n_2 = 2/3$.

η	$-U_1(-a+2\eta a)$	$U_2(-a+2\eta a)$	$Q_1(-a+2\eta a)$	$-Q_2(-a+2\eta a)$	$P_1(-a+2\eta a)$	$P_2(-a+2\eta a)$
0.0	1.6697(−1)	6.1119(−2)	5.0453(−2)	5.4942(−3)	1.9191(−2)	−1.7106(−2)
0.1	1.7575(−1)	7.0455(−2)	5.3750(−2)	7.7513(−3)	1.1026(−2)	−1.1663(−2)
0.2	1.7865(−1)	7.4230(−2)	5.4732(−2)	8.3674(−3)	4.9631(−3)	−7.6208(−3)
0.3	1.7964(−1)	7.6312(−2)	5.5100(−2)	8.6958(−3)	−1.8189(−4)	−4.1908(−3)
0.4	1.7935(−1)	7.7304(−2)	5.5117(−2)	8.9489(−3)	−4.9955(−3)	−9.8168(−4)
0.5	1.7796(−1)	7.7372(−2)	5.4846(−2)	9.1930(−3)	−9.9215(−3)	2.3023(−3)
0.6	1.7537(−1)	7.6472(−2)	5.4254(−2)	9.4308(−3)	−1.5388(−2)	5.9467(−3)
0.7	1.7122(−1)	7.4366(−2)	5.3204(−2)	9.6061(−3)	−2.1898(−2)	1.0287(−2)
0.8	1.6470(−1)	7.0492(−2)	5.1378(−2)	9.5542(−3)	−3.0148(−2)	1.5786(−2)
0.9	1.5379(−1)	6.3444(−2)	4.7937(−2)	8.7972(−3)	−4.1284(−2)	2.3211(−2)
1.0	1.2756(−1)	4.5357(−2)	3.8020(−2)	4.3594(−3)	−5.7978(−2)	3.4340(−2)

TABLE 12

Concentration-driven flow: species-specific velocity, heat-flow, and shear-stress profiles for the He-Xe mixture with $2a = 1.0$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, *and* $n_1/n_2 = 2/3$.

η	$-U_1(-a+2\eta a)$	$U_2(-a+2\eta a)$	$Q_1(-a+2\eta a)$	$-Q_2(-a+2\eta a)$	$P_1(-a+2\eta a)$	$P_2(-a+2\eta a)$
0.0	1.9140(−1)	6.6006(−2)	8.6133(−2)	2.3421(−3)	2.0835(−2)	−1.8831(−2)
0.1	1.9968(−1)	7.6638(−2)	9.0237(−2)	4.9463(−3)	1.2501(−2)	−1.3275(−2)
0.2	2.0272(−1)	8.0883(−2)	9.1762(−2)	5.6234(−3)	5.6487(−3)	−8.7065(−3)
0.3	2.0383(−1)	8.3276(−2)	9.2365(−2)	6.0203(−3)	−5.6862(−4)	−4.5616(−3)
0.4	2.0357(−1)	8.4413(−2)	9.2316(−2)	6.3456(−3)	−6.6101(−3)	−5.3396(−4)
0.5	2.0206(−1)	8.4450(−2)	9.1676(−2)	6.6616(−3)	−1.2845(−2)	3.6227(−3)
0.6	1.9919(−1)	8.3338(−2)	9.0391(−2)	6.9674(−3)	−1.9640(−2)	8.1526(−3)
0.7	1.9461(−1)	8.0837(−2)	8.8283(−2)	7.2021(−3)	−2.7422(−2)	1.3341(−2)
0.8	1.8755(−1)	7.6392(−2)	8.4971(−2)	7.1972(−3)	−3.6768(−2)	1.9571(−2)
0.9	1.7615(−1)	6.8612(−2)	7.9529(−2)	6.4773(−3)	−4.8586(−2)	2.7450(−2)
1.0	1.5069(−1)	4.8837(−2)	6.7082(−2)	1.7050(−3)	−6.4896(−2)	3.8323(−2)

rates, as defined by (4.14)–(4.17), for several values of the channel width. We note that the composition and the wall interaction data used to generate Tables 13–15 were the same as those used for Tables 1–12, and that the numerical results for the flow rates are also thought to be correct to within ± 1 in the last reported digit.

While an implementation of our solutions for the three considered problems requires, in general, some hours of computer time to establish the high-quality results we are reporting in our tables, solutions good enough for graphical presentation require very modest computational expense. To have an idea of the CPU time for what we might consider "practical results," we found, for example, that all of the He-Xe results given in Tables 1–15 could be obtained with essentially three figures of accuracy in less than one minute on an Apple MacBook running at 2 GHz.

Finally, we note that we have (for the three considered problems) compared numerical results from our approach based on the LBE for binary mixtures with those of the McCormack model, as developed and implemented in [23]. Due to different ways of the defining the dimensionless space variables in [23] and in this work, we have used the relationship

$$a = \xi_M a_M,$$

TABLE 13

Pressure-driven flow: mass- and heat-flow rates for the case $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, and $n_1/n_2 = 2/3$.

	Ne-Ar mixture			
$2a$	$-U_1$	$-U_2$	Q_1	Q_2
1.0(−2)	7.02875	4.57921	3.05198	1.90616
1.0(−1)	3.97126	3.34605	1.31567	1.01481
5.0(−1)	2.88363	3.26237	5.15113(−1)	4.87750(−1)
1.0	2.80472	3.50334	3.07392(−1)	3.07713(−1)
2.0	3.02335	4.02064	1.72501(−1)	1.77501(−1)
5.0	3.94713	5.45691	7.46936(−2)	7.77126(−2)
1.0(1)	5.49952	7.68822	3.82931(−2)	3.99837(−2)
1.0(2)	3.27996(1)	4.61398(1)	3.90604(−3)	4.09657(−3)
	He-Xe mixture			
$2a$	$-U_1$	$-U_2$	Q_1	Q_2
1.0(−2)	7.28191	4.43169	3.45590	1.72889
1.0(−1)	3.42560	3.62500	1.46027	1.05889
5.0(−1)	1.68434	3.80750	5.26626(−1)	5.60196(−1)
1.0	1.31140	4.18704	2.99703(−1)	3.63014(−1)
2.0	1.16837	4.90801	1.61420(−1)	2.13010(−1)
5.0	1.32920	6.86910	6.77017(−2)	9.41384(−2)
1.0(1)	1.80118	9.93900	3.43884(−2)	4.85365(−2)
1.0(2)	1.10778(1)	6.34092(1)	3.48652(−3)	4.98132(−3)

TABLE 14

Temperature-driven flow: mass- and heat-flow rates for the case $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, and $n_1/n_2 = 2/3$.

	Ne-Ar mixture			
$2a$	U_1	U_2	$-Q_1$	$-Q_2$
1.0(−2)	3.08637	1.87391	1.49378(1)	9.36209
1.0(−1)	1.38130	9.53251(−1)	6.68507	4.71728
5.0(−1)	5.65163(−1)	4.40808(−1)	2.59532	2.04733
1.0	3.41095(−1)	2.76103(−1)	1.51266	1.23229
2.0	1.91966(−1)	1.59245(−1)	8.25888(−1)	6.85471(−1)
5.0	8.29642(−2)	6.99555(−2)	3.48778(−1)	2.92661(−1)
1.0(1)	4.25182(−2)	3.60209(−2)	1.77504(−1)	1.49452(−1)
1.0(2)	4.34037(−3)	3.68920(−3)	1.80306(−2)	1.52261(−2)
	He-Xe mixture			
$2a$	U_1	U_2	$-Q_1$	$-Q_2$
1.0(−2)	3.47992	1.63717	1.59823(1)	8.73594
1.0(−1)	1.51411	8.53282(−1)	6.79083	4.41149
5.0(−1)	5.71005(−1)	3.90743(−1)	2.45191	1.85972
1.0	3.30863(−1)	2.44038(−1)	1.39410	1.10712
2.0	1.80382(−1)	1.40608(−1)	7.49405(−1)	6.11321(−1)
5.0	7.61935(−2)	6.17139(−2)	3.13429(−1)	2.59709(−1)
1.0(1)	3.87787(−2)	3.17730(−2)	1.59023(−1)	1.32410(−1)
1.0(2)	3.93853(−3)	3.25543(−3)	1.61100(−2)	1.34710(−2)

where ξ_M is the conversion factor defined by equation (7.19) of [9], to relate the channel half-width a used in this work with the a_M used in [23]. For the mass- and heat-flow rates, this is the only conversion that is required; for the profiles, in addition to the channel half-width conversion, the LBE results must be divided by ξ_M, in order to be properly compared to the results of [23]. Thus, concerning the mass-

TABLE 15

Concentration-driven flow: mass- and heat-flow rates for the case $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\beta_1 = 0.6$, $\beta_2 = 0.8$, and $n_1/n_2 = 2/3$.

	Ne-Ar mixture			
$2a$	$-U_1$	U_2	Q_1	$-Q_2$
1.0(−2)	3.96894	1.59880	1.75432	6.58115(−1)
1.0(−1)	1.68853	6.87292(−1)	6.40069(−1)	2.04304(−1)
5.0(−1)	6.06366(−1)	2.50910(−1)	1.95828(−1)	4.19489(−2)
1.0	3.41892(−1)	1.43652(−1)	1.05381(−1)	1.73285(−2)
2.0	1.82236(−1)	7.77918(−2)	5.47167(−2)	6.98903(−3)
5.0	7.57830(−2)	3.25972(−2)	2.23853(−2)	2.28975(−3)
1.0(1)	3.84251(−2)	1.64751(−2)	1.12765(−2)	1.05773(−3)
1.0(2)	3.89650(−3)	1.65480(−3)	1.13525(−3)	9.79098(−5)
	He-Xe mixture			
$2a$	$-U_1$	U_2	Q_1	Q_2
1.0(−2)	4.30486	1.52721	2.07209	−5.94298(−1)
1.0(−1)	1.85963	7.02653(−1)	8.71981(−1)	−2.01990(−1)
5.0(−1)	6.81859(−1)	2.67760(−1)	3.11395(−1)	−3.71345(−2)
1.0	3.89145(−1)	1.56308(−1)	1.76153(−1)	−1.22203(−2)
2.0	2.09485(−1)	8.63759(−2)	9.43221(−2)	−3.14174(−3)
5.0	8.76311(−2)	3.70753(−2)	3.93363(−2)	−3.25368(−4)
1.0(1)	4.44688(−2)	1.89313(−2)	1.99386(−2)	6.323 (−7)
1.0(2)	4.50724(−3)	1.91922(−3)	2.01820(−3)	1.48122(−5)

flow rates reported in Tables 1–9 of [23], we have found maximum relative deviations (with respect to our LBE results) of 33%, 62%, and 33% for the problems driven by pressure, temperature, and concentration gradients, respectively. For the heat-flow rates reported in these same tables of [23], we have found maximum relative deviations of 40%, 34%, and 300%, respectively, for the pressure-, temperature-, and concentration-driven problems. In all cases but one, the maximum deviations were found to occur for the following input parameters considered in [23]: the heaviest gas particle (Xe), the widest channel ($2a_M = 100$), and the largest concentration of the lighter species ($c_1 = 0.9$). Large maximum relative deviations between the McCormack profiles reported in Tables 10–18 of [23] and those computed with our current (LBE) approach were also observed.

6. Onsager relationships. In [23], Siewert and Valougeorgis established three independent (generalized) Onsager relationships relevant to the flow of binary gas mixtures in a plane-parallel channel driven by pressure, temperature, and concentration gradients. While the derivations reported in [23] were based on the McCormack kinetic model [15], little work is required to establish those same relationships [23] for the LBE (for rigid-sphere interactions) used in this work. For that purpose, we follow here a procedure described in detail for half-space flow problems in [9]. However, before starting our derivation, we should mention that, to denote the solutions and the driving terms of two different problems (among the three that can be defined by considering separately pressure, temperature, and concentration gradients), we attach subscripts X and Y to $\boldsymbol{\Psi}(\tau, c, \mu)$ and to $\boldsymbol{\Upsilon}(c)$.

In short, using the fact that the kernel defined by (2.25) is such that

$$\boldsymbol{S}\boldsymbol{K}^T(\boldsymbol{c} : \boldsymbol{c}') = \boldsymbol{K}(\boldsymbol{c}' : \boldsymbol{c})\boldsymbol{S}, \tag{6.1}$$

where

$$\boldsymbol{S} = \begin{bmatrix} c_2 & 0 \\ 0 & c_1 a_{1,2} \end{bmatrix} \tag{6.2}$$

and $a_{1,2}$ is given by (2.23), we can multiply (2.47) with μ changed to $-\mu$ and subscript Y added to $\boldsymbol{\Psi}(\tau, c, \mu)$ and $\boldsymbol{\Upsilon}(c)$ by

$$c^2(1-\mu^2)\mathrm{e}^{-c^2}\boldsymbol{\Psi}_X^T(\tau, c, \mu)\boldsymbol{S}^{-1},$$

multiply (2.47) with subscript X added to $\boldsymbol{\Psi}(\tau, c, \mu)$ and $\boldsymbol{\Upsilon}(c)$ by

$$c^2(1-\mu^2)\mathrm{e}^{-c^2}\boldsymbol{\Psi}_Y^T(\tau, c, -\mu)\boldsymbol{S}^{-1},$$

subtract the resulting equations, one from the other, and integrate the result of this operation over all μ, over all c, and over τ from $-a$ to a to find, after using (2.62),

$$\int_{-a}^{a}\int_{0}^{\infty}\int_{-1}^{1} \mathrm{e}^{-c^2}c^2(1-\mu^2)\big[\boldsymbol{\Psi}_X^T(\tau, c, \mu)\boldsymbol{S}^{-1}\boldsymbol{\Upsilon}_Y(c) - \boldsymbol{\Psi}_Y^T(\tau, c, -\mu)\boldsymbol{S}^{-1}\boldsymbol{\Upsilon}_X(c)\big]\mathrm{d}\mu\mathrm{d}c\mathrm{d}\tau = 0. \tag{6.3}$$

Taking all possible combinations of X and Y (with the restriction that $X \neq Y$) when these subscripts are set equal to P, T, and C in (6.3) and using the forms of the driving terms appropriate to each problem, we find the relationships

$$K_T\,[c_1 a_{1,2} \quad c_2]\,\boldsymbol{Q}_P = K_P\,[c_1 a_{1,2} \quad c_2]\,\boldsymbol{U}_T, \tag{6.4a}$$

$$K_T\,[c_1 a_{1,2} \quad c_2]\,\boldsymbol{Q}_C = c_1 c_2 K_C\,[a_{1,2} \quad -1]\,\boldsymbol{U}_T, \tag{6.4b}$$

and

$$c_1 c_2 K_C\,[a_{1,2} \quad -1]\,\boldsymbol{U}_P = K_P\,[c_1 a_{1,2} \quad c_2]\,\boldsymbol{U}_C, \tag{6.4c}$$

where we have added subscripts P, T, C to the quantities defined by (4.14) and (4.15) as tags for the problems driven, respectively, by pressure, temperature, and concentration gradients. As a (minor) test of our computations, we have confirmed the three identities listed as (6.4) for the data sets used to define the numerical results reported in this work. Moreover, since Kosuge et al. [13] have tabulated numerical results related to our (6.4), we include in Table 16 our numerical results for the quantities used in [13] to express the (generalized) Onsager relationships. We note that the results listed in Table 16 are relevant to the special case [13] of strictly diffuse reflection at both walls, equal-diameter particles, mass ratio $m_2/m_1 = 2$, and density ratio $n_2/n_1 = 1$ at equilibrium. In order to compare with [13], we used the channel half-width

$$a = \left(\frac{1}{k}\right)\left[\frac{(c_1 + c_2 d_2/d_1)^2}{4(2^{1/2})c_1 + c_2(1 + d_1/d_2)^2(1 + m_1/m_2)^{1/2}}\right], \tag{6.5}$$

where k is the Knudsen number used in [13], and the following expressions (valid for

TABLE 16
The quantities Λ_{XY}, $X \neq Y$, $X, Y = P, T, C$ as defined in [13] *for various values of k with $\alpha_1 = 1.0$, $\alpha_2 = 1.0$, $\beta_1 = 1.0$, $\beta_2 = 1.0$, $m_2/m_1 = 2$, $d_2/d_1 = 1$, and $n_2/n_1 = 1$.*

k	Λ_{TP}	Λ_{PT}	$-\Lambda_{CP}$	$-\Lambda_{PC}$	Λ_{CT}	Λ_{TC}
0.05	4.622713(−2)	4.622713(−2)	1.248352(−2)	1.248352(−2)	8.743064(−3)	8.743064(−3)
0.10	8.584004(−2)	8.584004(−2)	2.371949(−2)	2.371949(−2)	1.646907(−2)	1.646907(−2)
1.00	3.497536(−1)	3.497536(−1)	1.202181(−1)	1.202181(−1)	7.322853(−2)	7.322853(−2)
10.0	7.076139(−1)	7.076139(−1)	2.785816(−1)	2.785816(−1)	1.472278(−1)	1.472278(−1)
20.0	8.340453(−1)	8.340453(−1)	3.317800(−1)	3.317800(−1)	1.717606(−1)	1.717606(−1)

K_P, K_T, and K_C set equal to ε_0) for the quantities defined in [13]:

$$\Lambda_{PT} = \frac{1}{2c_1} a_{2,1} \begin{bmatrix} c_1 a_{1,2} & c_2 \end{bmatrix} \boldsymbol{U}_T, \tag{6.6a}$$

$$\Lambda_{PC} = \frac{1}{2c_1 c_2} a_{2,1} \begin{bmatrix} c_1 a_{1,2} & c_2 \end{bmatrix} \boldsymbol{U}_C, \tag{6.6b}$$

$$\Lambda_{TP} = \frac{1}{2c_1} a_{2,1} \begin{bmatrix} c_1 a_{1,2} & c_2 \end{bmatrix} \boldsymbol{Q}_P, \tag{6.6c}$$

$$\Lambda_{TC} = \frac{1}{2c_1 c_2} a_{2,1} \begin{bmatrix} c_1 a_{1,2} & c_2 \end{bmatrix} \boldsymbol{Q}_C, \tag{6.6d}$$

$$\Lambda_{CP} = \frac{1}{2} a_{2,1} \begin{bmatrix} a_{1,2} & -1 \end{bmatrix} \boldsymbol{U}_P, \tag{6.6e}$$

and

$$\Lambda_{CT} = \frac{1}{2} a_{2,1} \begin{bmatrix} a_{1,2} & -1 \end{bmatrix} \boldsymbol{U}_T. \tag{6.6f}$$

Note that subscript D is used in [13] with the same meaning as subscript C in this work (i.e., a tag for the concentration-driven problem). To be clear, we have listed identical results in various columns of Table 16 in order to emphasize that all quantities were computed as defined.

Finally, we note that we have also confirmed that

$$p_* = \begin{bmatrix} c_1 & c_2 \end{bmatrix} \boldsymbol{P}(\tau) + (K_P/\varepsilon_0)\tau,$$

where the second term on the right-hand side should not be taken into account for the cases of temperature and concentration gradients, is a (problem-dependent) constant.

7. Concluding remarks. We have reported in this work what we believe to be a compact, fast, and accurate method of solving channel-flow problems driven by pressure, temperature, and concentration gradients and described by the (vector) LBE for a binary mixture of rigid spheres. Accurate numerical results were given for the velocity, heat-flow, and shear-stress profiles, as well as for the mass- and heat-flow rates, for selected cases based on Ne-Ar and He-Xe mixtures.

TABLE 17
Refined results for Tables 10, 11, *and* 12 *of* [21] *in the notation of* [21].

	$-U_P$			Q_P		
$2a$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1.0$
0.10	2.0244(1)	4.3874	1.9504	4.1702	1.5684	7.9969(−1)
1.00	1.7564(1)	3.3264	1.5067	7.1258(−1)	5.2875(−1)	3.8908(−1)
10.0	1.8743(1)	4.5346	2.7296	7.9139(−2)	8.4299(−2)	8.9950(−2)
	U_T			$-Q_T$		
$2a$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1.0$
0.10	4.1702	1.5684	7.9969(−1)	2.0650(1)	7.7804	3.9044
1.00	7.1258(−1)	5.2875(−1)	3.8908(−1)	3.4557	2.5138	1.7830
10.0	7.9139(−2)	8.4299(−2)	8.9950(−2)	3.7488(−1)	3.6167(−1)	3.4674(−1)

In addition to the comparisons with the numerical results of the McCormack model that are discussed in section 5, we have also performed comparisons with the single-gas LBE results of [21], using three different ways of achieving the single-gas limit in our formulation:

$$\text{(i) } c_1 = 0, \quad \text{(ii) } c_2 = 0, \quad \text{or} \quad \text{(iii) } m_1 = m_2,\ d_1 = d_2,\ \alpha_1 = \alpha_2, \text{ and } \beta_1 = \beta_2.$$

We note that to convert our results to the same spatial units used in [21] we made use of the factor

$$\xi_{S,p} = 0.449027806\ldots,$$

which (for a single-species case) is the ratio between our dimensionless spatial variable, as defined by (2.41) and (2.42), and that used in [21] for channel-flow problems. Doing this, we found good but not perfect agreement with the five-figure results for the mass- and heat-flow rates and for the velocity and heat-flow profiles that are tabulated in [21]. In regard to the flow rates, while we found at most a difference of one unit in the fifth digit listed in Table 10 of [21], where the accommodation coefficients are taken to be equal to 0.1, we did find a maximum difference of 7 units in the fifth digit listed in Table 11 of [21] (case with accommodation coefficients equal to 0.5) and a maximum difference of 4 units in the fourth digit listed in Table 12 of [21] (case with accommodation coefficients equal to 1.0). The largest differences always occurred for the smallest channel width considered in Tables 10–12 of [21]. For the velocity and heat-flow profiles listed in Tables 13 and 14 of [21], we have observed, respectively, maximum differences of 5 and 3 units in the fifth digit listed in these tables. The maximum differences for the profiles were found to always occur at the channel walls. We have confirmed that the loss of accuracy in Tables 10–14 of [21] was due to using $L = 8$ in those computations, and so we list in Tables 17–19 our improved results (based on $L = 30$) for the cases studied in Tables 10–14 of [21].

Finally, we should like to mention that, considering the large deviations between the numerical results from the LBE and those from the McCormack model that were observed in this and other [9, 10] works, we are of the opinion that the McCormack model has a limited value as an economical alternative to the LBE for gas mixtures.

TABLE 18
Refined results for Table 13 *of* [21] *in the notation of* [21].

	$\alpha = 0.1$		$\alpha = 0.5$		$\alpha = 1.0$	
τ/a	$-u_P(\tau)$	$q_P(\tau)$	$-u_P(\tau)$	$q_P(\tau)$	$-u_P(\tau)$	$q_P(\tau)$
0.0	8.8693	3.7271(−1)	1.7574	2.8921(−1)	8.5378(−1)	2.2669(−1)
0.1	8.8671	3.7230(−1)	1.7549	2.8859(−1)	8.5116(−1)	2.2589(−1)
0.2	8.8602	3.7106(−1)	1.7475	2.8672(−1)	8.4327(−1)	2.2348(−1)
0.3	8.8486	3.6895(−1)	1.7350	2.8355(−1)	8.2994(−1)	2.1938(−1)
0.4	8.8320	3.6592(−1)	1.7172	2.7898(−1)	8.1090(−1)	2.1348(−1)
0.5	8.8101	3.6187(−1)	1.6935	2.7288(−1)	7.8568(−1)	2.0559(−1)
0.6	8.7822	3.5667(−1)	1.6635	2.6501(−1)	7.5357(−1)	1.9539(−1)
0.7	8.7473	3.5006(−1)	1.6258	2.5499(−1)	7.1335(−1)	1.8239(−1)
0.8	8.7035	3.4160(−1)	1.5785	2.4212(−1)	6.6281(−1)	1.6568(−1)
0.9	8.6461	3.3023(−1)	1.5167	2.2483(−1)	5.9696(−1)	1.4323(−1)
1.0	8.5500	3.1009(−1)	1.4143	1.9464(−1)	4.8982(−1)	1.0466(−1)

TABLE 19
Refined results for Table 14 *of* [21] *in the notation of* [21].

	$\alpha = 0.1$		$\alpha = 0.5$		$\alpha = 1.0$	
τ/a	$u_T(\tau)$	$-q_T(\tau)$	$u_T(\tau)$	$-q_T(\tau)$	$u_T(\tau)$	$-q_T(\tau)$
0.0	3.6061(−1)	1.7429	2.8168(−1)	1.3193	2.2268(−1)	9.9636(−1)
0.1	3.6050(−1)	1.7425	2.8125(−1)	1.3178	2.2198(−1)	9.9383(−1)
0.2	3.6018(−1)	1.7414	2.7995(−1)	1.3132	2.1987(−1)	9.8616(−1)
0.3	3.5963(−1)	1.7395	2.7775(−1)	1.3054	2.1629(−1)	9.7311(−1)
0.4	3.5883(−1)	1.7368	2.7457(−1)	1.2942	2.1113(−1)	9.5424(−1)
0.5	3.5777(−1)	1.7332	2.7032(−1)	1.2790	2.0422(−1)	9.2884(−1)
0.6	3.5640(−1)	1.7284	2.6484(−1)	1.2593	1.9530(−1)	8.9575(−1)
0.7	3.5466(−1)	1.7223	2.5785(−1)	1.2340	1.8392(−1)	8.5314(−1)
0.8	3.5242(−1)	1.7144	2.4886(−1)	1.2011	1.6928(−1)	7.9764(−1)
0.9	3.4941(−1)	1.7036	2.3677(−1)	1.1561	1.4960(−1)	7.2184(−1)
1.0	3.4412(−1)	1.6844	2.1575(−1)	1.0763	1.1583(−1)	5.8840(−1)

Acknowledgment. The authors would like to thank A. M. Yacout for his help with the computer systems used to perform the calculations reported in this work.

REFERENCES

[1] L. B. BARICHELLO, M. CAMARGO, P. RODRIGUES, AND C. E. SIEWERT, *Unified solutions to classical flow problems based on the BGK model*, Z. Angew. Math. Phys., 52 (2001), pp. 517–534.

[2] L. B. BARICHELLO AND C. E. SIEWERT, *A discrete-ordinates solution for a non-grey model with complete frequency redistribution*, J. Quant. Spectros. Radiat. Transfer, 62 (1999), pp. 665–675.

[3] C. CERCIGNANI, *The Boltzmann Equation and Its Applications*, Springer, New York, 1988.

[4] S. CHAPMAN AND T. G. COWLING, *The Mathematical Theory of Non-Uniform Gases*, Cambridge University Press, Cambridge, UK, 1952.

[5] J. H. FERZIGER AND H. G. KAPER, *Mathematical Theory of Transport Processes in Gases*, North–Holland, Amsterdam, 1972.

[6] R. D. M. GARCIA AND C. E. SIEWERT, *Some exact results basic to the linearized Boltzmann equations for a binary mixture of rigid spheres*, Z. Angew. Math. Phys., 57 (2006), pp. 999–1010.

[7] R. D. M. GARCIA AND C. E. SIEWERT, *The temperature-jump problem based on the linearized Boltzmann equation for a binary mixture of rigid spheres*, Eur. J. Mech. B Fluids, 26 (2007), pp. 132–153.

[8] R. D. M. Garcia and C. E. Siewert, *Some solutions (linear in the spatial variables) and generalized Chapman-Enskog functions basic to the linearized Boltzmann equations for a binary mixture of rigid spheres*, Z. Angew. Math. Phys., 58 (2007), pp. 262–288.

[9] R. D. M. Garcia and C. E. Siewert, *The viscous-slip, diffusion-slip, and thermal-creep problems for a binary mixture of rigid spheres described by the linearized Boltzmann equation*, Eur. J. Mech. B Fluids, to appear.

[10] R. D. M. Garcia and C. E. Siewert, *Heat transfer between parallel plates: An approach based on the linearized Boltzmann equation for a binary mixture of rigid-sphere gases*, Phys. Fluids, 19 (2007), 027102.

[11] R. D. M. Garcia and C. E. Siewert, *Particular solutions of the linearized Boltzmann equation for a binary mixture of rigid spheres*, Z. Angew. Math. Phys., submitted.

[12] R. D. M. Garcia, C. E. Siewert, and M. M. R. Williams, *A formulation of the linearized Boltzmann equations for a binary mixture of rigid spheres*, Eur. J. Mech. B Fluids, 24 (2005), pp. 614–620.

[13] S. Kosuge, K. Sato, S. Takata, and K. Aoki, *Flows of a binary mixture of rarefied gases between two parallel plates*, in Rarefied Gas Dynamics, M. Capitelli, ed., AIP, New York, 2005, pp. 150–155.

[14] S. K. Loyalka and K. A. Hickey, *Kinetic theory of thermal transpiration and the mechanocaloric effect: Planar flow of a rigid sphere gas with arbitrary accommodation at the surface*, J. Vac. Sci. Technol. A, 9 (1991), pp. 158–163.

[15] F. J. McCormack, *Construction of linearized kinetic models for gaseous mixtures and molecular gases*, Phys. Fluids, 16 (1973), pp. 2095–2105.

[16] S. Naris, D. Valougeorgis, F. Sharipov, and D. Kalempa, *Discrete velocity modelling of gaseous mixture flow in MEMS*, Superlattices Microstruct., 35 (2004), pp. 629–643.

[17] T. Ohwada, Y. Sone, and K. Aoki, *Numerical analysis of the Poiseuille and thermal transpiration flows between two parallel plates on the basis of the Boltzmann equation for hard-sphere molecules*, Phys. Fluids A, 1 (1989), pp. 2042–2049.

[18] C. L. Pekeris, *Solution of the Boltzmann-Hilbert integral equation*, Proc. Nat. Acad. Sci. U.S.A., 41 (1955), pp. 661–669.

[19] F. Sharipov and D. Kalempa, *Velocity slip and temperature jump coefficients for gaseous mixtures.* I. *Viscous slip coefficient*, Phys. Fluids, 15 (2003), pp. 1800–1806.

[20] F. Sharipov and V. Seleznev, *Data on internal rarefied gas flows*, J. Phys. Chem. Ref. Data, 27 (1998), pp. 657–706.

[21] C. E. Siewert, *The linearized Boltzmann equation: Concise and accurate solutions to basic flow problems*, Z. Angew. Math. Phys., 54 (2003), pp. 273–303.

[22] C. E. Siewert, R. D. M. Garcia, and P. Grandjean, *A concise and accurate solution for Poiseuille flow in a plane channel*, J. Math. Phys., 21 (1980), pp. 2760–2763.

[23] C. E. Siewert and D. Valougeorgis, *The McCormack model: Channel flow of a binary gas mixture driven by temperature, pressure and density gradients*, Eur. J. Mech. B Fluids, 23 (2004), pp. 645–664.

[24] M. M. R. Williams, *Mathematical Methods in Particle Transport Theory*, Butterworth, London, 1971.

[25] M. M. R. Williams, *A review of the rarefied gas dynamics theory associated with some classical problems in flow and heat transfer*, Z. Angew. Math. Phys., 52 (2001), pp. 500–516.

SIAM J. Appl. Math.
Vol. 67, No. 4, pp. 1064–1100

PATTERN SELECTION FOR FARADAY WAVES IN AN INCOMPRESSIBLE VISCOUS FLUID*

A. C. SKELDON† AND G. GUIDOBONI‡

Abstract. When a layer of fluid is oscillated up and down with a sufficiently large amplitude, patterns form on the surface, a phenomenon first observed by Faraday. A wide variety of such patterns have been observed from regular squares and hexagons to superlattice and quasipatterns and more exotic patterns such as oscillons. Previous work has investigated the mechanisms of pattern selection using the tools of symmetry and bifurcation theory. The hypotheses produced by these generic arguments have been tested against an equation derived by Zhang and Viñals in the weakly viscous and large depth limit. However, in contrast, many of the experiments use shallow viscous layers of fluid to counteract the presence of high frequency weakly damped modes that can make patterns hard to observe. Here we develop a weakly nonlinear analysis of the full Navier–Stokes equations for the two-frequency excitation Faraday experiment. The problem is formulated for general depth, although results are presented only for the infinite depth limit. We focus on a few particular cases where detailed experimental results exist and compare our analytical results with the experimental observations. Good agreement with the experimental results is found.

Key words. Faraday waves, superlattice patterns, weakly nonlinear analysis

AMS subject classification. 37N10

DOI. 10.1137/050639223

1. Introduction.

Waves on the surface of a fluid excited by a vertical oscillation were first observed by Faraday [1]. Subsequently, in the 1980's, the so-called Faraday crispation experiment became one of the first fluid experiments where mode interactions and chaos were observed [2]. Over the last decade, this experiment has become a testbed for ideas of pattern selection in systems under parametric excitation, and a large variety of patterns have been observed including not just regular patterns of squares and hexagons but many more exotic patterns such as superlattice patterns, quasipatterns, and oscillons. These more recent studies were initiated by the results of Edwards and Fauve [3], who used a two-frequency, rather than a single-frequency, excitation, thereby increasing the number of parameters in the problem and breaking the subharmonic time symmetry. Further two-frequency experiments have been performed by Kudrolli, Pier, and Gollub [4] and Arbell and Fineberg [5, 6, 7]. Subsequently, many of the patterns have been observed in experiments with only a single frequency of excitation [8]. Meanwhile, in practical applications of Faraday waves, the phenomenon has been investigated as a tool to produce patterns on films [9, 10], investigated as a mechanism for transporting gas across an air/water boundary [11], and seen as oscillations on the surface of bubbles [12].

In a container, if the amplitude of the vertical excitation is not too large, then no waves form on the surface of the fluid and the fluid is merely translated up and down. As the amplitude of the excitation is increased, waves appear at a critical amplitude of excitation. This onset of waves was first described theoretically by Benjamin and

*Received by the editors August 30, 2005; accepted for publication (in revised form) February 13, 2007; published electronically May 10, 2007.
http://www.siam.org/journals/siap/67-4/63922.html
†Department of Mathematics, University of Surrey, Guildford GU2 7XH, UK (a.skeldon@surrey.ac.uk).
‡Department of Mathematics, University of Houston, Houston, TX 77204-3008 (gio@math.uh.edu). This author was partially supported by a Marie Curie Training Fellowship.

Ursell [13], who showed that for an inviscid infinite layer of fluid the problem reduces to a Mathieu equation. Kumar and Tuckerman [14] developed a method for solving the linear stability problem for the viscous, finite depth fluid problem for a single frequency of excitation. This work was extended to two-frequency excitation [15] and gives excellent agreement with experimental measurements of the onset of patterns.

Understanding not just the onset of patterns but the type of patterns is challenging. The full mathematical description of the fluid problem involves the Navier–Stokes equations in a domain with a free surface, and the excitation makes the problem nonautonomous. Symmetry arguments along with the notion of resonant trial interactions have been used to uncover some of the pattern selection mechanisms [18]. This showed that weakly damped harmonic modes play a key role, with the wavenumber of the weakly damped mode relative to the critical wavenumber being an indicator of what patterns are likely to be seen. Since the wavenumbers of weakly damped modes are determined by the particular forcing function, this in turn has led to theoretical work in the nearly Hamiltonian limit on controlling pattern selection [19, 20]. In this work, they showed how multiple frequency components in the forcing can be used to enhance particular resonant triad interactions that in turn promote the stability of particular superlattice patterns. The theoretical ideas in [18, 19, 20] were all tested by calculating the coefficients of the relevant amplitude equations for a two-coupled scalar partial differential equation model derived and analyzed by Zhang and Viñals describing the Faraday problem in a weakly viscous and large depth limit [16, 17]. While the theory and the results calculated from the Zhang–Viñals equation agree well, it is harder to establish to what degree these pattern selection mechanisms can be used to explain experimental findings. This is because many of the experimental studies use a fluid that is either moderately viscous or a container that is shallow, neither of which is within the range of validity of the Zhang–Viñals model. The reason that experiments tend to focus on these cases is because of the presence of long wavelength modes that can make it difficult to observe regular patterns: these modes can be damped either by increasing the viscosity or by increasing the dissipation from the lower boundary by making the container shallower [21]. The large viscosity also minimizes the impact of the lateral boundaries on the patterns and the effect of patterns formed by meniscus waves emitted from the sidewalls.

Weakly nonlinear analysis from the full fluid equations for single-frequency excitation in an infinite fluid layer has been carried out by Chen and Viñals [22]. In this paper, we extend the formulation of the weakly nonlinear problem to two-frequency excitation and to finite fluid depth. The former involves a significantly different approach to the derivation of a solvability condition: for a single frequency of excitation, the form of the linear problem may be written as a recursion relation, and an adjoint to this recursion relation may be defined. This works because, in the linear Faraday problem, modes with different frequencies are coupled only through the excitation. Specifically, a frequency component $\cos\omega t$ in the excitation couples the nth Fourier mode to the $n-1$ and the $n+1$ mode. When an N mode truncation is taken, then the equation for the Nth mode is coupled only to the $(N-1)$th mode. Consequently, one can solve for the Nth mode in terms of the $(N-1)$th mode. In turn, this then allows one to solve for the $(N-1)$th mode and, recursively, for all modes. If instead multiple frequency forcing, for example, $\cos\chi\cos M_1\omega t+\sin\chi\cos M_2\omega t$, is used, then the nth Fourier mode is coupled to four other modes, $n-M_1$, $n-M_2$, $n+M_1$, $n+M_2$. Truncating at N modes leaves the Nth mode coupled to both $N-M_1$ and $N-M_2$, and so a recursion relation cannot readily be defined.

In the future, we will investigate the effect of depth on the coefficients of the

amplitude equations; however, for the purposes of this paper, we have focused on carrying out the calculations and detailed results for infinite depth only.

The pattern selection problem is further complicated by several issues. First, above onset not just a single wavenumber but a band of wavenumbers is unstable. Allowing for variation of the spatial scale to account for this typically leads to Ginzburg–Landau-type amplitude equations. Second, in the viscous Faraday problem, there are weakly damped long wavelength modes. These are coupled to the free surface deformation so that the larger the amplitude of the Faraday waves the more significant the effect. Both of these effects are discussed in [23] for Faraday waves in two space dimensions, but as yet there has been no attempt to include these effects in three space dimensions. Note that since the weakly damped long wavelength modes are coupled to the surface deformation, they do not effect the pattern selection at onset but could have an effect thereafter. Finally, at onset the wavenumber specifies the magnitude but not the direction of the associated wavevector. The spatial scale is therefore determined but not the particular pattern. Typically, a finite number of wavevectors are considered and amplitude equations derived for the amplitude associated with each wavevector. Two approaches are taken. In the first, an integer number of eigenvectors corresponding to modes that are equispaced around a circle are considered. Depending on the number of modes used, this leads to an amplitude equation describing squares, hexagons, or quasipatterns. The amplitude equations are of gradient form, and a Lyapunov function can be written down. The relative stability of the different patterns is then inferred from the relative value of the energy for the different states. A clear discussion of some of the issues involved in using amplitude equations to describe quasipatterns is given in [24]. Alternatively, eigenvectors that generate different spatially periodic lattices are considered. Amplitude equations may again be derived, but this time the eigenvalues indicating the relative stability for different patterns that are supported by the same lattice are considered. The methods are closely related, as discussed further in sections 4 and 6. In their single-frequency study, Chen and Vinãls [22] focus on squares, hexagons, and quasipatterns. Here in our two-frequency approach we, at least initially, consider spatially periodic patterns on a lattice. We apply our results to the particular two-frequency experimental results of Kudrolli, Pier, and Gollub [4] and find good agreement with their observations.

The layout of this paper is as follows. In section 2, we set up the mathematical problem. In section 3, a weakly nonlinear expansion about the critical wavenumber is carried out and the weakly nonlinear equations at each order derived. In section 4, we briefly discuss the pattern formation context within which we work and specify the general form of the solutions in the horizontal direction. This leads to a sequence of problems for the surface height and the vertical dependence of the velocity. These equations are solved in section 5, leading to the evaluation of the coefficients for the amplitude equations describing the weakly nonlinear pattern formation. The problem contains a number of physical parameters, and the coefficients are calculated for a range of values relevant to the experimental results in [4]. The calculations in section 5 are performed only in the case of infinite depth, although all early sections are not restricted in this way. The justification for this and the implications of the values of the coefficients for the pattern selection are discussed in section 6. Our conclusions are drawn in section 7.

2. Mathematical model. We consider an infinite horizontal layer of viscous incompressible fluid of finite depth that is subjected to gravity g and to a vertical periodic acceleration of amplitude a. At the lower boundary the fluid is in contact

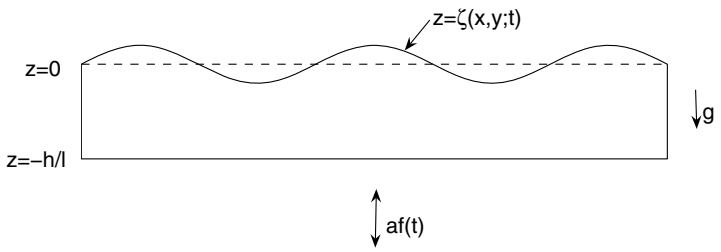


FIG. 2.1. *Sketch of a cross-section through the layer of fluid.*

with a rigid plane, while at the upper boundary the surface is open to the external ambient conditions. This means that the upper surface is a free boundary whose shape and evolution is an unknown of the problem.

We consider a frame of reference which is moving with the periodic excitation whose z-axis is perpendicular to the rigid plane at the bottom at $z = -h/l$, where h/l is the nondimensional depth of the layer. A sketch of the geometry is shown in Figure 2.1. We suppose the free surface is regular enough to be written in the Cartesian representation $z = \zeta(x, y; t)$; then the fluid motion is described by the dimensionless Navier–Stokes equations

$$\begin{aligned} \nabla \cdot \mathbf{u} &= 0, \\ \partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} &= -\nabla \mathcal{P} + C\Delta \mathbf{u} - (1 + af(t))\mathbf{e}_3, \end{aligned} \tag{2.1}$$

where $\mathbf{u} = (u, v, w)$ is the velocity field, $\mathcal{P}$ the pressure, and

$$f(t) \equiv f_1(t) = \cos(\omega t) \tag{2.2}$$

for single-frequency excitation and

$$f(t) = f_2(t) = \cos(\chi)\cos(M_1\omega t) + \sin(\chi)\cos(M_2\omega t + \phi) \tag{2.3}$$

for two-frequency excitation, where M_1 and M_2 are integers and χ and ϕ are real. The units of length, time, velocity, and pressure have been taken as l, $\sqrt{l/g}$, $\sqrt{gl}$, and ϱgl, respectively. The amplitude of the acceleration due to the excitation, a, is measured in units of g. Here l is taken as k_c^{-1}, where k_c is the wavenumber of the pattern at onset. The parameter $C = \nu/(gl^3)^{1/2}$ is the square of the inverse of the Galileo number, where ν is the kinematic viscosity of the fluid. We have used the notation $\nabla = (\nabla_H, \partial_z)$, with $\nabla_H = (\partial_x, \partial_y)$. Equations (2.1) apply in a domain $\Omega = \Sigma \times (-h/l, \zeta(x, y; t))$, where Σ is the horizontal periodicity cell. The bottom of the container, at $z = -h/l$, is rigid, and therefore we take no-slip boundary conditions here:

$$u = v = w = 0. \tag{2.4}$$

At the free surface $z = \zeta(x, y; t)$ we have the kinematic condition, which says that the surface is advected by the fluid, and two further conditions, one for the balance of the tangential stresses and one for the balance of normal stresses. This leads to three conditions at $z = \zeta(x, y; t)$, namely

$$\begin{aligned} \partial_t \zeta + u\partial_x \zeta + v\partial_y \zeta &= w, \\ \mathbf{t}_1 \cdot \mathbf{T}\mathbf{n} = \mathbf{t}_2 \cdot \mathbf{T}\mathbf{n} &= 0, \\ -\mathcal{P} + 2C\mathbf{n}\mathbf{D}(\mathbf{u})\mathbf{n} &= B\mathcal{H} - p_e, \end{aligned} \tag{2.5}$$

where $\mathbf{T} = -\mathcal{P}\mathbf{I} + 2C\mathbf{D}(\mathbf{u})$ is the stress tensor, $\mathbf{D}(\mathbf{u}) = (\nabla\mathbf{u} + \nabla^T\mathbf{u})/2$ is the rate-of-strain tensor, $\mathcal{H} = \nabla_H \cdot (\nabla_H\zeta/\sqrt{1+|\nabla_H\zeta|^2})$ is the double mean curvature, and the unit normal and tangent vectors are defined as

$$\mathbf{n}(x,y;t) = \left(-\frac{\partial_x\zeta}{\sqrt{1+|\nabla_H\zeta|^2}}, -\frac{\partial_y\zeta}{\sqrt{1+|\nabla_H\zeta|^2}}, \frac{1}{\sqrt{1+|\nabla_H\zeta|^2}}\right),$$
$$\mathbf{t}_1(x,y;t) = \left(\frac{1}{\sqrt{1+|\partial_x\zeta|^2}}, 0, \frac{\partial_x\zeta}{\sqrt{1+|\partial_x\zeta|^2}}\right),$$
$$\mathbf{t}_2(x,y;t) = \left(0, \frac{1}{\sqrt{1+|\partial_y\zeta|^2}}, \frac{\partial_y\zeta}{\sqrt{1+|\partial_y\zeta|^2}}\right).$$

Here p_e is the dimensionless pressure of the external ambient fluid and is assumed known. The parameter $B = \sigma/\varrho g l^2$, where σ is the surface tension and ϱ the density of the fluid, is the inverse Bond number and is a nondimensional measure of the relative importance of surface tension and gravity.

It is convenient to define a new pressure,

$$p = \mathcal{P} + (1 + af(t))z, \tag{2.6}$$

and this has the effect of shifting the acceleration term from the momentum equation to the normal stress condition. In addition, we eliminate the pressure from the momentum equation by taking $-(\nabla\times\nabla\times)$. Using the relation $\nabla\times\nabla\times\mathbf{u} = \nabla(\nabla\cdot\mathbf{u}) - \Delta\mathbf{u}$ and the fact that $\nabla\cdot\mathbf{u} = 0$, the problem then becomes

$$\begin{aligned} \nabla\cdot\mathbf{u} &= 0, \\ \partial_t\Delta\mathbf{u} - C\Delta\Delta\mathbf{u} &= \nabla\times\nabla\times(\mathbf{u}\cdot\nabla\mathbf{u}), \end{aligned} \tag{2.7}$$

with boundary conditions on $z = -h/l$,

$$u = v = w = 0, \tag{2.8}$$

and on $z = \zeta$,

$$\begin{aligned} \partial_t\zeta + u\partial_x\zeta + v\partial_y\zeta &= w, \\ \mathbf{t}_1\cdot\mathbf{T}\mathbf{n} = \mathbf{t}_2\cdot\mathbf{T}\mathbf{n} &= 0, \\ 2C\mathbf{n}\mathbf{D}(\mathbf{u})\mathbf{n} &= B\mathcal{H} + p - p_e - (1+af(t))\zeta. \end{aligned} \tag{2.9}$$

Equations (2.7) with boundary conditions (2.8) and (2.9) have a trivial solution,

$$\mathbf{u} = \mathbf{0}, \quad p = p_e, \quad \zeta = 0. \tag{2.10}$$

This solution corresponds to a flat-surface state where there is no relative motion of the fluid with respect to the moving frame.

3. Weakly nonlinear analysis. The flat-surface state loses stability at a critical amplitude of the excitation frequency to regular patterns of standing waves. We use a multiple timescale approach to derive equations describing the amplitude of these standing waves near threshold. In order to do this, the governing equations and the boundary conditions are expanded in a power series of the dimensionless distance

away from the threshold, ε, and solved order by order in ε. So for the driving dimensionless amplitude a, we let $a = a_0 + \varepsilon a_1 + \varepsilon^2 a_2$ and expand the flow variables as

$$\begin{aligned}
\mathbf{u} &= \varepsilon \mathbf{u}_1 + \varepsilon^2 \mathbf{u}_2 + \varepsilon^3 \mathbf{u}_3 + \cdots, \\
p &= p_e + \varepsilon p_1 + \varepsilon^2 p_2 + \varepsilon^3 p_3 + \cdots, \\
\zeta &= \varepsilon \zeta_1 + \varepsilon^2 \zeta_2 + \varepsilon^3 \zeta_3 + \cdots.
\end{aligned}$$

At each order in ε the solution is defined in a different domain since each ζ_i is different. In order to overcome this difficulty, Chen and Vinãls [22] take a Taylor expansion of the boundary conditions at the free surface around the flat surface state $z = 0$, so that they consider the solution in $\Sigma \times [-h/l, 0]$ at each order. We follow the same approach here. Near threshold, $\varepsilon \ll 1$, we separate fast and slow timescales: $t = \tau + T_1/\varepsilon + T_2/\varepsilon^2$ such that $\partial_t = \partial_\tau + \varepsilon \partial_{T_1} + \varepsilon^2 \partial_{T_2}$. The fast timescale is the timescale of the excitation, while the slower timescales describe the evolution of the amplitude of the patterns over many periods of the excitation. In sections 3.1, 3.2, and 3.3, we list the problem for each of the first three orders in ε. These agree with those used in the computations of [22], although note that there is a typographical error in their paper for the normal stress boundary condition at third order. In section 3.4, we derive the linear adjoint problem that is needed in order to find the solvability conditions that lead to the amplitude equations. The general form for the solvability conditions themselves are given in section 3.5. As found in [14] for the linear problem, the linear operator on the left-hand side of the hierarchy of problems for different ε depends only on the vertical velocity w and on the height of surface ζ. The horizontal components of the velocity, u and v, and the pressure, p, are needed to evaluate the nonlinear terms that appear on the right-hand side. These may be computed from w and ζ: details are given in the appendices.

3.1. Linear problem (first order problem).

$$\partial_\tau \Delta w_1 - C \Delta \Delta w_1 = 0, \tag{3.1}$$

with boundary conditions on $z = -h/l$,

$$w_1 = \partial_z w_1 = 0, \tag{3.2}$$

and on $z = 0$,

$$\begin{aligned}
\partial_\tau \zeta_1 - w_1 &= 0, \\
\Delta_H w_1 - \partial_z^2 w_1 &= 0, \\
-\partial_\tau \partial_z w_1 + C \partial_z^3 w_1 + 3C \Delta_H \partial_z w_1 & \\
-B \Delta_H \Delta_H \zeta_1 + (1 + a_0 f(\tau)) \Delta_H \zeta_1 &= 0.
\end{aligned}$$

Here $\Delta_H = \partial_x^2 + \partial_y^2$.

3.2. Second order problem.

$$\partial_\tau \Delta w_2 - C \Delta \Delta w_2 = N_{eq}^{(2)}, \tag{3.3}$$

with boundary conditions on $z = -h/l$,

$$w_2 = \partial_z w_2 = 0, \tag{3.4}$$

and on $z = 0$,

$$\begin{aligned}\partial_\tau \zeta_2 - w_2 &= N_{kc}^{(2)},\\ \Delta_H w_2 - \partial_z^2 w_2 &= N_{ts}^{(2)},\\ -\partial_\tau \partial_z w_2 + C\partial_z^3 w_2 + 3C\Delta_H \partial_z w_2 &\\ -B\Delta_H \Delta_H \zeta_2 + (1 + a_0 f(\tau))\Delta_H \zeta_2 &= N_{ns}^{(2)}.\end{aligned}$$

Here

$$\begin{aligned}N_{eq}^{(2)} &= [\nabla \times \nabla \times (\mathbf{u}_1 \cdot \nabla)\mathbf{u}_1] \cdot \mathbf{e}_3 - \partial_{T_1} \Delta w_1,\\ N_{kc}^{(2)} &= -\partial_{T_1}\zeta_1 - u_1\partial_x\zeta_1 - v_1\partial_y\zeta_1 + \partial_z w_1 \zeta_1,\\ N_{ts}^{(2)} &= \partial_x\Big[-\partial_{zz}u_1\zeta_1 - \partial_{xz}w_1\zeta_1 + 2(\partial_x u_1 - \partial_z w_1)\partial_x\zeta_1 + (\partial_y u_1 + \partial_x v_1)\partial_y\zeta_1\Big]\\ &\quad +\partial_y\Big[-\partial_{zz}v_1\zeta_1 - \partial_{yz}w_1\zeta_1 + 2(\partial_y v_1 - \partial_z w_1)\partial_y\zeta_1 + (\partial_y u_1 + \partial_x v_1)\partial_x\zeta_1\Big],\\ N_{ns}^{(2)} &= \partial_{T_1}\partial_z w_1 - \nabla_H \cdot (\mathbf{u}_1 \cdot \nabla)\mathbf{u}_1 + \Delta_H(-2C\partial_{zz}w_1\zeta_1 + \partial_z p_1 \zeta_1) - a_1 f(\tau)\Delta_H\zeta_1.\end{aligned}$$

3.3. Third order problem.

$$\partial_\tau \Delta w_3 - C\Delta\Delta w_3 = N_{eq}^{(3)}, \tag{3.5}$$

with boundary conditions on $z = -h/l$,

$$w_3 = \partial_z w_3 = 0, \tag{3.6}$$

and on $z = 0$,

$$\begin{aligned}\partial_\tau \zeta_3 - w_3 &= N_{kc}^{(3)},\\ \Delta_H w_3 - \partial_z^2 w_3 &= N_{ts}^{(3)},\\ -\partial_\tau \partial_z w_3 + C\partial_z^3 w_3 + 3C\Delta_H \partial_z w_3 &\\ -B\Delta_H \Delta_H \zeta_3 + (1 + a_0 f(\tau))\Delta_H \zeta_3 &= N_{ns}^{(3)}.\end{aligned}$$

Here

$$\begin{aligned}N_{eq}^{(3)} &= [\nabla \times \nabla \times (\mathbf{u}_1 \cdot \nabla)\mathbf{u}_2] \cdot \mathbf{e}_3 + [\nabla \times \nabla \times (\mathbf{u}_2 \cdot \nabla)\mathbf{u}_1] \cdot \mathbf{e}_3 - \partial_{T_2}\Delta w_1 - \partial_{T_1}\Delta w_2,\\ N_{kc}^{(3)} &= -\partial_{T_2}\zeta_1 - \partial_{T_1}\zeta_2 + \partial_z w_1 \zeta_2 + \partial_z w_2 \zeta_1 + \frac{1}{2}\partial_{zz} w_1 \zeta_1^2\\ &\quad -u_1\partial_x\zeta_2 - u_2\partial_x\zeta_1 - \partial_z u_1 \zeta_1 \partial_x \zeta_1 - v_1\partial_y\zeta_2 - v_2\partial_y\zeta_1 - \partial_z v_1 \zeta_1 \partial_y \zeta_1,\\ N_{ts}^{(3)} &= \partial_x\Big[-\partial_{zz}u_2\zeta_1 - \partial_{zz}u_1\zeta_2 - \frac{1}{2}\partial_{zzz}u_1\zeta_1^2 - \partial_{xz}w_2\zeta_1 - \partial_{xz}w_1\zeta_2 - \frac{1}{2}\partial_{xzz}w_1\zeta_1^2\\ &\quad -2(\partial_z w_2 - \partial_x u_2)\partial_x\zeta_1 - 2(\partial_z w_1 - \partial_x u_1)\partial_x\zeta_2 - 2\partial_z(\partial_z w_1 - \partial_x u_1)\zeta_1\partial_x\zeta_1\\ &\quad +(\partial_y u_2 + \partial_x v_2)\partial_y\zeta_1 + (\partial_y u_1 + \partial_x v_1)\partial_y\zeta_2 + \partial_z(\partial_y u_1 + \partial_x v_1)\zeta_1\partial_y\zeta_1\Big]\\ &\quad +\partial_y\Big[-\partial_{zz}v_2\zeta_1 - \partial_{zz}v_1\zeta_2 - \frac{1}{2}\partial_{zzz}v_1\zeta_1^2 - \partial_{yz}w_2\zeta_1 - \partial_{yz}w_1\zeta_2 - \frac{1}{2}\partial_{yzz}w_1\zeta_1^2\\ &\quad -2(\partial_z w_2 - \partial_y v_2)\partial_y\zeta_1 - 2(\partial_z w_1 - \partial_y v_1)\partial_y\zeta_2 - 2\partial_z(\partial_z w_1 - \partial_y v_1)\zeta_1\partial_y\zeta_1\\ &\quad +(\partial_y u_2 + \partial_x v_2)\partial_x\zeta_1 + (\partial_y u_1 + \partial_x v_1)\partial_x\zeta_2 + \partial_z(\partial_y u_1 + \partial_x v_1)\zeta_1\partial_x\zeta_1\Big],\end{aligned}$$

$$
\begin{aligned}
N_{ns}^{(3)} = {} & \partial_{T_2}\partial_z w_1 + \partial_{T_1}\partial_z w_2 - a_2 f(\tau)\Delta_H \zeta_1 - a_1 f(\tau)\Delta_H \zeta_2 - \nabla_H \cdot [\mathbf{u}_1 \cdot \nabla \mathbf{u}_2 + \mathbf{u}_2 \cdot \nabla \mathbf{u}_1] \\
& + \Delta_H \Big[\partial_z p_2 \zeta_1 + \partial_z p_1 \zeta_2 + \frac{1}{2}\partial_{zz} p_1 \zeta_1^2 - 2C\partial_{zz} w_1 \zeta_2 - 2C\partial_{zz} w_2 \zeta_1 - C\partial_{zzz} w_1 \zeta_1^2 \\
& + 2C(\partial_z u_2 + \partial_x w_2)\partial_x \zeta_1 + 2C(\partial_z w_1 - \partial_x u_1)(\partial_x \zeta_1)^2 + 2C(\partial_z v_2 + \partial_y w_2)\partial_y \zeta_1 \\
& + 2C\partial_z(\partial_x w_1 + \partial_z u_1)\partial_x \zeta_1 \zeta_1 + 2C\partial_z(\partial_y w_1 + \partial_z v_1)\partial_y \zeta_1 \zeta_1 \\
& + 2C(\partial_z w_1 - \partial_y v_1)(\partial_y \zeta_1)^2 - 2C(\partial_y u_1 + \partial_x v_1)\partial_x \zeta_1 \partial_y \zeta_1 - \frac{3}{2}B\partial_{xx}\zeta_1(\partial_x \zeta_1)^2 \\
& - \frac{3}{2}B\partial_{yy}\zeta_1(\partial_y \zeta_1)^2 - \frac{1}{2}B\partial_{xx}\zeta_1(\partial_y \zeta_1)^2 - \frac{1}{2}B\partial_{yy}\zeta_1(\partial_x \zeta_1)^2 - 2B\partial_x \zeta_1 \partial_y \zeta_1 \partial_{xy}\zeta_1 \Big].
\end{aligned}
$$

3.4. Linear adjoint problem. In order to use the Fredholm alternative and derive a solvability condition, the solution to the linear adjoint problem is needed. We suppose that $S_1 = (w_1, \zeta_1)$ is the solution of the linear problem and denote by $S^* = (w^*, \zeta^*)$ the solution of the linear adjoint problem. Then S_1 and S^* satisfy

$$
(S^*, \mathcal{L}S_1) = 0 = (\mathcal{L}^* S^*, S_1), \tag{3.7}
$$

where $\mathcal{L}$ and $\mathcal{L}^*$ are the linear and the linear adjoint operators, respectively, and $(\cdot, \cdot)$ means the following scalar product:

$$
\int_0^{2\pi/\omega} \int_\Omega w^* \Big(\partial_\tau \Delta w_1 - C\Delta\Delta w_1 \Big) d\Omega d\tau + \int_0^{2\pi/\omega} \int_\Sigma \zeta_* \Big[\partial_\tau \zeta_1 - w_1 \Big]_{z=0} d\Sigma d\tau = 0, \tag{3.8}
$$

where $\Omega = \Sigma \times (-h/l, 0)$ and Σ is the horizontal periodicity cell.

3.5. Solvability conditions. From the Fredholm alternative theorem it follows that at second order the solvability condition takes the form of

$$
\begin{aligned}
& \int_0^{2\pi/\omega} \int_\Omega w^* (\partial_\tau \Delta w_2 - C\Delta\Delta w_2 - N_{eq}^{(2)}) d\Omega d\tau \\
& \quad + \int_0^{2\pi/\omega} \int_\Sigma [\zeta^* (\partial_\tau \zeta_2 - w_2 - N_{kc}^{(2)})]_{z=0} d\Sigma d\tau = 0.
\end{aligned} \tag{3.9}
$$

This implies that

$$
\begin{aligned}
& \int_0^{2\pi/\omega} \int_\Omega w^* N_{eq}^{(2)} d\Omega d\tau + \int_0^{2\pi/\omega} \int_\Sigma [\zeta^* N_{kc}^{(2)}]_{z=0} d\Sigma d\tau \\
& + \int_0^{2\pi/\omega} \int_\Sigma [w^* N_{ns}^{(2)}]_{z=0} d\Sigma d\tau + C \int_0^{2\pi/\omega} \int_\Sigma [\partial_z w^* N_{ts}^{(2)}]_{z=0} d\Sigma d\tau = 0.
\end{aligned}
$$

Similarly at third order, we have

$$\int_0^{2\pi/\omega}\int_\Omega w^* N_{eq}^{(3)} d\Omega d\tau + \int_0^{2\pi/\omega}\int_\Sigma [\zeta^* N_{kc}^{(3)}]_{z=0} d\Sigma d\tau$$
$$+ \int_0^{2\pi/\omega}\int_\Sigma [w^* N_{ns}^{(3)}]_{z=0} d\Sigma d\tau + C\int_0^{2\pi/\omega}\int_\Sigma [\partial_z w^* N_{ts}^{(3)}]_{z=0} d\Sigma d\tau = 0.$$

4. Patterns. In order to proceed further, we need to first solve the linear problem set out in section 3.1. However, in an unbounded horizontal domain, while the linear problem predicts the onset of spatially periodic patterns at a given excitation frequency and excitation amplitude with a wavenumber k_c, it does not uniquely determine the pattern that is produced. This is related to the fact that, in an unbounded horizontal domain, the Faraday problem is isotropic so that no particular direction is preferred: any wavevector with wavenumber k_c would give an allowable solution; for example, stripes with any orientation would be possible. Furthermore, within the linear problem, linear superposition of different wavevectors with the critical wavenumber also give solutions. In this way, solutions such as squares, hexagons, superlattice patterns, and quasipatterns may be constructed by adding together stripe solutions of the appropriate orientation. However, the fact that these are solutions to the linear problem does not guarantee their existence or stability for the nonlinear problem. Indeed, only particular combinations of patterns are observed in experiments. Here we consider patterns that are spatially periodic, and this is implicit in our choice of domain in sections 2 and 3.

For patterns that are spatially periodic in two space dimensions, previous work has used equivariant bifurcation theory to find the generic types of solutions that exist [25], the generic amplitude equations that these patterns satisfy, and the stability of each pattern in terms of the coefficients of these amplitude equations [26]. For example, on the family of lattices with hexagonal symmetry, the generic amplitude equations are

$$\begin{aligned}
\dot{z}_1 &= \lambda z_1 + \epsilon \bar{z}_2 \bar{z}_3 \\
&\quad + (b_1|z_1|^2 + b_2|z_2|^2 + b_2|z_3|^2 + b_4|z_4|^2 + b_5|z_5|^2 + b_6|z_6|^2) z_1 + \mathcal{O}(|\mathbf{z}|^4), \\
\dot{z}_2 &= \lambda z_2 + \epsilon \bar{z}_3 \bar{z}_1 \\
&\quad + (b_2|z_1|^2 + b_1|z_2|^2 + b_2|z_3|^2 + b_6|z_4|^2 + b_4|z_5|^2 + b_5|z_6|^2) z_2 + \mathcal{O}(|\mathbf{z}|^4), \\
\dot{z}_3 &= \lambda z_3 + \epsilon \bar{z}_1 \bar{z}_2 \\
&\quad + (b_2|z_1|^2 + b_2|z_2|^2 + b_1|z_3|^2 + b_5|z_4|^2 + b_6|z_5|^2 + b_4|z_6|^2) z_3 + \mathcal{O}(|\mathbf{z}|^4), \\
\dot{z}_4 &= \lambda z_4 + \epsilon \bar{z}_6 \bar{z}_5 \\
&\quad + (b_4|z_1|^2 + b_6|z_2|^2 + b_5|z_3|^2 + b_1|z_4|^2 + b_2|z_5|^2 + b_2|z_6|^2) z_4 + \mathcal{O}(|\mathbf{z}|^4), \\
\dot{z}_5 &= \lambda z_5 + \epsilon \bar{z}_4 \bar{z}_6 \\
&\quad + (b_5|z_1|^2 + b_4|z_2|^2 + b_6|z_3|^2 + b_2|z_4|^2 + b_1|z_5|^2 + b_2|z_6|^2) z_5 + \mathcal{O}(|\mathbf{z}|^4), \\
\dot{z}_6 &= \lambda z_6 + \epsilon \bar{z}_5 \bar{z}_4 \\
&\quad + (b_6|z_1|^2 + b_5|z_2|^2 + b_4|z_3|^2 + b_2|z_4|^2 + b_2|z_5|^2 + b_1|z_6|^2) z_6 + \mathcal{O}(|\mathbf{z}|^4),
\end{aligned} \tag{4.1}$$

where the z_i are complex amplitudes and λ, ϵ, and b_i are real. An example of one of the family of such lattices is shown in Figure 4.1, where $\mathbf{K_{i_h}}$ is the mode with amplitude z_i. Different lattices correspond to different choices for θ. In terms of the Faraday problem considered in this paper, these equations arise by representing the horizontal spatial dependence of the linear problem for the surface height ζ_1 and the

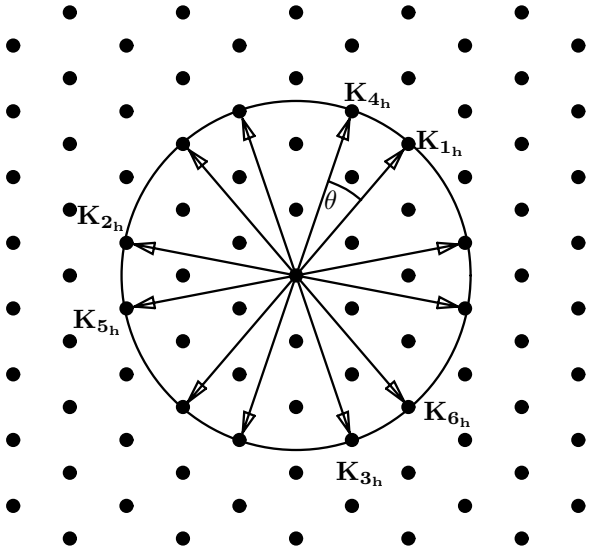


FIG. 4.1. *Hexagonal lattice generated by* 12 *wavevectors on the critical circle.*

vertical velocity w_1 as a sum of six modes where the ith mode has amplitude z_i and wavevector K_{i_h}.

For each hexagonal lattice, the equivariant branching lemma gives six patterns that bifurcate from the trivial state, and these are listed in Table 4.1 along with their branching equations and stability assignments. A further pattern has been found to exist and bifurcate from the trivial solution as discussed in [27]. We have not included this in Table 4.1 since its eigenvalues are indistinguishable from those for superhexagons at cubic order. We refer to both the superhexagons and the patterns discussed in [27] as superlattice patterns.

Similar results exist for families of square lattices.

Previously, the coefficients of the amplitude equations have been calculated for a long wavelength scalar partial differential equation describing a range of convection problems [28], for Turing patterns [29], and more recently for the Zhang–Viñals model of the Faraday problem [18]. These calculations allow inferences on the relative stability of different spatially periodic patterns to be made.

The cubic truncation of the amplitude equations (4.1) can be written in gradient form:

$$\dot{z}_i = -\frac{\partial \mathcal{F}}{\partial \bar{z}_i}, \tag{4.2}$$

where the Lyapunov function is given by

$$\begin{aligned}\mathcal{F} = &-\sum_{i=1..6}\left[\lambda|z_i|^2 - \frac{1}{2}b_1|z_i|^4\right]\\ &-\epsilon(z_1z_2z_3 + z_4z_5z_6 + \bar{z}_1\bar{z}_2\bar{z}_3 + \bar{z}_4\bar{z}_5\bar{z}_6)\\ &-b_2\left(|z_1|^2|z_2|^2 + |z_1|^2|z_3|^2 + |z_4|^2|z_5|^2 + |z_4|^2|z_6|^2 + |z_5|^2|z_6|^2\right)\\ &-b_4\left(|z_1|^2|z_4|^2 + |z_2|^2|z_5|^2 + |z_3|^2|z_6|^2\right) - b_5\left(|z_1|^2|z_5|^2 + |z_2|^2|z_6|^2 + |z_3|^2|z_4|^2\right)\\ &-b_6\left(|z_1|^2|z_6|^2 + |z_2|^2|z_4|^2 + |z_3|^2|z_5|^2\right).\end{aligned} \tag{4.3}$$

TABLE 4.1

Branching equations and signs of eigenvalues for primary bifurcation branches on the hexagonal lattice; $\epsilon, b_1, \ldots, b_6$ are coefficients in the bifurcation equation (4.1).

Branching equation	Signs of nonzero eigenvalues
Stripes (S) $\mathbf{z} = (A_S, 0, 0, 0, 0, 0)$ $0 = \lambda A_S + b_1 A_S^3 + \mathcal{O}(A_S^5)$	$\mathrm{sgn}(b_1)$, $\mathrm{sgn}(\tilde{\epsilon} A_S + (b_2 - b_1) A_S^2)$, $\mathrm{sgn}(-\tilde{\epsilon} A_S + (b_2 - b_1) A_S^2)$, $\mathrm{sgn}(b_4 - b_1)$, $\mathrm{sgn}(b_5 - b_1)$, $\mathrm{sgn}(b_6 - b_1)$.
Simple hexagons ($\mathrm{H}^{\pm}$) $\mathbf{z} = (A_H, A_H, A_H, 0, 0, 0)$ $0 = \lambda A_H + \tilde{\epsilon} A_H^2$ $+(b_1 + 2b_2) A_H^3 + \mathcal{O}(A_H^4)$	$\mathrm{sgn}(\tilde{\epsilon} A_H + 2(b_1 + 2b_2) A_H^2)$, $\mathrm{sgn}(-\tilde{\epsilon} A_H + (b_1 - b_2) A_H^2)$, $\mathrm{sgn}(-\tilde{\epsilon} A_H + (b_4 + b_5 + b_6 - b_1 - 2b_2) A_H^2)$, $\mathrm{sgn}(-\tilde{\epsilon} A_H + \mathcal{O}(A_H^3))$.
Rectangles ($\mathrm{R}_{h1,m,n}$) $\mathbf{z} = (A_R, 0, 0, A_R, 0, 0)$ $0 = \lambda A_R + (b_1 + b_4) A_R^3 + \mathcal{O}(A_R^5)$	$\mathrm{sgn}(b_1 + b_4)$, $\mathrm{sgn}(b_1 - b_4)$, $\mathrm{sgn}(\mu_1)$, $\mathrm{sgn}(\mu_2)$, where $\mu_1 + \mu_2 = (-2b_1 - 2b_4 + 2b_2 + b_5 + b_6) A_R^2$, $\mu_1 \mu_2 = -\tilde{\epsilon}^2 A_R^2 + (b_1 + b_4 - b_2 - b_5)$ $(b_1 + b_4 - b_2 - b_6) A_R^4$.
Rectangles ($\mathrm{R}_{h2,m,n}$) $\mathbf{z} = (A_R, 0, 0, 0, A_R, 0)$ $0 = \lambda A_R + (b_1 + b_5) A_R^3 + \mathcal{O}(A_R^5)$	$\mathrm{sgn}(b_1 + b_5)$, $\mathrm{sgn}(b_1 - b_5)$, $\mathrm{sgn}(\mu_1)$, $\mathrm{sgn}(\mu_2)$, where $\mu_1 + \mu_2 = (-2b_1 - 2b_5 + 2b_2 + b_4 + b_6) A_R^2$, $\mu_1 \mu_2 = -\tilde{\epsilon}^2 A_R^2 + (b_1 + b_5 - b_2 - b_4)$ $(b_1 + b_5 - b_2 - b_6) A_R^4$.
Rectangles ($\mathrm{R}_{h3,m,n}$) $\mathbf{z} = (A_R, 0, 0, 0, 0, A_R)$ $0 = \lambda A_R + (b_1 + b_6) A_R^3 + \mathcal{O}(A_R^5)$	$\mathrm{sgn}(b_1 + b_6)$, $\mathrm{sgn}(b_1 - b_6)$, $\mathrm{sgn}(\mu_1)$, $\mathrm{sgn}(\mu_2)$, where $\mu_1 + \mu_2 = (-2b_1 - 2b_6 + 2b_2 + b_4 + b_5) A_R^2$, $\mu_1 \mu_2 = -\tilde{\epsilon}^2 A_R^2 + (b_1 + b_6 - b_2 - b_4)$ $(b_1 + b_6 - b_2 - b_5) A_R^4$.
Superhexagons ($\mathrm{SH}^{\pm}_{m,n}$) $\mathbf{z} = (A_{SH}, A_{SH}, A_{SH},$ $A_{SH}, A_{SH}, A_{SH})$ $0 = \lambda A_{SH} + \tilde{\epsilon} A_{SH}^2 + (b_1 + 2b_2) A_{SH}^3$ $+(b_4 + b_5 + b_6) A_{SH}^3 + \mathcal{O}(A_{SH}^4)$	$\mathrm{sgn}(\tilde{\epsilon} A_{SH} + 2(b_1 + 2b_2 + b_4 + b_5 + b_6) A_{SH}^2)$, $\mathrm{sgn}(\tilde{\epsilon} A_{SH} + 2(b_1 + 2b_2 - b_4 - b_5 - b_6) A_{SH}^2)$, $\mathrm{sgn}(-\tilde{\epsilon} A_{SH} + \mathcal{O}(A_{SH}^3))$, $\mathrm{sgn}(-\tilde{\epsilon} A_{SH} + \mathcal{O}(A_{SH}^3))$,* $\mathrm{sgn}(\mu_1)$, $\mathrm{sgn}(\mu_2)$, where $\mu_1 + \mu_2 = -4\tilde{\epsilon} A_{SH} + 4(b_1 - b_2) A_{SH}^2$, $\mu_1 \mu_2 = 4(\tilde{\epsilon} A_{SH} - (b_1 - b_2) A_{SH}^2)^2$ $-2((b_4 - b_5)^2 + (b_4 - b_6)^2$ $+(b_5 - b_6)^2)) A_{SH}^4$, $\mathrm{sgn}(\mu_0)$, where $\mu_0 = \mathcal{O}(A_{SH}^{2(m-1)})$.

*These two eigenvalues differ at $\mathcal{O}(A_{SH}^3)$.

The different planforms then correspond to minima of the Lyapunov functional, and an "energy" for each state may be computed. In Table 4.2, we list the different planforms and the corresponding value of the Lyapunov function (4.3). The information given by the eigenvalues of the amplitude equations given in Table 4.1 and that given by the energy of the different states as given in Table 4.2 is complementary. Below we calculate the coefficients for the amplitude equations (4.1) from the full Navier–Stokes equation formulation of the Faraday problem. We then calculate the eigenvalues to examine relative stability. For those states that are relatively stable, we calculate the value of the Lyapunov function to find which have the lowest energy.

The coefficients of the amplitude equations are found by focusing on three calculations: one for stripes, one for rectangular patterns, and one for hexagons; these correspond to considering the three subspaces, $\mathbf{z} = (A_S, 0, 0, 0, 0, 0)$, $\mathbf{z} = (A_R, 0, 0, A_R, 0, 0)$,

TABLE 4.2

Value of the Lyapunov function $\mathcal{F}$ for each of the primary bifurcation branches on the hexagonal lattice; $\epsilon, b_1, \ldots, b_6$ are coefficients in the bifurcation equation (4.1). Only one of the rectangular states has been included: the other two may be obtained by cyclic permutation.

Planform	$\mathcal{F}$
Stripes (S) $\mathbf{z} = (A_S, 0, 0, 0, 0, 0)$	$\frac{\lambda^2}{b_1}$
Simple hexagons ($H^{\pm}$) $\mathbf{z} = (A_H, A_H, A_H, 0, 0, 0)$	$-\left(3\lambda A_H^2 + 2\epsilon A_H^3 + \frac{3}{2}(b_1 + 2b_2)A_H^4\right)$, where $0 = \lambda + \epsilon A_H + (b_1 + 2b_2)A_H^2$
Rectangles ($R_{h1,m,n}$) $\mathbf{z} = (A_R, 0, 0, A_R, 0, 0)$	$\frac{\lambda^2}{(b_1+b_4)}$
Superhexagons ($SH^{\pm}_{m,n}$) $\mathbf{z} = (A_{SH}, A_{SH}, A_{SH}, A_{SH}, A_{SH}, A_{SH})$	$-\left(6\lambda A_{SH}^2 + 4\epsilon A_{SH}^3 + 3(b_1 + 2b_2 + b_4 + b_5 + b_6)A_{SH}^4\right)$, where $0 = \lambda + \tilde{\epsilon} A_{SH} + (b_1 + 2b_2 + b_4 + b_5 + b_6)A_{SH}^2$

and $\mathbf{z} = (A_H, A_H, A_H, 0, 0, 0)$, respectively. Focusing on a particular pattern means that we make an assumption about the particular form of the horizontal behavior of the fluid variables. We can use this to reformulate the weakly nonlinear analysis in section 3 that is in terms of functions of x, y, and z to a simpler set of problems for sets of functions that depend only on z. It is this reformulation of the problem that is carried out in this section for each of the stripes, rectangles, and hexagons. Since stripes and rectangles arise through a symmetry-breaking bifurcation, there are no quadratic terms in the amplitude equations for these patterns. A result of this is that the solvability condition at second order necessarily leads to $a_1 = 0$. This fact can be included from the beginning of the analysis, and then we need only scale on two timescales; that is, we let $t = \tau + T/\varepsilon^2$ so that $\partial_t = \partial_\tau + \varepsilon^2 \partial_T$. (The more general formulation on three timescales is needed for hexagons.)

For stripes, we consider a solution to the first order problem given in section 3.1 of the form

$$w_1(x, z, \tau, T) = A_S(T)(e^{ikx} + e^{-ikx}) \sum_n W_{1,n}(z) e^{i(n\omega+\alpha)\tau},$$
$$\zeta_1(x, \tau, T) = A_S(T)(e^{ikx} + e^{-ikx}) \sum_n Z_{1,n} e^{i(n\omega+\alpha)\tau},$$

and for rectangular patterns, we consider

$$w_1(x, y, z, \tau, T) = A_R(T)[e^{ikx} + e^{ik(cx+sy)} + c.c.] \sum_n W_{1,n}(z) e^{i(n\omega+\alpha)\tau},$$
$$\zeta_1(x, y, \tau, T) = A_R(T)[e^{ikx} + e^{ik(cx+sy)} + c.c.] \sum_n Z_{1,n} e^{i(n\omega+\alpha)\tau}, \tag{4.4}$$

where $s = \sin\theta$ and $c = \cos\theta$ and θ is the angle between the wavevectors that make up the rectangular pattern. Here k is the wavenumber of the pattern, and a Floquet expansion in the basic frequency ω has been used as in [14]. When $\alpha = 0$, the expansion gives a harmonic solution, and when $\alpha = \omega/2$, the expansion gives a subharmonic

solution. Here and below we sum from $n = -\infty$ to $n = +\infty$. Consequently, since both the w_1 and ζ_1 are real, we have in addition

$$\begin{aligned} W_{1,n} &= \bar{W}_{1,-n}, \quad \alpha = 0, \\ W_{1,n} &= \bar{W}_{1,-n-1}, \quad \alpha = \omega/2. \end{aligned} \tag{4.5}$$

Although we do not list them, there are analogous reality conditions for the velocity components and surface height at each order. Similar choices for the expansion are made for the behavior of the horizontal velocity components u_1 and v_1, and these are listed in Appendix A. Since the results for stripes can be obtained by setting $\theta = 0$ and careful consideration of some factors of two, in what follows we include stripes as a special case in our formulation of the problem for rectangles. In order to proceed, the general form (4.4) for the pattern is substituted into the first order problem given in section 3.1. The result is a homogeneous fourth order linear differential equation for the vertical dependence of the vertical velocity component $W_{1,n}(z)$ along with the appropriate boundary conditions at $z = 0$ and $z = -h/l$. This is given in section 4.1.

At second order, as for the first order problem, assuming that we are interested in particular patterns means that we know the form for the horizontal behavior of the fluid. Specifically, we take the general form of the second order solution for rectangles as

$$\begin{aligned} w_2(x,y,z,\tau,T) = {} & A_R^2(T)[e^{2ikx} + e^{2ik(cx+sy)} + c.c.] \sum_n W_{2,1,n}(z) e^{i(n\omega+2\alpha)\tau} \\ & + A_R^2(T)[e^{ik[(1+c)x+sy]} + c.c.] \sum_n W_{2,2,n}(z) e^{i(n\omega+2\alpha)\tau} \\ & + A_R^2(T)[e^{ik[(1-c)x-sy]} + c.c.] \sum_n W_{2,3,n}(z) e^{i(n\omega+2\alpha)\tau}, \\ \zeta_2(x,y,\tau,T) = {} & A_R^2(T)[e^{2ikx} + e^{2ik(cx+sy)} + c.c.] \sum_n Z_{2,1,n} e^{i(n\omega+2\alpha)\tau} \\ & + A_R^2(T)[e^{ik[(1+c)x+sy]} + c.c.] \sum_n Z_{2,2,n} e^{i(n\omega+2\alpha)\tau} \\ & + A_R^2(T)[e^{ik[(1-c)x-sy]} + c.c.] \sum_n Z_{2,3,n} e^{i(n\omega+2\alpha)\tau}. \end{aligned} \tag{4.6}$$

The forms that are taken for the velocity components u_2 and v_2 are given in Appendix A. The expressions for the velocity components are substituted into the equations and boundary conditions at second order given in section 3.2, and this leads to an inhomogeneous fourth order ordinary differential equation for $W_{2,i,n}(z)$ along with boundary conditions. These are given in section 4.2.

As discussed above, if we take $a_1 = 0$, then there is no solvability condition at second order for rectangles. However, at third order there is a solvability condition. In order to derive this, the general form for the adjoint problem is needed. In section 4.3, we give the formulation for the adjoint problem, derived from the adjoint problem given in (3.8) along with the assumption that patterns to the adjoint problem take the general form

$$\begin{aligned} w^*(x,z,\tau,T) &= A^*(T)(e^{ikx} + e^{-ikx}) \sum_n W_n^*(z) e^{i(n\omega+\alpha)\tau}, \\ \zeta^*(x,\tau,T) &= A^*(T)(e^{ikx} + e^{-ikx}) \sum_n Z_n^* e^{i(n\omega+\alpha)\tau}. \end{aligned}$$

This then allows us to formulate the solvability condition at third order in section 4.4. The result is an amplitude equation for rectangles whose coefficients may be determined from $W_{1,n}(z), Z_{1,n}, W_{2,i,n}(z), Z_{2,i,n}$ and their derivatives along with the adjoint eigenfunctions $W^*_{1,n}(z)$ and $Z^*_{1,n}$ and their derivatives.

In the case of hexagons, we consider a first order solution of the form

$$\begin{aligned} &w_1(x,y,z,\tau,T_1,T_2) \\ &\quad = A_H(T_1,T_2)[e^{ikx} + e^{ik(-x+\sqrt{3}y)/2} + e^{ik(-x-\sqrt{3}y)/2} + c.c.] \sum_n W_{1,n}(z) e^{i(n\omega+\alpha)\tau}, \\ &\zeta_1(x,y,z,\tau,T_1,T_2) \\ &\quad = A_H(T_1,T_2)[e^{ikx} + e^{ik(-x+\sqrt{3}y)/2} + e^{ik(-x-\sqrt{3}y)/2} + c.c.] \sum_n Z_{1,n} e^{i(n\omega+\alpha)\tau} \end{aligned} \tag{4.7}$$

with similar choices for u_1 and v_1 that are listed in Appendix A. This leads to the same first order problem as for stripes and rectangles, as given in section 4.1 below.

At second order hexagons differ from rectangles. Generically, in problems that have E(2) symmetry, hexagons arise in a transcritical bifurcation and are necessarily locally unstable. In the weakly nonlinear analysis, this appears as a quadratic amplitude equation that results from the solvability condition for the second order problem. Two cases of interest arise that can result in stable hexagonal solutions: first, when there is an extra symmetry in the problem that removes the quadratic term, and second, when the coefficient of the quadratic term is sufficiently small so that the quadratic terms may formally be included at cubic order [30]. The Faraday problem is an example of a system that has E(2) symmetry. As we shall see below, for some values of the parameter χ in the drive (2.3), the response is subharmonic, and for some values it is harmonic. When the response is subharmonic, then there is an extra time symmetry in the problem and there are no quadratic terms in the amplitude equations. When $\chi = 0$, the response is harmonic, but since there is no M_2 component in the drive, there is again an extra symmetry in the problem, and again there are no quadratic terms in the amplitude equations. However, as the parameter χ is increased from 0, this extra symmetry is broken, and there is a gradual increase from zero in the size of the coefficient of the quadratic term. There is therefore at least some range in parameter space where it is reasonable to include the quadratic terms at cubic order. The way we proceed with the hexagon calculation is therefore as follows. First, we consider three timescales and formulate the solvability condition at second order. This is done using the solvability condition given in (3.9) and the specific form for the hexagonal pattern at first order (4.7); the result is given in section 4.5. This gives us a quadratic amplitude equation and enables us to compute the size of the quadratic term.

Next, we make the assumption that the coefficient of the quadratic term is either zero or sufficiently small ($O(\varepsilon)$) so that we may formally include the terms at cubic order. We therefore set $a_1 = 0$, rescale the problem on two timescales, and formulate the problem for the solution at second order by taking as a general form for the second order problem

$$\begin{aligned} w_2(x,y,z,\tau,T) &= A_H^2[e^{2ikx} + e^{ik(-x+\sqrt{3}y)} + e^{ik(-x-\sqrt{3}y)} + c.c.] \sum_n W_{2,1,n}(z) e^{i(n\omega+2\alpha)\tau} \\ &\quad + A_H^2[e^{ik\sqrt{3}y} + e^{ik(3x+\sqrt{3}y)/2} + e^{ik(3x-\sqrt{3}y)/2} + c.c.] \sum_n W_{2,2,n}(z) e^{i(n\omega+2\alpha)\tau}, \end{aligned}$$

$$\zeta_2(x,y,\tau,T) = A_H^2[e^{2ikx} + e^{ik(-x+\sqrt{3}y)} + e^{ik(-x-\sqrt{3}y)} + c.c.]\sum_n Z_{2,1,n}e^{i(n\omega+2\alpha)\tau}$$

$$+A_H^2[e^{ik\sqrt{3}y} + e^{ik(3x+\sqrt{3}y)/2} + e^{ik(3x-\sqrt{3}y)/2} + c.c.]\sum_n Z_{2,2,n}e^{i(n\omega+2\alpha)\tau}. \tag{4.8}$$

Expressions for u_2 and v_2 are listed in Appendix A. Substitution of these expressions into the second order equation and boundary conditions given in section 3.2 leads to an inhomogeneous fourth order ordinary differential equation for $W_{2,i,n}$ along with boundary conditions. These are given in section 4.6. Finally, including the quadratic terms at cubic order, we formulate the solvability condition for the third order hexagonal problem in section 4.7.

4.1. The linear problem. The linear problem is the same for all periodic patterns and is given by

$$[i(n\omega+\alpha) - C(D^2-k^2)](D^2-k^2)W_{1,n}(z) = 0, \tag{4.9}$$

where D indicates the derivative with respect to z, with boundary conditions

$$W_{1,n} = DW_{1,n} = 0, \tag{4.10}$$

on $z = -h/l$, and on $z = 0$,

$$i(n\omega+\alpha)Z_{1,n} - W_{1,n} = 0, \tag{4.11}$$

$$(D^2+k^2)W_{1,n} = 0, \tag{4.12}$$

$$\Big(i(n\omega+\alpha) + 3Ck^2\Big)DW_{1,n} - CD^3W_{1,n} + k^2(Bk^2+1)Z_{1,n} = -\frac{1}{2}a_0k^2Z_{1,f,n}. \tag{4.13}$$

For a single frequency of excitation,

$$Z_{1,f,n} = Z_{1,n-1} + Z_{1,n+1}, \tag{4.14}$$

and for two frequencies,

$$Z_{1,f,n} = \cos(\chi)(Z_{1,n-M_1} + Z_{1,n+M_1}) + \sin(\chi)(e^{i\phi}Z_{1,n-M_2} + e^{-i\phi}Z_{1,n+M_2}). \tag{4.15}$$

These equations are supplemented by the reality conditions

$$W_{1,-n}(0) = \bar{W}_{1,n}(0), \quad Z_{1,-n} = \bar{Z}_{1,n}$$

for harmonic modes and

$$W_{1,-n}(0) = \bar{W}_{1,n-1}(0), \quad Z_{1,-n} = \bar{Z}_{1,n-1}$$

for subharmonic modes, where the bar indicates complex conjugation.

4.2. Second order problem for rectangles and stripes. The functions $W_{2,i,n}$ and $Z_{2,i,n}$, $i = 1,2,3$, can all be found from the same system of equations with different choices made for the parameters θ and d. For $i = 1$, we take $\theta = 0$ and $d = 1/2$: this would be the same as solving for stripes. For $i = 2$, we take $\theta = \tilde{\theta}$ and $d = 1$. For $i = 3$, we take $\theta = \pi + \tilde{\theta}$ and $d = 1$, where $\tilde{\theta} \in (0, \pi/2]$. The solutions $W_{2,i,n}$ satisfy

$$\begin{aligned}[i(n\omega+2\alpha) - C(D^2-2k^2(1+c))](D^2-2k^2(1+c))W_{2,i,n}(z) \\ = d\sum_{l+m=n}\big[4k^2s^2DW_{1,l}(z)W_{1,m}(z) - 2(1+c)D^3W_{1,l}(z)W_{1,m}(z) \\ +2(-1+c+2c^2)D^2W_{1,l}(z)DW_{1,m}(z)\big]\end{aligned} \tag{4.16}$$

at $z = -h/l$,

$$W_{2,i,n} = DW_{2,i,n} = 0, \tag{4.17}$$

while at $z = 0$ we have the kinematic condition and the tangential stress condition,

$$i(n\omega + 2\alpha)Z_{2,i,n} - W_{2,i,n} = 2d(1+c)\sum_{l+m=n} DW_{1,l}Z_{1,m}, \tag{4.18}$$

$$\begin{aligned}(D^2 + 2k^2(1+c))W_{2,i,n} = -d\sum_{l+m=n}\big[&2(1+c)Z_{1,l}D^3W_{1,m}\\ &+2(3+2c)(1+c)k^2Z_{1,l}DW_{1,m}\big],\end{aligned} \tag{4.19}$$

and the normal stress condition,

$$\begin{aligned}&\left[i(n\omega+2\alpha) + 6(1+c)Ck^2\right]DW_{2,i,n} - CD^3W_{2,i,n}\\ &\qquad +2k^2(1+c)(2B(1+c)k^2+1)Z_{2,i,n} = S_{1,i,n} - a_0k^2(1+c)Z_{2,f,i,n},\end{aligned} \tag{4.20}$$

where

$$\begin{aligned}S_{1,i,n} = d\sum_{l+m=n}\big[&2c(1+c)DW_{1,l}DW_{1,m}\\ &+4(1+c)k^2Z_{1,l}DP_{1,m} - 8(1+c)Ck^2Z_{1,l}D^2W_{1,m} - 2(1+c)W_{1,l}D^2W_{1,m}\big].\end{aligned}$$

For a single frequency,

$$Z_{2,f,i,n} = Z_{2,i,n-1} + Z_{2,i,n+1},$$

and for two frequencies,

$$Z_{2,f,i,n} = \cos(\chi)(Z_{2,i,n-M_1} + Z_{2,i,n+M_1}) + \sin(\chi)(e^{i\phi}Z_{2,i,n-M_2} + e^{-i\phi}Z_{2,i,n+M_2}). \tag{4.21}$$

Note that if there is no extra symmetry that suppresses it, then the resonant triad interaction at second order causes this calculation to blow up at $\theta = \pi/3$.

4.3. Linear adjoint problem. The adjoint problem is

$$\left[i(n\omega+\alpha) + C(D^2-k^2)\right](D^2-k^2)W_n^*(z) = 0, \tag{4.22}$$

with boundary conditions

$$W_n^* = DW_n^* = 0,$$

on $z = -h/l$, and on $z = 0$,

$$\left[i(n\omega+\alpha) - 3Ck^2\right]DW_n^* + CD^3W_n^* = Z_n^*, \tag{4.23}$$

$$(D^2+k^2)W_n^* = 0, \tag{4.24}$$

$$i(n\omega+\alpha)Z_n^* + k^2(Bk^2+1)W_n^* = -\frac{1}{2}a_0k^2W_{f,n}^*. \tag{4.25}$$

For a single frequency,

$$W_{f,n}^* = W_{n-1}^* + W_{n+1}^*,$$

and for two frequencies,

$$W_{f,n}^* = \cos(\chi)(W_{n-M_1}^* + W_{n+M_1}^*) + \sin(\chi)(e^{i\phi}W_{n-M_2}^* + e^{-i\phi}W_{n+M_2}^*).$$

4.4. Solvability condition for rectangles and stripes. The solvability condition is

$$\delta \frac{dA}{dT} = a_2 \beta A + \gamma A^3, \tag{4.26}$$

where $A \equiv A_S$ in Table 4.1 for stripes and $A \equiv A_R$ in Table 4.1 for rectangles and

$$\begin{aligned} \delta = \sum_{l,m}^{(1)} \Bigg[& -2 Z_l^* Z_{1,m} + 2 W_l^*(0) D W_{1,m}(0) \\ & + 2 \int_{-h/l}^{0} W_l^*(z) \Big(-D^2 W_{1,m}(z) + k^2 W_{1,m}(z) \Big) dz \Bigg]. \end{aligned} \tag{4.27}$$

For one frequency,

$$\beta = -k^2 \left[\sum_{l,m}^{(4)} W_l^*(0) Z_{1,m} + \sum_{l,m}^{(5)} W_l^*(0) Z_{1,m} \right], \tag{4.28}$$

and for two frequencies,

$$\begin{aligned} \beta = -k^2 \Bigg[& \cos(\chi) \left(\sum_{l,m}^{(6)} W_l^*(0) Z_{1,m} + \sum_{l,m}^{(7)} W_l^*(0) Z_{1,m} \right) \\ & + \sin(\chi) \left(\sum_{l,m}^{(8)} W_l^*(0) e^{i\psi} Z_{1,m} + \sum_{l,m}^{(9)} W_l^*(0) e^{-i\psi} Z_{1,m} \right) \Bigg]. \end{aligned} \tag{4.29}$$

The coefficient γ is given by the sum of three separate components corresponding to contributions from each of $W_{2,i,n}$ and $Z_{2,i,n}$. They may each be calculated from a single function $\gamma_i(\theta_s, d)$ by taking each of the three functions in turn and different θ_s and d. Specifically,

$$\gamma(\theta_s, d) = \gamma_1 \left(0, \frac{1}{2} \right) + \gamma_2(\theta_s, 1) + \gamma_3(\pi + \theta_s, 1).$$

The function γ_i is given in Appendix B.

4.5. Second order solvability condition for hexagons. The second order solvability condition for hexagons leads to the amplitude equation

$$\delta \frac{dA_H}{dT_1} = a_1 \beta A_H + \gamma_2 A_H^2, \tag{4.30}$$

and γ_2 is given by

$$\begin{aligned} \gamma_2 = & -2 \sum_{l,m,n}^{(2)} Z_l^* Z_{1,m} D W_{1,n}(0) \\ & - \sum_{l,m,n}^{(2)} W_l^*(0) \left[D W_{1,m}(0) D W_{1,n}(0) + 8 C k^2 Z_{1,m} D^2 W_{1,n}(0) \right] \\ & + \sum_{l,m,n}^{(2)} W_l^*(0) \left[4 k^2 Z_{1,m} D P_{1,n}(0) - 2 W_{1,m}(0) D^2 W_{1,n}(0) \right] \end{aligned}$$

$$-C\sum_{l,m,n}^{(2)} DW_l^*(0)Z_{1,m}\left[2D^3W_{1,n}(0)+4k^2DW_{1,n}(0)\right]$$
$$-\sum_{l,m,n}^{(2)}\int_{-h/l}^{0} W_l^*(z)\Big[W_{1,m}(z)(6k^2DW_{1,n}(z)-2D^3W_{1,n}(z))$$
$$-4DW_{1,m}(z)D^2W_{1,n}(z)\Big]\,dz. \tag{4.31}$$

4.6. Second order problem for hexagons. The second order problem for hexagons consists of $W_{2,1,n}(z) = W_{2,n}(z)$, which solves the same problem as that obtained for stripes, and $W_{2,2,n}$, which satisfies

$$[i(n\omega+2\alpha)-C(D^2-3k^2)](D^2-3k^2)W_{2,2,n}(z)$$
$$=\sum_{l+m=n}(3k^2W_{1,l}(z)DW_{1,m}(z)-3D^3W_{1,l}(z)W_{1,m}(z)).$$

At $z=-h/l$ we have $W_{2,2,n}=DW_{2,2,n}=0$, while at $z=0$ we have the kinematic condition and the tangential stress conditions, namely,

$$i(n\omega+2\alpha)Z_{2,2,n}-W_{2,2,n}=3\sum_{l+m=n}Z_{1,l}DW_{1,m},$$
$$(3k^2+D^2)W_{2,2,n}=-\sum_{l+m=n}\Big[3Z_{1,l}D^3W_{1,m}+12k^2Z_{1,l}DW_{1,m}\Big],$$

and the normal stress condition,

$$(i(n\omega+2\alpha)+9Ck^2)DW_{2,2,n}-CD^3W_{2,2,n}+3k^2(3Bk^2+1)Z_{2,2,n}$$
$$=S_{1,h,n}-\frac{3}{2}a_0k^2Z_{2,f,n},$$

where

$$S_{1,h,n}=\sum_{l+m=n}\left(\frac{3}{2}DW_{1,l}DW_{1,m}-12Ck^2Z_{1,l}D^2W_{1,m}\right.$$
$$\left.+6k^2Z_{1,l}DP_{1,m}-3W_{1,l}D^2W_{1,m}\right).$$

For a single frequency of excitation,

$$Z_{2,f,n}=Z_{2,2,n-1}+Z_{2,2,n+1},$$

and for two frequencies,

$$Z_{2,f,n}=\cos(\chi)(Z_{2,2,n-M_1}+Z_{2,2,n+M_1})+\sin(\chi)(e^{i\phi}Z_{2,2,n-M_2}+e^{-i\phi}Z_{2,2,n+M_2}).$$

4.7. Solvability condition for hexagons at third order. The solvability condition at third order problem takes the form

$$\delta\frac{dA_H}{dT}=a_2\beta A_H+\gamma_2A_H^2+(\gamma_1+\gamma_3)A_H^3, \tag{4.32}$$

where δ, β, and γ_1 are the same as for stripes, γ_2 is given by (4.31), and γ_3 is given in Appendix C.

5. Calculation of the coefficients of the amplitude equations. In section 4, we formulated a hierarchy of problems for the z dependence of the fluid parameters. In this section, we outline how we solve this sequence of problems and calculate the coefficients of the amplitude equations. We focus on the harmonic case, $\alpha = 0$, but the calculations for the subharmonic case are similar. We also note that using the transformation $\omega \to \omega/2$, $M_1 \to 2M_1$, $M_2 \to 2M_2$, the subharmonic case can be incorporated into the formulation for the harmonic problem. As explained in section 5.1, the calculations are substantially simpler in the specific case of an infinite layer, and while we have solved the linear problem for both the finite and infinite cases, we have calculated the coefficients of the amplitude equations only for infinite depth. We justify why this is a reasonable approximation for the specific fluid parameter choices we use when we discuss the results in section 6.

5.1. Linear problem. The general solution for $W_{1,n}(z)$ that satisfies (4.9) is

$$\begin{aligned} W_{1,0}(z) &= (a_{1,0} + zc_{1,0})e^{kz} + (b_{1,0} + zd_{1,0})e^{-kz}, \\ W_{1,n}(z) &= a_{1,n}e^{kz} + b_{1,n}e^{-kz} + c_{1,n}e^{q_{1,n}z} + d_{1,n}e^{-q_{1,n}z}, \qquad n \neq 0, \end{aligned} \tag{5.1}$$

where

$$q_{1,n}^2 = \frac{in\omega}{C} + k^2.$$

Applying the boundary conditions allows values for the coefficients $a_{1,n}$, $b_{1,n}$, $c_{1,n}$, and $d_{1,n}$ to be found. However, the subsequent analysis is substantially easier in the particular case of an infinite layer, where the lower boundary conditions are replaced with the requirement that the solution be bounded as $z \to -\infty$. In this case, $b_{1,n} = d_{1,n} = 0$, and it is this case we discuss below.

In order to find $a_{1,n}$ and $c_{1,n}$, they are expressed first in terms of $Z_{1,n}$ using the kinematic condition (4.11) and the tangential stress condition (4.12) to give

$$\begin{aligned} a_{1,0} &= 0, \\ a_{1,n} &= (in\omega + 2Ck^2)Z_{1,n}, \qquad n \neq 0, \\ c_{1,0} &= 0, \\ c_{1,n} &= -2Ck^2 Z_{1,n}, \qquad n \neq 0. \end{aligned} \tag{5.2}$$

Then the normal stress condition (4.13) is used to eliminate $W_{1,n}$ to give

$$\begin{aligned} k^2\left(Bk^2+1\right)Z_{1,0} &= -\frac{1}{2}a_0k^2Z_{1,f,0}, \qquad n = 0, \\ \left[k(2Ck^2+in\omega)^2 - 4C^2k^4q_{1,n} + k^2\left(Bk^2+1\right)\right]Z_{1,n} &= -\frac{1}{2}a_0k^2Z_{1,f,n}, \qquad n \neq 0. \end{aligned} \tag{5.3}$$

This is a generalized eigenvalue problem for a_0 of the form

$$\mathbf{AZ} = a_0\mathbf{BZ},$$

where $\mathbf{Z} = (Z_{1,0}, Z_{1,1}, \ldots, Z_{1,N})^T$. This generalized eigenvalue problem is the same as that solved in [15]. The minimum real positive eigenvalue a_0 gives the critical amplitude of onset of patterns, and the corresponding eigenvector gives the values for $Z_{1,n}$. Hence $W_{1,n}$ can be found from (5.2) and (5.1). These calculations and those that follow were carried out using MATLAB. For the majority of the harmonic calculations, we take $N = 20$, and for the subharmonic calculations, we take $N = 40$. Doubling the number of modes typically changed the results by less than 0.2%.

5.2. Linear adjoint problem. The general solution for the linear adjoint problem (4.22) for a fluid of infinite depth is

$$W_0^*(z) = (a_0^* + zc_0^*)e^{kz},$$
$$W_n^*(z) = a_n^* e^{kz} + c_n^* e^{q_n^* z}, \qquad n \neq 0,$$

where

$$q_n^{*\,2} = -\frac{in\omega}{C} + k^2.$$

Note that q_n^* is the complex conjugate of $q_{1,n}$. This can be solved analytically in terms of Z_n^* using the adjoint equivalents to the kinematic condition (4.23) and the tangential stress condition (4.24), and after some manipulation we find

$$Z_0^* = -2Ck^3 W_0^*(0),$$
$$Z_n^* = \frac{1}{ind_n\omega} W_n^*(0), \qquad n \neq 0, \tag{5.4}$$

where

$$d_n = \left(k(2Ck^2 - in\omega)^2 - 4C^2k^4 q_n^*\right)^{-1}.$$

Substitution of these expressions for Z_n^* into the normal stress condition (4.25) gives

$$k^2(Bk^2+1)W_0^*(0) = -\frac{1}{2}a_0^* k^2 W_{f,0}^*(0), \tag{5.5}$$
$$\left[k(2Ck^2 - in\omega)^2 - 4C^2k^4 q_n^* + k^2(Bk^2+1)\right] W_n^*(0) = -\frac{1}{2}a_0^* k^2 W_{f,n}^*(0), \qquad n \neq 0.$$

This results in a second generalized eigenvalue problem of the form

$$\mathbf{A}^*\mathbf{W}^* = a_0^*\mathbf{B}^*\mathbf{W}^*.$$

As expected, when we solve this, we find $a_0 = a_0^*$, and $W_n^*(0)$ is the complex conjugate of $Z_{1,n}$. Once W_n^* has been found, then Z_n^* follows from (5.4).

5.3. Second order problem for stripes and rectangles. Next, we consider the problem for $W_{2,i,n}(z)$, (4.16). The homogeneous equation has the solution

$$W_{2,i,0}(z) = (a_{2,i,0} + zc_{2,i,0})e^{\tilde{k}z},$$
$$W_{2,i,n}(z) = a_{2,i,n}e^{\tilde{k}z} + c_{2,i,n}e^{q_{2,i,n}z}, \qquad n \neq 0,$$

where

$$\tilde{k}^2 = 2(1+c)k^2,$$
$$q_{2,i,n}^2 = \frac{in\omega}{C} + \tilde{k}^2.$$

In the inhomogeneous equation (4.16), the right-hand side generates terms of the form $\alpha_{l,m}e^{Q_{l,m}z}$. Each of these contributes to the solution a term $\delta_{l,m}e^{Q_{l,m}z}$, where

$$\delta_{l,m} = \frac{\alpha_{l,m}}{\left(in\omega - C\left(Q_{l,m}^2 - \tilde{k}^2\right)\right)\left(Q_{l,m}^2 - \tilde{k}^2\right)}.$$

The general solution to (4.16) is therefore

$$W_{2,i,0}(z) = (a_{2,i,0} + c_{2,i,0}z)e^{\tilde{k}z} + \sum_{l+m=0} \delta_{l,m} e^{Q_{l,m}z},$$
$$W_{2,i,n}(z) = a_{2,i,n}e^{\tilde{k}z} + c_{2,i,n}e^{q_{2,i,n}z} + \sum_{l+m=n} \delta_{l,m} e^{Q_{l,m}z}, \qquad n \neq 0. \tag{5.6}$$

The coefficients $a_{2,i,n}$ and $c_{2,i,n}$ can be found in terms of $Z_{2,i,n}$, using the tangential stress condition (4.19) and the kinematic condition (4.18), to give

$$\begin{aligned} a_{2,i,n} &= \gamma_{1,i,n} Z_{2,i,n} + v_{1,i,n}, \\ c_{2,i,n} &= \gamma_{2,i,n} Z_{2,i,n} + v_{2,i,n}, \end{aligned} \tag{5.7}$$

where

$$\begin{aligned} &\gamma_{1,i,0} = 0, \\ &\gamma_{1,i,n} = (in\omega + 2C\tilde{k}^2), \qquad n \neq 0, \\ &\gamma_{2,i,0} = 0, \\ &\gamma_{2,i,n} = -2C\tilde{k}^2, \qquad n \neq 0, \end{aligned}$$

and

$$\begin{aligned} v_{1,i,0} &= S_{2,i,0}, \\ v_{1,i,n} &= \frac{C}{in\omega}\left(\left(\frac{in\omega}{C} + 2\tilde{k}^2\right) S_{2,i,n} - S_{3,i,n}\right), \qquad n \neq 0, \\ v_{2,i,0} &= \frac{1}{2\tilde{k}}\left(S_{3,i,0} - 2\tilde{k}^2 S_{2,i,0}\right), \\ v_{2,i,n} &= \frac{C}{in\omega}\left(-2\tilde{k}^2 S_{2,i,n} + S_{3,i,n}\right), \qquad n \neq 0, \end{aligned}$$

and

$$\begin{aligned} S_{2,i,n} &= -2(1+c)d \sum_{l+m=n} Z_{1,l} D W_{1,m} - \sum_{l+m=n} \delta_{l,m}, \\ S_{3,i,n} &= -d \sum_{l+m=n} \left(2(1+c) Z_{1,l} D^3 W_{1,m} + (3+2c)\tilde{k}^2 Z_{1,l} D W_{1,m}\right) \\ &\quad - \sum_{l+m=n} Q_{l,m}^2 \delta_{l,m} - \tilde{k}^2 \sum_{l+m=n} \delta_{l,m}. \end{aligned}$$

Substitution of the general solution (5.6) into the normal stress condition, (4.20), gives

$$\begin{aligned} \tilde{k}^2\left(B\tilde{k}^2 + 1\right) Z_{2,i,0} + \frac{1}{2} a_0 \tilde{k}^2 Z_{2,f,i,0} &= -2C\tilde{k}^3 S_{2,i,0} + S_{1,i,0} \\ &\quad -3C\tilde{k}^2 \sum_{l+m=0} Q_{l,m}\delta_{l,m} + C \sum_{l+m=0} Q_{l,m}^3 \delta_{l,m}, \end{aligned}$$

$$\Big(\tilde{k}(in\omega + 2C\tilde{k}^2)\gamma_{1,i,n} + 2C\tilde{k}^2 q_{2,i,n}\gamma_{2,i,n}$$
$$+\tilde{k}^2\left(B\tilde{k}^2+1\right)\Big) Z_{2,i,n} + \frac{1}{2}a_0\tilde{k}^2 Z_{2,f,i,n} = -\tilde{k}(in\omega + 2C\tilde{k}^2)v_{1,i,n} - 2C\tilde{k}^2 q_{2,i,n}v_{2,i,n}$$
$$+S_{1,i,n} - (in\omega + 3C\tilde{k}^2)\sum_{l+m=n} Q_{l,m}\delta_{l,m}$$
$$+C\sum_{l+m=n} Q^3_{l,m}\delta_{l,m}, \qquad n \neq 0.$$

This is of the form

$$\mathbf{A}\mathbf{Z}_2 = \mathbf{b},$$

where $\mathbf{Z}_2 = (Z_{2,i,0}, Z_{2,i,1}, \ldots, Z_{2,i,N})$. This can be solved for $Z_{2,i,n}$, and hence $W_{2,i,n}$ can be found from (5.7) and (5.6).

5.4. Solvability condition for stripes and rectangles. Once the calculations for the linear, linear adjoint, and second order problem have been completed, the coefficients for the solvability condition are calculated from (4.27), (4.28), (4.29), and (B.1).

5.5. Second order problem and solvability condition for hexagons. The linear and linear adjoint problems for hexagons are the same as for stripes and rectangles. Once Z_1, W_1, Z^*, and W^* are known, the size of the quadratic coefficient in the amplitude equation for hexagons may be computed from (4.31). The calculations of the second order solution and solvability condition at third order then follow in a very similar fashion to the calculation for rectangles: the same products appear but with different coefficients.

5.6. Evaluating the coefficients of the amplitude equations. In order to compute the stability of the patterns as given in Table 4.1, we need to calculate each of the cubic coefficients b_i that appear in the generic amplitude equations (4.1) along with the value of the quadratic coefficient $\tilde{\epsilon}$. From the calculations for rectangles, stripes, and hexagons this may be done as follows. The solvability condition for rectangles and stripes takes the form given in (4.26), that is,

$$\delta\frac{dA}{dT} = a_2\beta A + \gamma A^3. \tag{5.8}$$

The values of δ and β are independent of the type of pattern considered, but the value of γ is not. By comparing the amplitude equations with the equations given for each state given in Table 4.1, we see that when γ is computed in the case of stripes, it gives us the value for b_1. If the solvability condition is calculated for rectangles, then γ gives us $b_1 + b_4$, and hence the value of b_4 (and similarly b_5 and b_6) may be found. In the case of hexagons, the solvability condition takes the form

$$\delta\frac{\partial A_H}{\partial T} = a_2\beta A_H + \gamma_2 A_H^2 + (\gamma_1 + \gamma_3)A_H^3, \tag{5.9}$$

where δ, β, and γ_1 are the same as for stripes and γ_2 is given by (4.31) and γ_3 comes from (C.1). By comparing with the amplitude equation for hexagons given in Table 4.1, we see that γ_3 gives the value of $2b_2$. Note that while b_1 and b_2 are fixed for a given set of fluid parameters, the values of b_4, b_5, and b_6 depend on the lattice angle θ.

In the results we present in the next section, we rescale the amplitude equations by letting $A \mapsto \sqrt{|\beta/b_1|}A$, $A_H \mapsto \sqrt{|\beta/b_1|}A_H$, and $T \mapsto |\delta/\beta|$ so that the cubic coefficient for stripes b_1 in the rescaled equations is always ± 1.

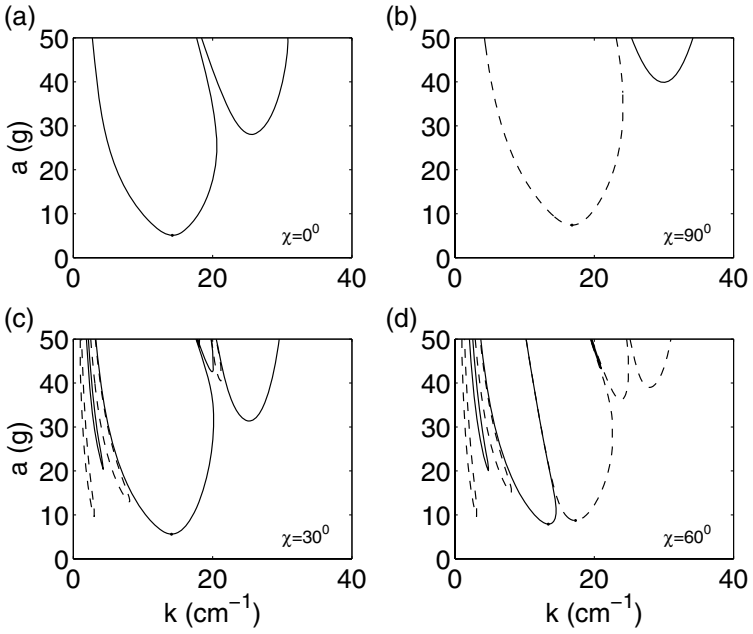


FIG. 6.1. *Linear stability curves for different values of* χ. $\rho = 0.95cm^{-3}$, $\nu = 20.9cS$, $\sigma = 20.6dyn/cm$. *Excitation is* $a(\cos\chi\cos M_1\omega t + \sin\chi\cos M_2\omega t)$ *with* $(M_1, M_2) = (4, 5)$, $\omega = 44\pi$. *Harmonic tongues are marked with solid lines and subharmonic tongues with dashed lines.*

6. Results. Here we present the results of our calculations for the particular fluid parameters used in the experimental results given in [4].

6.1. Linear stability results. First, the generalized eigenvalue problem (5.3) is solved, and this gives the critical amplitude as a function of k. Typical curves are shown in Figure 6.1. These are analogous to those computed by Besson, Edwards, and Tuckerman [15] but focus on the particular parameter values that are used by Kudrolli, Pier, and Gollub [4], namely $\rho = 0.95\text{cm}^{-3}$, $\nu = 20.9$cS, and $\sigma = 20.6$dyn/cm. Curves that have a subharmonic response with the excitation are indicated with dotted lines, and those that are harmonic are marked with a solid line. The critical onset occurs at the minimum value of a, and this occurs at a critical wavenumber k_c. For ease of comparison with the experimental results, the curves are plotted in dimensional rather than nondimensional units: with our choice of nondimensionalization, the value of k_c is always 1. The minimum value point is indicated by a dot. The richness of the dynamics that is seen with two-frequency forcing is partly due to the fact that the parameter χ allows one to tune between the critical value for a occurring for either a harmonic tongue, as is the case for $0^\circ < \chi < 60^\circ$, or a subharmonic tongue, as is the case for $60^\circ < \chi < 90^\circ$. There is a bicritical point that occurs when χ is approximately 60°. The phase ϕ has little impact on the position of the minimum value of the lowest tongue and therefore in the position of the bicritical point (changes in the position of the mimima are typically less than 0.001%). The phase does, however, alter the position of the tongues for some of the other harmonics.

Experimentally, it is the onset of patterns as a function of χ that is observed rather than the linear stability curves directly. In Figure 6.2, we show how the minimum of the linear stability curves varies with the relative importance of the amplitude of

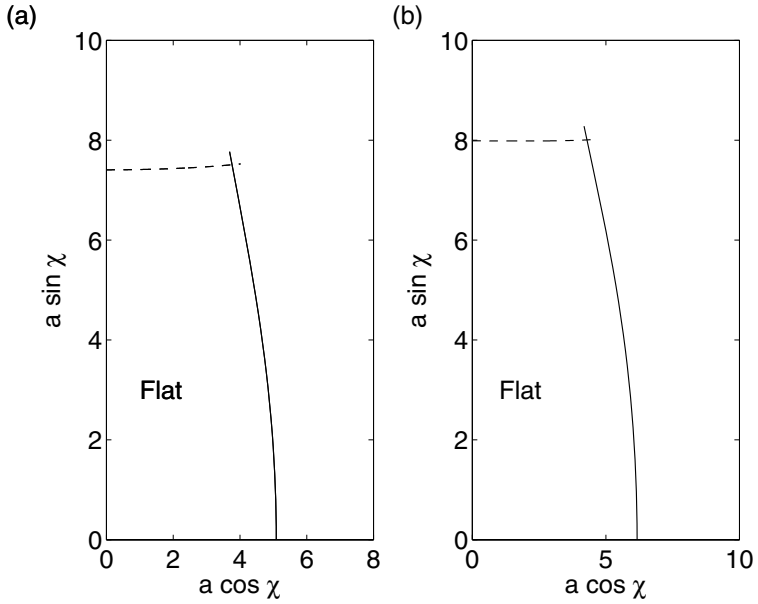


FIG. 6.2. *Path of a_0 as a function of χ.* (a) $(M_1, M_2) = (4, 5)$, $\phi = 16°$, *and* $\omega = 44\pi$; (b) $(M_1, M_2) = (6, 7)$, $\phi = 20°$, *and* $\omega = 32.88\pi$.

the two components of the excitation as given by $a\cos\chi$ and $a\sin\chi$. Two cases are shown: Figure 6.2(a) is for the same parameter values as for Figure 6.1. Figure 6.2(b) is for the same fluid parameters but different excitation parameters. These two cases correspond to the two cases considered in detail in Kudrolli, Pier, and Gollub, and the linear stability boundaries compare well with the corresponding experimental results shown in Figures 1 and 6 of their paper [4]. The bicritical point occurs for $\chi = 63.4$ in the case $(M_1, M_2) = (4, 5)$ and $\chi = 61.8$ when $(M_1, M_2) = (6, 7)$. These compare well with the value quoted by Kudrolli, Pier, and Gollub of $\chi = 61.5$.

As discussed in the introduction, high viscosity or shallow depth are used to damp modes with a small wavenumber that can make regular patterns hard to observe. The fluid used by Kudrolli, Pier, and Gollub was of moderate viscosity ($C \approx 0.4$ rather than $C \ll 1$). How "shallow" a container is depends on the product $k_c h$, in particular whether $e^{-k_c h}$ is negligible when compared with $e^{k_c h}$. For the Kudrolli–Pier–Gollub experiments, $h = 0.3$cm, giving $k_c \approx 14$ and $e^{-k_c h}/e^{k_c h} \approx 0.0002$. In the weakly nonlinear calculations, we shall see that it is not just the main harmonic tongue that is of importance but the weakly damped harmonic tongue that has a minimum at $k \approx 6\text{cm}^{-1}$. For this tongue, $e^{-k_c h}/e^{k_c h} \approx 0.03$. Since this is also small, we believe that an infinite depth approximation is reasonable. Consequently, in the nonlinear results that we present below, the calculations are performed only for infinite depth.

6.2. Single-frequency results ($\chi = 0°$ and $\chi = 90°$). The two-frequency excitation term we are interested in is of the form

$$f(t) = f_2(t) = \cos(\chi)\cos(M_1\omega t) + \sin(\chi)\cos(M_2\omega t + \phi).$$

When $\chi = 0°$ or $90°$, this reduces to a single-frequency excitation. When $\chi = 0°$, this corresponds to a pure M_1 excitation, and when $\chi = 90°$, this is a pure M_2 excitation.

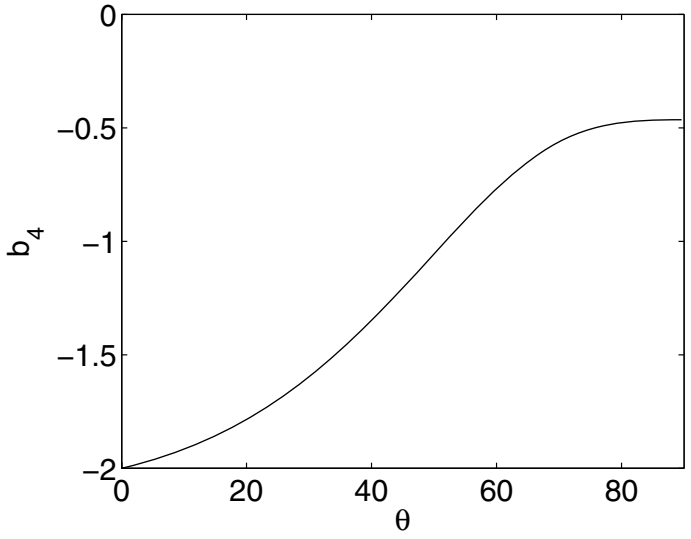


FIG. 6.3. *For* $(M_1, M_2) = (4, 5)$ *and* $\chi = 0°$ *the value of the coefficient* b_4 *as a function of the lattice angle* θ.

Since the two cases we consider have M_1 even and M_2 odd, the pure M_1 response is harmonic and the pure M_2 response is subharmonic. In each case, the overall picture is similar. The quadratic term in the amplitude equations is zero and we find the following:

- All the eigenvalues for hexagons given in Table 4.1 are negative for all values of the lattice angle θ. This suggests that hexagons are a stable state.
- Stripes and superlattice patterns have at least one unstable eigenvalue.
- Rectangular patterns are stable, on the lattice on which they occur, if they are "sufficiently square."

This last point may be seen by considering the eigenvalues of the family of rectangles $R_{h1,m,n}$ listed in Table 4.1. The eigenvalues $b_1 + b_4$, μ_1, and μ_2 are all negative. The remaining eigenvalue, $b_1 - b_4$, is negative only if the lattice angle θ is sufficiently large. In the case $(M_1, M_2) = (4, 5)$ and $\chi = 0°$ in Figure 6.3, we plot the scaled value of b_4 as a function of the lattice angle θ. Since in the scaled units $b_1 = -1$, $b_1 - b_4 < 0$ only if $\theta > 52°$. The aspect ratio of the rectangles is given by $\sqrt{(1 - \cos\theta)/(1 + \cos\theta)}$ so that stability for rectangles with $\theta > 52°$ means that rectangles with an aspect ratio between 0.48 and 1 are stable (on the lattice on which they occur).

By considering the eigenvalues on spatially periodic lattices alone, we find that both hexagons and rectangles are possible stable states. If we find the values of the Lyapunov function $\mathcal{F}$ (see Table 4.2) for hexagons and rectangles, then we find that those rectangles with an aspect ratio closest to 1 have the lowest value.

The results are similar for the three other single frequency cases that are relevant to the bifurcation sets shown in Figure 6.2, namely $(M_1, M_2) = (4, 5)$ and $\chi = 90°$ and $(M_1, M_2) = (6, 7)$ and $\chi = 0°$ or $\chi = 90°$. The conclusion is that the calculations for solutions on periodic lattices show that stripe patterns are unstable. Hexagons and "sufficiently square" rectangular patterns are relatively stable states with square patterns having the lowest value of the Lyapunov function. This suggests that square patterns will be observed at the points where $\chi = 0°$ and $\chi = 90°$. These square

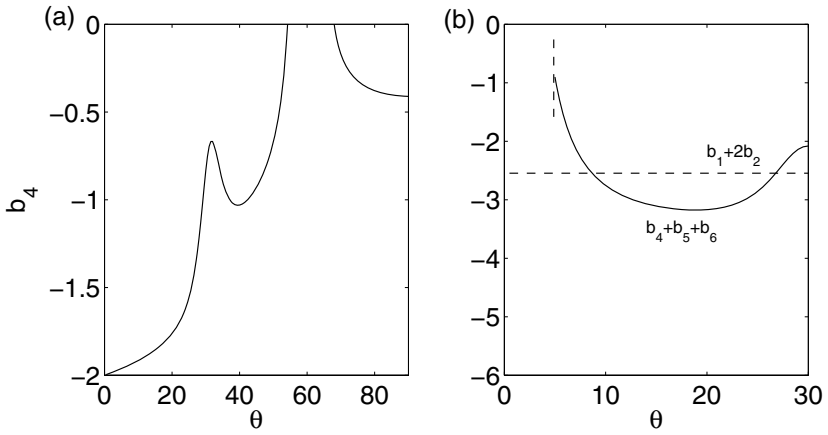


FIG. 6.4. $(M_1, M_2) = (4, 5)$ *and* $\chi = 60°$, $\phi = 20°$*:* (a) *The value of the coefficient* b_4 *as a function of the lattice angle* θ*. At* $\theta = 60°$ *quadratic resonance occurs, and the calculation for* b_4 *is not valid. For this reason, this point has been excluded from the calculation.* (b) *The value of the coefficient* $b_4 + b_5 + b_6$ *compared with* $b_1 + 2b_2$ *as a function of the lattice angle* θ*. Just as* b_4 *has a singularity at* $60°$*,* b_5 *has a singularity at* $0°$*, and so the region around* $0°$ *has been excluded.*

patterns will be harmonic when $\chi = 0°$ and subharmonic when $\chi = 90°$. This agrees with the experimental findings of Kudrolli, Pier, and Gollub.

Note that if we use the same parameter values as used by Chen and Viñals in [22] in their single-frequency study, then we get excellent agreement with their work.

6.3. Two-frequency results ($0 < \chi < 90°$). For two-frequency excitation, we have performed a systematic study of the coefficients $\tilde{\epsilon}$, b_1, b_2, b_3, b_4, b_5, and b_6 as a function of χ and the lattice angle θ for the same fluid parameters as used above and for the two cases $(M_1, M_2) = (4, 5)$ and $(M_1, M_2) = (6, 7)$. In the first case, most results are presented for $\phi = 16°$ and in the second case for $\phi = 20°$: these are the values for which the majority of the results in [4] are presented. We focus on the onset of harmonic patterns since along the subharmonic branch squares the absence of the quadratic term means that hexagons and rectangles remain the only stable states, with squares having the lowest energy.

From the values of the coefficients we have computed the stability of the different planforms based on the eigenvalues for each state given in Table 4.1 as a function of the lattice angle. If the eigenvalues indicate that the state is relatively stable, then we compute the value of the Lyapunov function, as given in Table 4.2.

When χ is zero, b_4 increases monotonically with θ, as shown in Figure 6.3. However, this changes as χ is increased and peaks develop. In Figure 6.4(a), we plot $b_4(\theta)$ for $(M_1, M_2) = (4, 5)$ and for $\chi = 60°$, $\phi = 16°$. The angle $\theta = 60°$ has been excluded because this corresponds to the quadratic resonance point, and the calculation for rectangles breaks down here. From Table 4.1 it can be seen that for superlattice patterns to become stable, one needs $b_4 + b_5 + b_6 - b_1 - 2b_2 > 0$. This same quantity causes the destabilization of hexagons. In Figure 6.4(b), we plot $b_4 + b_5 + b_6$ along with $b_1 + 2b_2$. For most values of θ, and so on most lattices, $b_4 + b_5 + b_6 - b_1 - 2b_2 < 0$, and superlattice patterns are unstable. However, the peak in $b_4(\theta)$ at $31.8°$ leads to a small region centered at $30°$ for which $b_4 + b_5 + b_6 - b_1 - 2b_2 > 0$, and it is possible for superlattice patterns to be stable.

The corresponding graphs of b_4 and $b_4 + b_5 + b_6$ for $(M_1, M_2) = (6, 7)$ are

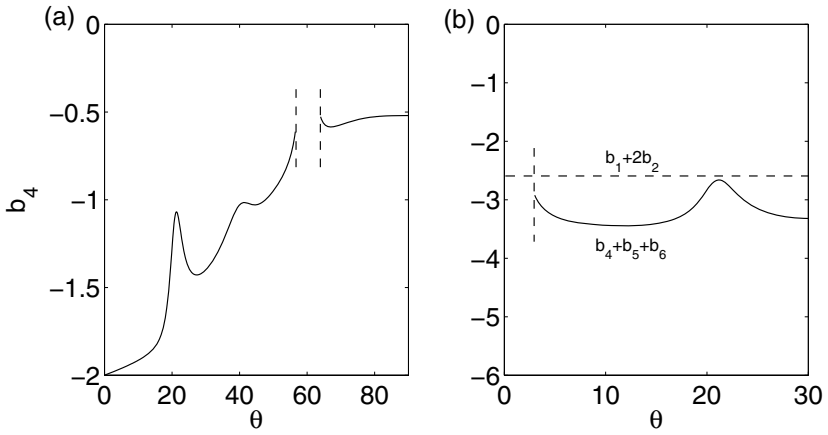


FIG. 6.5. $(M_1, M_2) = (6, 7)$ *and* $\chi = 60°$*:* (a) *The value of the coefficient* b_4 *as a function of the lattice angle* θ*;* (b) *the value of the coefficient* $b_4 + b_5 + b_6$ *compared with* $b_1 + 2b_2$ *as a function of the lattice angle* θ*.*

shown in Figure 6.5. In this case, there is a peak in b_4 at 21.2°, and this leads to a peak in $b_4 + b_5 + b_6$ with a maximum at 21.2°. Note that even at $\chi = 60°$, $b_4 + b_5 + b_6 - b_1 - 2b_2 < 0$, and all superlattice patterns are unstable. It is only once, $\chi > 61°$, that stable superlattice patterns occur, the first to stabilize being those that occur on a lattice with lattice angle $\theta = 21.2°$.

We have shown for $\chi = 0°$ that both hexagons and rectangles may be stable with rectangles having the lowest value of the Lyapunov function. Figure 6.4 suggests that in the case $(M_1, M_2) = (4, 5)$, stable superlattice patterns may occur for $\chi = 60°$. In Figure 6.6(a), we show a bifurcation set that summarizes the regions where different states are stable for the λ, χ plane. Note that hexagons are a planform that exists on all the hexagonal lattices and so that where hexagons are shown as stable they are stable to perturbations on all hexagonal lattices. Where they are unstable, there is at least one lattice on which they are unstable. The peak in $b_4(\theta)$ at 30° means that the maximal region of instability for hexagons occurs as $\theta_h \to 30°$. For rectangles and superlattices, the story is more complicated. Different lattices support different superlattice patterns and different rectangles. Since it is on lattices with $\theta \to 30°$ that hexagons become unstable first (and superlattices onset first), we plot the regions for the stability of rectangles and superlattice patterns for the specific case $\theta = 30°$.

At the actual value of $\theta = 30°$ the center manifold reduction that leads to the amplitude equations is not formally valid, as discussed in [24], and the "superlattice patterns" are in fact quasipatterns. However, in practical terms, there is little real difference between taking a periodic lattice that has a lattice angle close to 30° and taking 30° itself: although on the periodic lattice the amplitude equations may be formally justified, the spectral gap between critical and noncritical eigenvalues will be small, and thus the formal region of validity for the center manifold is likely to be small. Visually, it is impossible to distinguish between a quasipattern and a superlattice pattern with a very large lengthscale. Similarly, the regions of stability would be indistinguishable whether we took $\theta = 30°$ or a value of θ close to 30° that results in a spatially periodic lattice.

As shown in Figure 6.6(a), for most values of λ and χ, there is bistability where more than one pattern is stable. In Figure 6.6(b), we show which of the stable

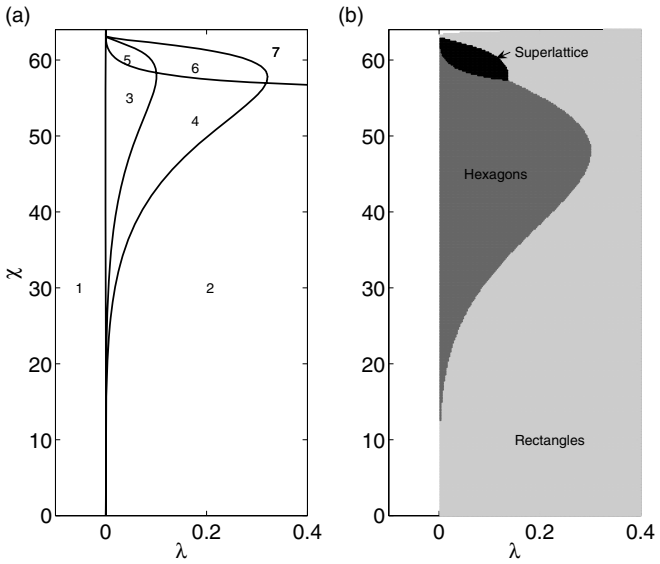


FIG. 6.6. *Results in the λ, χ plane for $(M_1, M_2) = (4, 5)$ and $\phi = 16°$.* (a) *Eigenvalue results. The regions correspond to* 1. *trivial state stable,* 2. *stable rectangles and hexagons,* 3. *stable hexagons and superlattice patterns,* 4. *stable rectangles, hexagons, and superlattice patterns,* 5. *stable superlattice patterns,* 6. *stable superlattice patterns and rectangles,* 7. *stable rectangles.* (b) *Patterns with the lowest energy as computed from the Lyapunov function. Black: superlattice patterns. Dark grey: hexagons. Light grey: rectangles.*

patterns has the lowest value of the Lyapunov function. What we find is that, for small values of χ, rectangles and hexagons are both stable, with rectangles being the most stable state. Note that it is the $R_{h3,m,n}$ rectangles that are stable, and as $\theta \to 30°$, the aspect ratio of these particular rectangles tends to 1. As χ increases, the value of $\tilde{\epsilon}$ increases from zero, and this results in the destabilization of the rectangles so that hexagons become the preferred state at onset. Near the bicritical point, hexagons are themselves destabilized to superlattice patterns, and there are regions where superlattice patterns are the only stable state. A typical bifurcation diagram for $(M_1, M_2) = (4, 5)$ at $\chi = 60°$ is shown in Figure 6.7. Note that the values of $\tilde{\epsilon}$ are small so that higher order correction terms for the position of the secondary bifurcation points are unlikely to affect the overall qualitative bifurcation sequence.

The analogous bifurcation set and energy diagram for the case $(M_1, M_2) = (6, 7)$ are shown in Figure 6.8. In this case, the lattice angle $\theta = 21.2°$ has been used since in the case $(6, 7)$, it is on this lattice that hexagons are destabilized first. On this lattice, the stable rectangles have aspect ratio 0.86. The overall picture is similar to that for $(4, 5)$ but with the transitions from rectangle to hexagon and hexagon to superlattice pattern occurring for larger values of χ.

Figures 6.6 and 6.8 compare well with the experimental results of Kudrolli, Pier, and Gollub. As is suggested by our results, they find that rectangles are stable for low values of χ giving way to hexagons as χ is increased. As for our theoretical results, the transition from rectangles to hexagons occurs for higher values of χ in the $(6, 7)$ case than in the $(4, 5)$ case. In both cases, there are superlattice/quasipatterns near the bicritical point.

Kudrolli, Pier, and Gollub also investigated the dependence of their results on the

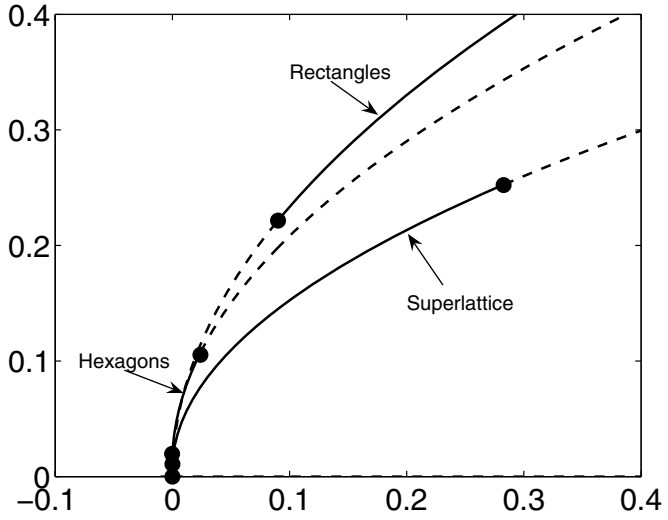


FIG. 6.7. *Bifurcation diagram for* $(M_1, M_2) = (4, 5)$ *and* $\chi = 60°$, $\phi = 16°$. *Stable branches are shown by a solid line and unstable lines by a dotted line. Only branches where some part is stable are shown.*

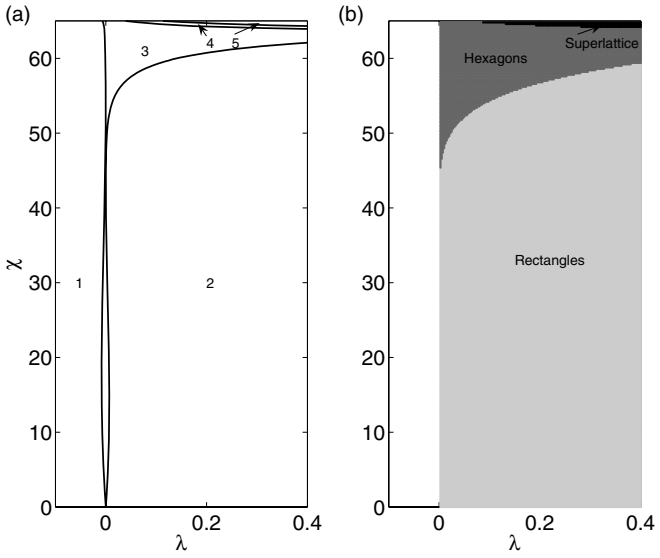


FIG. 6.8. *Results in the* λ, χ *plane for* $(M_1, M_2) = (6, 7)$. (a) *The regions correspond to* 1. *trivial state stable,* 2. *stable rectangles and hexagons,* 3. *hexagons,* 4. *stable hexagons and superlattice patterns,* 5. *stable superlattice patterns.* (b) *Patterns with the lowest energy as computed from the Lyapunov function. Black: superlattice patterns; dark grey: hexagons; light grey: rectangles.*

phase ϕ for $\chi = 61°$. In the $(6, 7)$ case, they found relatively little phase dependence. In the $(4, 5)$ case, they found that the largest region of superlattice patterns was for angles of ϕ close to $16°$, but they also found that for some values of ϕ there were

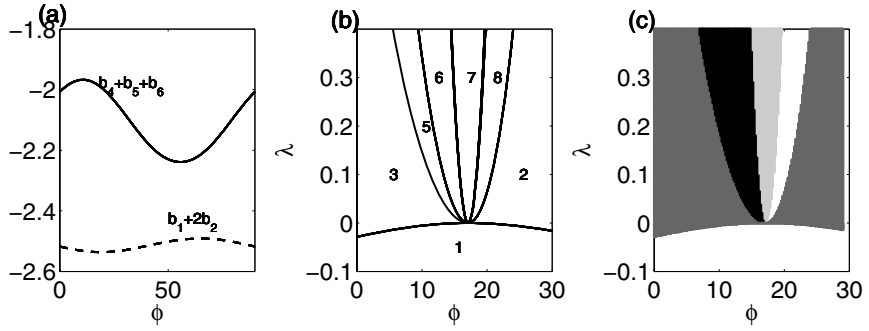


FIG. 6.9. $(M_1, M_2) = (4, 5)$, $\chi = 61°$, $\theta = 30°$: (a) *Dependence of* $b_1 + 2b_2$ *and* $b_4 + b_5 + b_6$ *on* ϕ. (b) *Bifurcation set:* 1. *trivial state stable,* 2. *stable rectangles and hexagons,* 3. *stable hexagons and superlattice patterns,* 5. *stable superlattice patterns,* 6. *stable superlattice patterns and rectangles,* 7. *stable rectangles,* 8. *no state stable.* (c) *Planforms with lowest energy as a function of* ϕ *and* λ. *Black: superlattice; dark grey: hexagons; light grey: rectangles; white (for* $\lambda > 0$*): no stable state found.*

no stable states near onset and for others the hexagons bifurcated to a pattern they called superlattice II. We cannot hope to capture this latter transition in our study since these patterns are time periodic and our amplitude equations have a gradient structure. A theoretical explanation for these patterns was given in [31]. Nevertheless, we illustrate how the phase does effect our results in Figure 6.9 for $(4, 5)$ and $\chi = 61°$. First, in Figure 6.9(a), the quantities $b_1 + 2b_2$ and $b_4 + b_5 + b_6$ are shown. These are $\pi/2$ periodic functions of ϕ. The consequence is that the regions of stability of different patterns depend on ϕ, as shown by the bifurcation set in Figure 6.9(b) and the corresponding plot of the Lyapunov function in Figure 6.9(c) for the angle $\theta = 30°$. These show that it is for phases close to 16° that stable superlattice patterns occur closest to onset. We also find values of the phase for which there are no stable spatially periodic pattern near onset (the white wedge region in Figure 6.9).

Overall, the results of Figures 6.4 through 6.8 show that the cases $(M_1, M_2) = (4, 5)$ and $(M_1, M_2) = (6, 7)$ are broadly similar. The experimental results show similar bifurcation sequences in both cases but very different forms for the planforms: in the case $(6, 7)$, superlattice patterns are observed with an easily visible regular periodic structure with two wavelengths. In [4], these are referred to as superlattice-I patterns. In the $(4, 5)$ case, quasipatterns are observed. The key difference in the two cases is the lattice angle for which stable superlattice patterns are possible. In [18], it was argued that the stabilization of the superlattice-I patterns followed from a resonant triad formed by the harmonic tongue with one of the weakly damped harmonic tongues. This was supported by a peak in the value of the rhombic coefficient (b_4 here) at around 22° for the Zhang–Vinãls model. In Figures 6.5 and 6.8, we have seen that the same mechanism operates in the full Navier–Stokes equations. Furthermore, it is the same fundamental mechanism that is at play in the $(4, 5)$ case, as shown in Figures 6.4 and 6.6, where now it is a peak in b_4 at 32.8° which is significant. This peak can again be traced to a resonant triad between the main harmonic mode and the first (that is, smallest k) weakly damped harmonic mode. The consequence is that while the analysis for the $(6, 7)$ case suggests that a superlattice pattern with angle close to 21.2°, such as that shown in Figure 6.10(a), will occur, for the $(4, 5)$ case a superlattice pattern/quasipattern with angle close to 30°, such as that shown in Figure 6.10(b), will occur.

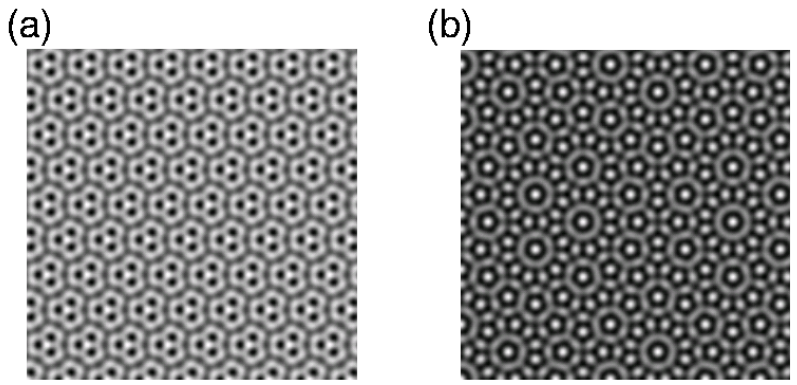


FIG. 6.10. *Typical superlattice patterns. Superlattice patterns have the form* $\mathbf{z} = (A_{SH}, A_{SH}, A_{SH}, A_{SH}, A_{SH}, A_{SH})$*; the associated planform has the form* $A_{SH}(\cos(\mathbf{K_{1_h}} \cdot \mathbf{r} + \psi) + \cos(\mathbf{K_{2_h}} \cdot \mathbf{r} + \psi) + \cos(\mathbf{K_{3_h}} \cdot \mathbf{r} + \psi) + \cos(\mathbf{K_{4_h}} \cdot \mathbf{r} + \psi) + \cos(\mathbf{K_{5_h}} \cdot \mathbf{r} + \psi) + \cos(\mathbf{K_{6_h}} \cdot \mathbf{r} + \psi))$*, where* $\mathbf{r} = (x, y)$*.* (a) *For* $\theta \approx 21.2^\circ$*,* $\psi = 2\pi/3$*.* (b) *For* $\theta \approx 30^\circ$*,* $\psi = 0$*. The superlattice pattern for* $\psi = 0$ *has a hexagonal rather than a triangular symmetry. Of these two states, only one is stable, but which one is not determined at cubic order. The experimental patterns for the* $(6, 7)$ *case clearly show a triangular structure, and it was this that motivated the theoretical work of* [27] *on the* $\psi = 2\pi/3$ *superlattice pattern.*

Note that there is a peak in the value of the coefficient b_4 for values of χ far from the bicritical point. For example, there is still a distinct peak in b_4 at approximately 21° for $\chi = 30^\circ$ in the $(4, 5)$ case. This can be related to the fact that the weakly damped harmonic tongue involved in the triad resonance is still prominent, as seen in Figure 6.1(c). However, it is only near the bicritical point that $b_4 + b_5 + b_6 - b_1 - 2b_2$ becomes positive, allowing superlattice patterns to stabilize.

7. Conclusion. In this paper, we have derived the form of the weakly nonlinear problem for a finite depth of fluid that is subject to a vertical oscillation from the full Navier–Stokes equations. Using the ideas of symmetry for patterns that tessellate the plane, we have found the coefficients of the amplitude equations and calculated the consequences for stability of different spatially periodic patterns in the infinite depth case. We have focused on the particular parameters that were used in experimental results presented in [4]. Good agreement has been found between the predictions of the weakly nonlinear analysis and the experimental results, without the use of any fitted parameters.

In the future, there are many interesting questions relating to regular patterns in the Faraday problem that we plan to use our method to explore. Our current code will allow us to perform a careful comparison between the Zhang–Viñals equations and the full Navier–Stokes equations for varying viscosity. We will also be able to investigate how the coefficients of the amplitude equations scale with the different parameters and compare with the scaling arguments given in [20] for multiple frequency forcing. This will enable us to see the degree to which the arguments for the control of patterns through multiple frequency in the weak viscosity case carry over to moderate and large viscosity. (With our formulation, it is straightforward to include more than two frequency components.)

The solvability condition is valid for both finite and infinite depth, although our subsequent calculations were carried out only for infinite depth. Our future plans also include coding up the finite depth case. This will allow us to explore how the coef-

ficients vary with depth and enable comparison with experimental results in shallow containers to be made.

Meanwhile, the Faraday wave experiment continues to be a rich source for striking and intriguing patterns, as shown by the recent results of Epstein and Fineberg [32].

Appendix A. Form for the horizontal components of the velocity. The form of the horizontal velocity components can be derived from the form of the vertical velocity and equations (2.1). Hence, the horizontal velocity components for rectangles at first order are

$$u_1 = \frac{iA_R}{k}[e^{ikx} + ce^{ik(cx+sy)} - c.c.]\sum_n DW_{1,n}(z)e^{i(n\omega+\alpha)\tau},$$
$$v_1 = \frac{isA_R}{k}[e^{ik(cx+sy)} - c.c.]\sum_n DW_{1,n}(z)e^{i(n\omega+\alpha)\tau}.$$

Horizontal velocity components for rectangles at second order are

$$\begin{aligned}
u_2 = {} & \frac{i}{2k}A_R^2[e^{2ikx} + ce^{2ik(cx+sy)} - c.c.]\sum_n DW_{2,1,n}(z)e^{i(n\omega+2\alpha)\tau} \\
& + \frac{i}{2k}A_R^2[e^{ik[(1+c)x+sy]} - c.c.]\sum_n DW_{2,2,n}(z)e^{i(n\omega+2\alpha)\tau} \\
& + \frac{i}{2k}A_R^2[e^{ik[(1-c)x-sy]} - c.c.]\sum_n DW_{2,3,n}(z)e^{i(n\omega+2\alpha)\tau}, \\
v_2 = {} & \frac{is}{2k}A_R^2[e^{2ik(cx+sy)} - c.c.]\sum_n DW_{2,1,n}(z)e^{i(n\omega+2\alpha)\tau} \\
& + \frac{is}{2k(1+c)}A_R^2[e^{ik[(1+c)x+sy]} - c.c.]\sum_n DW_{2,2,n}(z)e^{i(n\omega+2\alpha)\tau} \\
& - \frac{is}{2k(1-c)}A_R^2[e^{ik[(1-c)x-sy]} - c.c.]\sum_n DW_{2,3,n}(z)e^{i(n\omega+2\alpha)\tau}.
\end{aligned}$$

Horizontal velocity components for hexagons at first order are

$$u_1 = \frac{iA_H}{k}\left[e^{ikx} - \frac{1}{2}e^{ik(-x+\sqrt{3}y)/2} - \frac{1}{2}e^{ik(-x-\sqrt{3}y)/2} - c.c.\right]\sum_n DW_{1,n}(z)e^{i(n\omega+\alpha)\tau},$$
$$v_1 = \frac{i\sqrt{3}A_H}{2k}[e^{ik(-x+\sqrt{3}y)/2} - e^{ik(-x-\sqrt{3}y)/2} - c.c.]\sum_n DW_{1,n}(z)e^{i(n\omega+\alpha)\tau}.$$

Horizontal velocity components for hexagons at second order are

$$\begin{aligned}
u_2 = {} & \frac{iA_H^2}{2k}\left[e^{2ikx} - \frac{1}{2}e^{ik(-x+\sqrt{3}y)} - \frac{1}{2}e^{ik(-x-\sqrt{3}y)} - c.c.\right]\sum_n DW_{2,1,n}(z)e^{i(n\omega+2\alpha)\tau} \\
& + \frac{iA_H^2}{2k}[e^{ik(3x+\sqrt{3}y)/2} + e^{ik(3x-\sqrt{3}y)/2} - c.c.]\sum_n DW_{2,2,n}(z)e^{i(n\omega+2\alpha)\tau}, \\
v_2 = {} & \frac{i\sqrt{3}A_H^2}{4k}[e^{ik(-x+\sqrt{3}y)} - e^{ik(-x-\sqrt{3}y)} - c.c.]\sum_n DW_{2,1,n}(z)e^{i(n\omega+2\alpha)\tau} \\
& + \frac{iA_H^2}{\sqrt{3}k}\left[e^{ik\sqrt{3}y} + \frac{1}{2}e^{ik(3x+\sqrt{3}y)/2} - \frac{1}{2}e^{ik(3x-\sqrt{3}y)/2} - c.c.\right]\sum_n DW_{2,2,n}(z)e^{i(n\omega+2\alpha)\tau}.
\end{aligned}$$

Appendix B. Cubic coefficient for rectangles and stripes.

$$
\begin{aligned}
\gamma_i = &-\sum_{l,m,n}^{(2)} Z_l^* \left[Z_{1,m} DW_{2,i,n}(0) - 2cDW_{1,m}(0) Z_{2,i,n}\right] \\
&-\sum_{l,m,n}^{(2)} W_l^*(0) \left[W_{1,m}(0) D^2 W_{2,i,n}(0) - 2cD^2 W_{1,m}(0) W_{2,i,n}(0)\right] \\
&+2k^2 \sum_{l,m,n}^{(2)} W_l^*(0) \left[Z_{1,m} DP_{2,i,n}(0) + \frac{1}{4d} Z_{1,m} DP_{2,4,n}(0) + DP_{1,m}(0) Z_{2,i,n}\right] \\
&-2Ck^2 \sum_{l,m,n}^{(2)} W_l^*(0) \left[Z_{1,m} D^2 W_{2,i,n}(0) + 2D^2 W_{1,m}(0) Z_{2,i,n}\right] \\
&-\sum_{l,m,n}^{(2)} W_l^*(0) \left[DW_{1,m}(0) DW_{2,i,n}(0) - 4(1+c)Ck^4 Z_{1,m} W_{2,i,n}(0)\right] \\
&-C \sum_{l,m,n}^{(2)} DW_l^*(0) \left[Z_{1,m} D^3 W_{2,i,n}(0) - 2cD^3 W_{1,m}(0) Z_{2,i,n}\right] \\
&-2Ck^2 \sum_{l,m,n}^{(2)} DW_l^*(0) \left[(3+3c+c^2-s^2) DW_{1,m}(0) Z_{2,i,n} - 2cZ_{1,m} DW_{2,i,n}(0)\right] \\
&-\sum_{l,m,n,j}^{(3)} Z_l^* Z_{1,m} Z_{1,n} D^2 W_{1,j}(0) \\
&+\sum_{l,m,n,j}^{(3)} W_l^*(0) Z_{1,m} Z_{1,n} \left[Bk^6(3-2s^2) Z_{1,j} + 4Ck^4(2+c^2) DW_{1,j}(0)\right. \\
&\qquad\qquad \left. -2Ck^2 D^3 W_{1,j}(0) + 3k^2 D^2 P_{1,j}(0)\right] \\
&-C \sum_{l,m,n,j}^{(3)} DW_l^*(0) Z_{1,m} Z_{1,n} \left[(9-4s^2) k^2 D^2 W_{1,j}(0) + D^4 W_{1,j}(0)\right] \\
&-\sum_{l,m,n}^{(2)} \int_{-h/l}^{0} W_l^*(z) \left[2k^2(2+c) DW_{1,m}(z) W_{2,i,n}(z) + 3k^2 W_{1,m}(z) DW_{2,i,n}(z)\right] dz \\
&-\sum_{l,m,n}^{(2)} \int_{-h/l}^{0} W_l^*(z) \left[(2c-1) D^2 W_{1,m}(z) DW_{2,i,n}(z) - 2DW_{1,m}(z) D^2 W_{2,i,n}(z)\right] dz \\
&+\sum_{l,m,n}^{(2)} \int_{-h/l}^{0} W_l^*(z) \left[W_{1,m}(z) D^3 W_{2,i,n}(z) - 2cD^3 W_{1,m}(z) W_{2,i,n}(z))\right] dz. \qquad \text{(B.1)}
\end{aligned}
$$

The following notation has been used for the sums in (4.27), (4.28), (4.29), and (B.1):

$$
\begin{aligned}
\sum_{l,m}^{(1)} &= \begin{cases} l+m=0 \quad (\alpha=0), \\ l+m+1=0 \quad (\alpha=\omega/2), \end{cases} \\
\sum_{l,m,n}^{(2)} &= \begin{cases} l+m+n=0 \quad (\alpha=0), \\ l+m+n+2=0 \quad (\alpha=\omega/2), \end{cases} \\
\sum_{l,m,n,j}^{(3)} &= \begin{cases} l+m+n+j=0 \quad (\alpha=0), \\ l+m+n+j+2=0 \quad (\alpha=\omega/2), \end{cases} \\
\sum_{l,m}^{(4)} &= \begin{cases} l+m+1=0 \quad (\alpha=0), \\ l+m+2=0 \quad (\alpha=\omega/2), \end{cases}
\end{aligned}
$$

$$\sum_{l,m}^{(5)} = \begin{cases} l+m-1=0 \quad (\alpha=0), \\ l+m=0 \quad (\alpha=\omega/2), \end{cases}$$

$$\sum_{l,m}^{(6)} = \begin{cases} l+m+M_1=0 \quad (\alpha=0), \\ l+m+M_1+1=0 \quad (\alpha=\omega/2), \end{cases}$$

$$\sum_{l,m}^{(7)} = \begin{cases} l+m-M_1=0 \quad (\alpha=0), \\ l+m-M_1+1=0 \quad (\alpha=\omega/2), \end{cases}$$

$$\sum_{l,m}^{(8)} = \begin{cases} l+m+M_2=0 \quad (\alpha=0), \\ l+m+M_2+1=0 \quad (\alpha=\omega/2), \end{cases}$$

$$\sum_{l,m}^{(9)} = \begin{cases} l+m-M_2=0 \quad (\alpha=0), \\ l+m-M_2+1=0 \quad (\alpha=\omega/2). \end{cases}$$

Note that terms involving the derivative of the pressure with respect to z appear in the solvability condition. These may be computed from the flow variables, as discussed in Appendix D.

Appendix C. Cubic coefficient for hexagons.

$$\begin{aligned}
\gamma_3 = &\sum_{l,m,n}^{(2)} Z_l^* \left(2DW_{1,m}(0)Z_{2,2,n} - 2Z_{1,m}DW_{2,2,n}(0)\right) \\
&+ \sum_{l,m,n}^{(2)} W_l^*(0)\left(2D^2W_{1,m}(0)W_{2,2,n}(0) - 2W_{1,m}(0)D^2W_{2,2,n}(0)\right) \\
&+4k^2 \sum_{l,m,n}^{(2)} W_l^*(0)\left(DP_{1,m}(0)Z_{2,2,n}(0) + Z_{1,m}DP_{2,2,n}(0) + Z_{1,m}DP_{2,0,n}(0)\right. \\
&\qquad\qquad \left. +\frac{1}{3}Z_{1,m}DP_{2,3,n}(0)\right) \\
&-Ck^2 \sum_{l,m,n}^{(2)} W_l^*(0)\left(4Z_{1,m}D^2W_{2,2,n}(0) + 8D^2W_{1,m}(0)Z_{2,2,n}\right) \\
&- \sum_{l,m,n}^{(2)} W_l^*(0)\left(-12Ck^4Z_{1,m}W_{2,2,n}(0) + 2DW_{1,m}(0)DW_{2,2,n}(0)\right) \\
&+C \sum_{l,m,n}^{(2)} DW_l^*(0)\left(2D^3W_{1,m}(0)Z_{2,2,n} - 2Z_{1,m}D^3W_{2,2,n}(0)\right) \\
&+C \sum_{l,m,n}^{(2)} DW_l^*(0)\left(4k^2Z_{1,m}DW_{2,2,n}(0) - 16k^2DW_{1,m}(0)Z_{2,2,n}\right) \\
&-4 \sum_{l,m,n,j}^{(3)} Z_l^*Z_{1,m}Z_{1,n}D^2W_{1,j}(0) \\
&+k^2 \sum_{l,m,n,j}^{(3)} W_l^*(0)Z_{1,m}Z_{1,n}\left(12D^2P_{1,j}(0) + 36Ck^2DW_{1,j}(0) - 8CD^3W_{1,j}(0)\right. \\
&\qquad\qquad \left. +6Bk^4Z_{1,j}\right)
\end{aligned}$$

$$
\begin{aligned}
&-C\sum_{l,m,n,j}^{(3)} DW_l^*(0)Z_{1,m}Z_{1,n}\left(24k^2D^2W_{1,j}(0)+4D^4W_{1,j}(0)\right)\\
&-\sum_{l,m,n}^{(2)}\int_{-h/l}^{0} W_l^*(z)\left(10k^2DW_{1,m}(z)W_{2,2,n}(z)+6k^2W_{1,m}(z)DW_{2,2,n}(z)\right)\\
&+4\sum_{l,m,n}^{(2)}\int_{-h/l}^{0} W_l^*(z)DW_{1,m}(z)D^2W_{2,2,n}(z)\\
&+\sum_{l,m,n}^{(2)}\int_{-h/l}^{0} W_l^*(z)\left(2W_{1,m}(z)D^3W_{2,2,n}(z)-2D^3W_{1,m}(0)W_{2,2,n}(z)\right). \qquad \text{(C.1)}
\end{aligned}
$$

Appendix D. A posteriori computation of the pressure. We compute the pressure from the Navier–Stokes equations. As we need only the derivative of the pressure with respect to z, we need only consider the third component of the Navier–Stokes equations:

$$\partial_z p = -\partial_t w + C\Delta w - \mathbf{u}\cdot\nabla w. \qquad \text{(D.1)}$$

At the first order this reduces to

$$\partial_z p_1 = -\partial_\tau w_1 + C\Delta w_1, \qquad \text{(D.2)}$$

and then we can take $\partial_z p_1$ in the form

$$\partial_z p_1 = A_S(e^{ikx}+e^{-ikx})\sum_n DP_{1,n}(z)e^{i(n\omega+\alpha)\tau}, \qquad \text{(D.3)}$$

or the equivalent form for rectangles and hexagons, where $DP_{1,n}(z)$ solves

$$DP_{1,n}(z) = [-i(n\omega+\alpha)+C(D^2-k^2)]W_{1,n}(z). \qquad \text{(D.4)}$$

At the second order we have

$$\partial_z p_2 = -\partial_\tau w_2 + C\Delta w_2 - \mathbf{u}_1\cdot\nabla w_1 - \partial_{T_1} w_1. \qquad \text{(D.5)}$$

The solution for the rectangular pattern does not depend on T_1. Here we take $\partial_z p_2$ in the form

$$
\begin{aligned}
\partial_z p_2 = {}& A_R^2[e^{2ikx}+e^{2ik(cx+sy)}+c.c.]\sum_n DP_{2,1,n}(z)e^{i(n\omega+2\alpha)\tau}\\
&+A_R^2[e^{ik[(1+c)x+sy]}+c.c.]\sum_n DP_{2,2,n}(z)e^{i(n\omega+2\alpha)\tau}\\
&+A_R^2[e^{ik[(1-c)x-sy]}+c.c.]\sum_n DP_{2,3,n}(z)e^{i(n\omega+2\alpha)\tau}\\
&+A_R^2\sum_n DP_{2,4,n}(z)e^{i(n\omega+2\alpha)\tau},
\end{aligned}
$$

where

$$DP_{2,1,n}(z) = [-i(n\omega+2\alpha)+C(D^2-4k^2)]W_{2,1,n}(z),$$

$$
\begin{aligned}
DP_{2,2,n}(z) &= [-i(n\omega+2\alpha)+C(D^2-2(1+c)k^2)]W_{2,2,n}(z)\\
&\quad -2(1-c)\sum_{(l+m=n)} W_{1,l}(z)DW_{1,m}(z),\\
DP_{2,3,n}(z) &= [-i(n\omega+2\alpha)+C(D^2-2(1-c)k^2)]W_{2,3,n}(z)\\
&\quad -2(1+c)\sum_{(l+m=n)} W_{1,l}(z)DW_{1,m}(z),\\
DP_{2,4,n}(z) &= -8\sum_{(l+m=n)} W_{1,l}(z)DW_{1,m}(z).
\end{aligned}
$$

In the hexagonal pattern, the solution depends on both T_1 and T_2. We take the pressure as the sum of two terms:

$$\partial_z p_2 = \partial_z \hat{p}_2 + \partial_z \tilde{p}_2,$$

where $\hat{p}_z$ solves

$$\partial_z \hat{p}_2 = -\partial_\tau w_2 + C\Delta w_2 - \mathbf{u}_1\cdot\nabla w_1, \tag{D.6}$$

while $\tilde{p}_z = -\partial_{T_1} w_1$. We take $\hat{p}_z$ in the form of a hexagonal pattern:

$$
\begin{aligned}
\partial_z \hat{p}_2 &= A_H^2[e^{ikx}+e^{ik(-x+\sqrt{3}y)/2}+e^{ik(-x-\sqrt{3}y)/2}+c.c.]\sum_n DP_{2,0,n}(z)e^{i(n\omega+2\alpha)\tau}\\
&\quad +A_H^2[e^{2ikx}+e^{ik(-x+\sqrt{3}y)}+e^{ik(-x-\sqrt{3}y)}+c.c.]\sum_n DP_{2,1,n}(z)e^{i(n\omega+2\alpha)\tau}\\
&\quad +A_H^2[e^{ik\sqrt{3}y}+e^{ik(3x+\sqrt{3}y)/2}+e^{ik(3x-\sqrt{3}y)/2}+c.c.]\sum_n DP_{2,2,n}(z)e^{i(n\omega+2\alpha)\tau}\\
&\quad +A_H^2\sum_n DP_{2,3,n}(z)e^{i(n\omega+2\alpha)\tau},
\end{aligned}
$$

where

$$
\begin{aligned}
DP_{2,0,n}(z) &= -3\sum_{(l+m=n)} W_{1,l}(z)DW_{1,m}(z),\\
DP_{2,1,n}(z) &= [-i(n\omega+2\alpha)+C(D^2-4k^2)]W_{2,1,n}(z),\\
DP_{2,2,n}(z) &= [-i(n\omega+2\alpha)+C(D^2-3k^2)]W_{2,2,n}(z)\\
&\quad -\sum_{(l+m=n)} W_{1,l}(z)DW_{1,m}(z),\\
DP_{2,3,n}(z) &= -12\sum_{(l+m=n)} W_{1,l}(z)DW_{1,m}(z).
\end{aligned}
$$

Acknowledgments. We have benefited from discussions with a number of people during the course of writing this paper, including Paul Matthews, Alastair Rucklidge, and José Vega. We are very grateful to Peilong Chen for letting us use his code to compare our work with his and to the anonymous referees, who made a number of helpful suggestions.

REFERENCES

[1] M. Faraday, *On the forms and states assumed by fluids in contact with vibrating elastic surfaces*, Phil. Trans. Roy. Soc. Lond., 121 (1831), pp. 319–340.

[2] S. Ciliberto and J. P. Gollub, *Chaotic mode competition in parametrically forced surface waves*, J. Fluid Mech., 158 (1985), pp. 381–398.
[3] W. S. Edwards and S. Fauve, *Patterns and quasi-patterns in the Faraday experiment*, J. Fluid Mech., 278 (1994), pp. 123–148.
[4] A. Kudrolli, B. Pier, and J. P. Gollub, *Superlattice patterns in surface waves*, Phys. D, 123 (1998), pp. 99–111.
[5] H. Arbell and J. Fineberg, *Spatial and temporal dynamics of two interacting modes in parametrically driven surface waves*, Phys. Rev. Lett., 81 (1998), pp. 4384–4387.
[6] H. Arbell and J. Fineberg, *Two-mode rhomboidal states in driven surface waves*, Phys. Rev. Lett., 84 (2000), pp. 654–657.
[7] H. Arbell and J. Fineberg, *Temporally harmonic oscillons in Newtonian fluids*, Phys. Rev. Lett., 85 (2000), pp. 756–759.
[8] D. Binks and W. van der Water, *Nonlinear pattern formation of Faraday waves*, Phys. Rev. Lett., 78 (1997), pp. 4043–4046.
[9] P. H. Wright and J. R. Saylor, *Patterning of particulate films using Faraday waves*, Rev. Sci. Instrum., 74 (2003), pp. 4063–4070.
[10] J. R. Saylor and A. L. Kinard, *Simulation of particle deposition beneath Faraday waves in thin liquid films*, Phys. Fluids, 17 (2005), 047107.
[11] J. R. Saylor and R. A. Handler, *Gas transport across an air/water interface populated with capillary waves*, Phys. Fluids, 9 (1997), pp. 2529–2541.
[12] T. G. Leighton, *From sea to surgeries, from babbling brooks to baby scans: Bubble acoustics at ISVR*, Proc. Inst. Acoustics, 26 (2004), pp. 357–381.
[13] T. B. Benjamin and F. Ursell, *The stability of a plane free surface of a liquid in vertical periodic motion*, Proc. Roy. Soc. London Ser. A, 225 (1954), pp. 505–515.
[14] K. Kumar and L. Tuckerman, *Parametric instability of the interface between two fluids*, J. Fluid Mech., 279 (1994), pp. 49–68.
[15] T. Besson, W. S. Edwards, and L. S. Tuckerman, *Two-frequency parametric excitation of surface waves*, Phys. Rev. E (3), 54 (1996), pp. 507–513.
[16] W. Zhang and J. Viñals, *Pattern formation in weakly damped parametric surface waves*, J. Fluid Mech., 336 (1997), pp. 301–330.
[17] W. Zhang and J. Viñals, *Square patterns and quasipatterns in weakly damped Faraday waves*, Phys. Rev. E (3), 53 (1996), pp. R4283–R4286.
[18] M. Silber, C. Topaz, and A. C. Skeldon, *Two-frequency forced Faraday waves: Weakly damped modes and pattern selection*, Phys. D, 143 (2000), pp. 205–225.
[19] C. M. Topaz and M. Silber, *Resonances and superlattice pattern stabilization in two-frequency forced Faraday waves*, Phys. D, 172 (2002), pp. 1–29.
[20] J. Porter and M. Silber, *Broken symmetries and pattern formation in two-frequency forced Faraday waves*, Phys. Rev. Lett., 89 (2002), 084051.
[21] C. Martel and E. Knobloch, *Damping of nearly inviscid water waves*, Phys. Rev. E (3), 56 (1997), pp. 5544–5548.
[22] P. Chen and J. Viñals, *Amplitude equation and pattern selection in Faraday waves*, Phys. Rev. E (3), 60 (1999), pp. 559–570.
[23] F. J. Mancebo and J. M. Vega, *Viscous Faraday waves in two-dimensional large-aspect-ratio containers*, J. Fluid Mech., 560 (2006), pp. 369–393.
[24] A. M. Rucklidge, *Pattern formation in large domains*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci., 361 (2003), pp. 2649–2664.
[25] B. Dionne and M. Golubitsky, *Planforms in two and three dimensions*, Z. Angew. Math. Phys., 43 (1992), pp. 36–62.
[26] B. Dionne, M. Silber, and A. C. Skeldon, *Stability results for steady, spatially periodic planforms*, Nonlinearity, 10 (1997), pp. 321–353.
[27] M. Silber and M. R. E. Proctor, *Nonlinear competition between small and large hexagonal patterns*, Phys. Rev. Lett., 81 (1998), pp. 2450–2453.
[28] A. C. Skeldon and M. Silber, *New stability results for patterns in a model of long-wavelength convection*, Phys. D, 122 (1998), pp. 117–133.
[29] S. L. Judd and M. Silber, *Simple and superlattice Turing patterns in reaction-diffusion systems: Bifurcation, bistability and parameter collapse*, Phys. D, 136 (2000), pp. 45–65.
[30] M. Golubitsky, J. W. Swift, and E. Knobloch, *Symmetries and pattern selection in Rayleigh-Bénard convection*, Phys. D, 10 (1984), pp. 249–276.
[31] D. Tse, A. M. Rucklidge, R. Hoyle, and M. Silber, *Spatial period-multiplying instabilities of hexagonal Faraday waves*, Phys. D, 146 (2000), pp. 367–387.
[32] T. Epstein and J. Fineberg, *Grid states and nonlinear selection in parametrically excited surface waves*, Phys. Rev. E (3), 73 (2006), 055302.

SIAM J. APPL. MATH.
Vol. 67, No. 4, pp. 1101–1123

LOCATING TRANSPARENT REGIONS IN OPTICAL ABSORPTION AND SCATTERING TOMOGRAPHY*

NUUTTI HYVÖNEN†

Abstract. The aim of optical absorption and scattering tomography is to reconstruct the optical properties inside a physical body, e.g., a neonatal head, by illuminating it with near-infrared light and measuring the outward flux of photons on the object boundary. Because brain consists of strongly scattering tissue with imbedded cavities filled by weakly scattering cerebrospinal fluid, propagation of near-infrared photons in the human head can be treated by combining the diffusion approximation of the radiative transfer equation with geometrical optics to obtain the radiosity-diffusion forward model of optical tomography. Currently, a disadvantage with the radiosity-diffusion model is that the locations of the transparent cavities must be known in advance in order to be able to reconstruct the physiologically interesting quantities, i.e., the absorption and the scatterer in the strongly scattering brain tissue. In this work we show, both theoretically and numerically, that under suitable conditions the factorization method of Andreas Kirsch can be used for locating the transparent cavities through the boundary measurements of optical tomography if the background optical properties of the strongly scattering tissue are known.

Key words. optical absorption and scattering tomography, inverse boundary value problems, factorization method, nonscattering regions, transparent regions, inclusions

AMS subject classifications. 35R30, 35J25, 35R05

DOI. 10.1137/06066299X

1. Introduction.

In optical absorption and scattering tomography (OAST), a physical body is illuminated with a flux of near-infrared (NIR) photons and the outward flux is measured on the surface of the body. The idea is to reconstruct the optical properties, such as absorption and scatter, inside the body by using the measured pairs of input and output fluxes. OAST has a few possible clinical applications, the most important of which are, arguably, screening for breast cancer and the development of a cerebral imaging modality for mapping structure and function in newborn infants, and possibly adults too. For more medical and instrumental details we refer to the articles [2, 3, 5, 14, 16].

In a strongly scattering medium, e.g., brain tissue, propagation of NIR photons can be modeled to a good extent by the diffusion approximation of the radiative transfer equation (RTE) [2]. Since the diffusion approximation is not valid in weakly scattering regions [4, 11], e.g., in cavities that are filled with nearly transparent cerebrospinal fluid, some other approximation of the RTE is also needed when building up the forward model of OAST for the human head. By combining the diffusion approximation with geometrical optics, one obtains the radiosity-diffusion forward model [19, 29], which takes into account the effect of the weakly scattering regions.

As noted in [14], a disadvantage with the current implementation of the radiosity-diffusion model is that the boundaries of the transparent cavities must be known in advance when solving the actual inverse problem of OAST, i.e., reconstructing the absorption and the scatter in the strongly scattering tissue. If an anatomical

*Received by the editors June 15, 2006; accepted for publication (in revised form) January 3, 2007; published electronically May 14, 2007. This work was supported by the Finnish Funding Agency for Technology and Innovation (project 40084/06) and the Academy of Finland (project 115013).
http://www.siam.org/journals/siap/67-4/66299.html
†Institute of Mathematics, Helsinki University of Technology, PO Box 1100, FI-02015 HUT, Finland (nuutti.hyvonen@hut.fi).

magnetic resonance image is available, it is possible to segment the head into diffusive and transparent regions and proceed by reconstructing the optical properties in the diffusive region within the framework of the radiosity-diffusion model [14]. However, if there is no such a priori information on the locations of the transparent cavities, the natural way to solve the inverse problem of OAST is to locate the cavities and reconstruct the optical properties of the diffusive region simultaneously. In this work we tackle a preliminary simplified version of this inverse problem: We assume that the absorption and the scatter in the diffusive region of the examined body are known and try locate the transparent regions through boundary measurements. Notice that this is not an easy task for a state-of-the-art iterative Newton-type algorithm based on the output least squares formulation of the inverse problem (cf. [2, 14]) since differentiating the measurement mappings with respect to the shape of the transparent region is difficult (cf. [19, 29]).

The factorization method, introduced originally within inverse obstacle scattering by Kirsch [25], provides a tool that can be used for locating inhomogeneities in a diffusive background in noniterative fashion. The factorization technique has already been applied to electrical impedance tomography in [7, 8, 20] and to OAST with strongly scattering inclusions in [21, 23] and with thin transparent layers in [6] (see also [13]). In this work, we will use the factorization method to obtain a conditional characterization of the transparent cavities via boundary measurements. Based on this theoretical work, we will formulate a reconstruction algorithm and test it numerically with two-dimensional simulated data.

This text is organized as follows. In section 2, we introduce the radiosity-diffusion model. Since our formulation differs slightly from the material in some of the references, we do not use mere citations but provide the most essential details, as well. Section 3 introduces and proves the conditional characterization result. In section 4, we interpret the theoretical work of section 3 as a reconstruction algorithm and test it numerically with simulated data. Section 5 contains some concluding remarks.

2. Radiosity-diffusion model. Propagation of electromagnetic radiation in a medium is governed by Maxwell's equations. In particular, this holds for the case of interest to us, namely, NIR light traveling through some biological tissue. However, since the radiation within a strongly scattering medium is completely incoherent and the wavelength of NIR light is small compared to the characteristic distances of human tissue, the exact models are totally useless. Therefore, we will model light propagation by using approximations of the radiative transfer equation, also known as the Boltzmann equation. Because the human brain consists of strongly scattering tissue with weakly scattering cavities filled by cerebrospinal fluid [3, 28], our aim is to treat these two extremes separately and then bundle the models together to obtain the so-called radiosity-diffusion forward problem [3, 11, 12].

We begin our work with a short glance at transport theory. Let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, be a bounded body with a smooth enough boundary. The radiation flux density at $x \in \Omega$ at time $t \in \mathbb{R}$ to the infinitesimal solid angle ds in direction $\hat{\theta} \in S^{n-1}$ is written as

$$d\vec{J}(x, t, \hat{\theta}) = I(x, t, \hat{\theta})\hat{\theta} ds(\hat{\theta}),$$

where the amplitude $I(x, t, \hat{\theta})$ is called the radiance. In the framework of transport theory, this scalar function satisfies the RTE

$$\frac{1}{c}I_t(x,t,\hat{\theta}) + \hat{\theta}\cdot\nabla I(x,t,\hat{\theta}) + (\mu_a(x)+\mu_s(x))I(x,t,\hat{\theta}) - \mu_s(x)\int_{S^{n-1}} f(x,\hat{\theta},\hat{\omega})I(x,t,\hat{\omega})ds(\hat{\omega}) = q(x,t,\hat{\theta}), \tag{2.1}$$

where c is the speed of light (assumed to be constant), the positive scalar functions μ_a and μ_s are the absorption and scattering coefficients, respectively, and q denotes the source term, which is assumed to vanish in this discussion. The kernel f is the scattering phase function, satisfying the following three conditions:

$$\int_{S^{n-1}} f(x,\hat{\theta},\hat{\omega})ds(\hat{\theta}) = \int_{S^{n-1}} f(x,\hat{\theta},\hat{\omega})ds(\hat{\omega}) = 1,$$
$$f(x,\hat{\theta},\hat{\omega}) \geq 0, \qquad x\in\mathbb{R}^n, \quad \hat{\theta},\hat{\omega}\in S^{n-1},$$
$$f(x,\hat{\theta},\hat{\omega}) = f(x,-\hat{\omega},-\hat{\theta}), \quad \hat{\theta},\hat{\omega}\in S^{n-1}.$$

Due to the first two conditions, for fixed x, f may be regarded as a probability distribution on S^{n-1} with respect to either of the variables $\hat{\theta}$ and $\hat{\omega}$.

The net photon flux through an infinitesimal oriented surface patch $\nu(x)dS(x)$ is obtained by integrating the flux density over all radiation directions,

$$d\Phi(x,t) = \left(\int_{S^{n-1}} d\vec{J}(x,t,\hat{\theta})\right)\cdot\nu(x)dS(x) = \vec{J}(x,t)\cdot\nu(x)dS(x), \tag{2.2}$$

where the vector field $\vec{J}$ is the energy current density. In a similar manner, the outward $d\Phi_+$ and the inward flux $d\Phi_-$ through νdS can be computed by choosing the domain of integration in (2.2) to be the positive $\{\hat{\theta}\in S^{n-1} \mid \hat{\theta}\cdot\nu > 0\}$ or the negative hemisphere $\{\hat{\theta}\in S^{n-1} \mid \hat{\theta}\cdot\nu < 0\}$, respectively. Furthermore, the scalar function

$$\varphi(x,t) = \int_{S^{n-1}} I(x,t,\hat{\theta})ds(\hat{\theta})$$

is the photon density. Note that $\varphi(x,t)$ and $\vec{J}(x,t)$ are essentially the coefficients of the zeroth- and first order terms for the linearization of $I(x,t,\hat{\theta})$ with respect to $\hat{\theta}$. For more transport theory the reader may consult, e.g., [9].

2.1. Strong scattering. Being an integro-differential equation, the radiative transfer equation, as discussed above, can easily lead to numerical problems of prohibitive size if no simplifications are made. The commonly used simplification is called the diffusion approximation, which has been shown to be justified for materials that are much more scattering than absorbing [14, 17].

Let $P : L^2(S^{n-1}) \to \text{span}\{1,\theta_1,\ldots,\theta_n\}$ be an orthogonal projection which linearizes the dependence on the components of the scattering direction. Denoting the integro-differential operator induced by the left-hand side of (2.1) by $\mathcal{B}$, we define the diffusion approximation of the radiative transfer equation as

$$P\mathcal{B}PI = 0, \tag{2.3}$$

where I denotes the radiance. Due to the way that the projection P is defined, one should be able to write the diffusion approximation using only the photon density φ and the energy current density $\vec{J}$ defined above. Indeed, by a straightforward

calculation [3, 18], one sees that (2.3) is equivalent to the coupled system

$$\frac{1}{c}\varphi_t = -\nabla \cdot \vec{J} - \mu_a \varphi, \tag{2.4}$$

$$\frac{1}{c}\vec{J}_t = -\frac{1}{n}\nabla\varphi - (\mu_a I + (I-B)\mu_s)\vec{J}, \tag{2.5}$$

where $I \in \mathbb{R}^{n\times n}$ is the identity matrix and the symmetric matrix $B \in \mathbb{R}^{n\times n}$ is defined by

$$B_{jk} = \frac{n}{|S^{n-1}|}\int_{S^{n-1}}\int_{S^{n-1}} \theta_j \omega_k f(x,\hat{\theta},\hat{\omega}) ds(\hat{\theta}) ds(\hat{\omega}).$$

In order to be able to handle the boundary conditions corresponding to the diffusion approximation, we write out the total flux inwards ($-$) and outwards ($+$) on the boundary $\partial\Omega$ when the dependence on the scattering direction is linearized [18]:

$$\Phi_\pm = \left(\pm\gamma\varphi + \frac{1}{2}\nu\cdot\vec{J}\right)\Big|_{\partial\Omega}, \tag{2.6}$$

where $\nu = \nu(x)$ is the exterior unit normal of $\partial\Omega$, in two dimensions $\gamma = 1/\pi$ and in three dimensions $\gamma = 1/4$. Note that the expression for the fluxes $\Phi_\pm$ differs somewhat from the one given in most references (cf. [29]). However, it is carefully deduced from the mathematical model described above, and so it is one reasonable choice.

2.2. Transparent regions. In weakly scattering regions, the diffusion approximation ceases to be valid [11, 12], and so we will have to come up with something different. Let $D \subset \Omega$, with $\partial D \cap \partial\Omega = \emptyset$, be a nonscattering region with a C^2-boundary. Because in a nonscattering, or transparent, region all radiation is in the forward direction, in D equation (2.1) yields the relation

$$\frac{1}{c}I_t(x,t,\hat{\theta}) + \hat{\theta}\cdot\nabla I(x,t,\hat{\theta}) + \tilde{\mu}_a I(x,t,\hat{\theta}) = 0,$$

where $\tilde{\mu}_a > 0$ is the absorption coefficient that is assumed to be constant in D. For the time-harmonic case $I(x,t,\hat{\theta}) = \hat{I}(x,\hat{\theta})e^{-i\omega t}$, we have

$$(\tilde{\mu}_a - ik)\hat{I} + \hat{\theta}\cdot\nabla\hat{I} = 0,$$

yielding an attenuated plane wave solution

$$\hat{I} \sim e^{-(\tilde{\mu}_a - ik)\hat{\theta}\cdot x},$$

where $k = \omega/c$.

Let $x \in \partial D$ be a boundary point of the nonscattering region. Denote the unit normal vector of ∂D pointing into D by $\nu = \nu(x)$, and let $\hat{\theta} \in S^{n-1}$ satisfy $\hat{\theta}\cdot\nu(x) < 0$. Let $y(\hat{\theta}) \in \partial D$ be the first boundary point where the line emanating from x into direction $-\hat{\theta}$ hits the boundary ∂D. Since the radiation propagates with no scattering, the contribution from the direction $-\hat{\theta}$ to the time-harmonic amplitude of the radiation flux density at x is

$$d\vec{J}(x,\hat{\theta}) = \hat{\theta}\hat{I}(y,\hat{\theta})e^{-(\tilde{\mu}_a - ik)\hat{\theta}\cdot(x-y)} ds(\hat{\theta}).$$

Using this expression and assuming that the total flux into the nonscattering region distributes uniformly to all directions, i.e., $\hat{I}(y,\hat{\theta}) = \hat{I}_0(y)$, through a straightforward

geometrical consideration one obtains the following dependence between the time-harmonic amplitudes of the total fluxes in (+) and out (−) of the nonscattering region (for details, see [19]):

$$\begin{aligned}\Phi_-(x) &= \frac{n-1}{|S^{n-2}|}\int_{\partial D} v(x,y)\frac{(\nu(x)\cdot(x-y))(\nu(y)\cdot(x-y))}{|x-y|^{n+1}}\\ &\quad\times e^{-(\tilde{\mu}_a - ik)|x-y|}\Phi_+(y)dS(y)\\ &= (\mathcal{G}\Phi_+)(x),\end{aligned} \tag{2.7}$$

where $v(\cdot,\cdot)$ is a visibility function,

$$v(x,y) = \begin{cases} 1 & \text{if } tx+(1-t)y \in D \text{ for } 0<t<1,\\ 0 & \text{otherwise.}\end{cases} \tag{2.8}$$

Thus, we have obtained a relation between the inward and outward fluxes on the boundary of the nonscattering region, which gives us the means to handle the transparent regions using nonlocal boundary conditions.

The following lemma summarizes a few essential properties of $\mathcal{G}$ [19, 22].

LEMMA 2.1. *The linear integral operator $\mathcal{G} : L^2(\partial D) \to L^2(\partial D)$ is compact and*

$$\|\mathcal{G}\|_{L^2(\partial D)\to L^2(\partial D)} < 1.$$

In particular, $I - \mathcal{G} : L^2(\partial D) \to L^2(\partial D)$ is invertible. Furthermore, if $k = 0$, $\mathcal{G}$ is self-adjoint.

2.3. Forward problem. Let us consider the time-harmonic radiosity-diffusion forward problem of optical tomography in a bounded domain Ω consisting of a nonscattering open region D, with $\partial\Omega \cap \partial D = \emptyset$, and a strongly scattering region $\Omega \setminus \overline{D}$. Assume that the time-harmonic flux $\Phi_{\text{in}}(x)e^{-i\omega t}$ is conducted through the object boundary $\partial\Omega$. By solving (2.5) for the time-harmonic amplitude of the energy current density and substituting into (2.4), we see that the time-harmonic amplitude of the photon density (still denoted by φ) satisfies the equation

$$\nabla\cdot\kappa\nabla\varphi + (ik - \mu_a)\varphi = 0 \quad \text{in } \Omega\setminus\overline{D}, \tag{2.9}$$

where $k = \omega/c$ and

$$\kappa = \frac{1}{n}((\mu_a - ik)I + (I-B)\mu_s)^{-1}.$$

Further, by using identity (2.6), together with (2.5) and (2.7), one obtains the outer boundary condition

$$\gamma\varphi + \frac{1}{2}\nu\cdot\kappa\nabla\varphi = \Phi_{\text{in}} \quad \text{on } \partial\Omega, \tag{2.10}$$

where the sign of Φ_{in} is inverted for convenience, and the nonlocal inner boundary condition

$$G\varphi + \nu\cdot\kappa\nabla\varphi = 0 \quad \text{on } \partial D. \tag{2.11}$$

In (2.11), we have used the shorthand notation

$$G = 2\gamma(I-\mathcal{G})^{-1}(I+\mathcal{G}), \tag{2.12}$$

where $(I-\mathcal{G})^{-1}: L^2(\partial D) \to L^2(\partial D)$ denotes the $L^2(\partial D)$-inverse of $I-\mathcal{G}$. Notice that $G: L^2(\partial D) \to L^2(\partial D)$ is self-adjoint if $k=0$; this can be seen, for example, by expanding $(I-\mathcal{G})^{-1}$ as a Neumann series and using the self-adjointness of $\mathcal{G}$. In all above formulae and in what follows, the normal vectors point out of the strongly scattering region $\Omega \setminus \overline{D}$.

In [19] and [22], it has been shown that the radiosity-diffusion forward problem, obtained as a combination of (2.9), (2.10), and (2.11), has a unique solution under physically reasonable conditions, as follows.

THEOREM 2.2. *Assume that* $0 < c_a \le \mu_a \le C_a$, $0 \le \mu_s \le C_s$, $\tilde{\mu}_a > 0$, *and that* $\partial\Omega$ *and* ∂D *are smooth enough. Then* κ *is well defined and positive definite, and the time-harmonic radiosity-diffusion forward problem has a unique weak solution* $\varphi \in H^1(\Omega \setminus \overline{D})$ *for any input flux* $\Phi_{\text{in}} \in H^{-1/2}(\partial\Omega)$.

2.4. Inverse problem. According to the above derived mathematical model and under the assumption of time-harmonicity, to know all pairs of inward and outward photon fluxes on the object boundary $\partial\Omega$ is equivalent to knowing the Robin-to-Robin boundary map

$$\Upsilon : \Phi_{\text{in}} \mapsto \left(\gamma\varphi - \frac{1}{2}\nu \cdot \kappa\nabla\varphi\right)\Big|_{\partial\Omega},$$

where φ is the solution of (2.9) with the boundary conditions (2.10) and (2.11). If the assumptions of Theorem 2.2 hold, it is easy to see that Υ is a bounded linear operator from $H^{-1/2}(\partial\Omega)$ to itself (cf. [21]). The idealized time-harmonic inverse problem of OAST is to determine μ_a and κ from the knowledge of Υ.

Since collecting all Robin–Robin boundary value pairs is in a pure mathematical sense equivalent to collecting all Neumann–Dirichlet pairs, besides Υ we may assume to know the Neumann-to-Dirichlet boundary map

$$\Lambda : f \mapsto \varphi|_{\partial\Omega},$$

where $\varphi \in H^1(\Omega \setminus \overline{D})$ is the solution of (2.9) with the boundary conditions (2.11) and

$$\nu \cdot \kappa\nabla\varphi = f \quad \text{on } \partial\Omega.$$

It is easy to see that Λ is an isomorphism from $H^{-1/2}(\partial\Omega)$ to $H^{1/2}(\partial\Omega)$. When implementing the factorization method to the framework of OAST and transparent inclusions, we will work with Λ instead of Υ for the sake of convenience and readability.

In earlier work [11], it has been demonstrated that reconstructing the optical parameters inside $\Omega \setminus \overline{D}$ is practically impossible if the shape of the nonscattering region D is not known in advance. On the other hand, locating ∂D via boundary measurements by using some Newton-type iterative algorithm seems far-fetched due to the difficulties encountered when differentiating the boundary operator $\mathcal{G}$ with respect to the shape of D: It is easy to see that the differentiation results in awkward formulae even if the effect of the irregular visibility function (2.8) is not taken into account. In consequence, introducing noniterative algorithms for finding D is of great importance for the further development of OAST.

In the following section, we will assume that the absorption coefficient μ_a and the diffusion tensor κ inside Ω are known but the whereabouts of the possible nonscattering region D is unknown. We will show that under such circumstances there is reason to believe that the factorization method of Kirsch [25] can be used for locating D via boundary measurements. To be more precise, we will formulate and prove a

conditional characterization result. In section 4, the functionality of the method will be confirmed through numerical studies.

3. Factorization method and transparent regions. We begin by summarizing our framework. Let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, be a bounded body with a connected complement and $D \subset \Omega$, with $\partial D \cap \partial\Omega = \emptyset$, a nonscattering region for which $\Omega \setminus \overline{D}$ is connected. For simplicity, we assume that the measurements are static in time, i.e., $k = 0$ in the formulae of the preceding section. Furthermore, assume that the a priori known absorption coefficient $\mu_a : \Omega \to \mathbb{R}$ and symmetric diffusion tensor $\kappa : \Omega \to \mathbb{R}^{n\times n}$ are smooth enough and satisfy the the following conditions:

$$c_a \le \mu_a \le C_a, \quad c_\kappa I \le \kappa \le C_\kappa I, \tag{3.1}$$

where c_a, c_κ, C_a, and C_κ are positive constants and the latter inequality is to be understood in the sense of positive definiteness.

We denote the Neumann-to-Dirichlet map corresponding to the object containing the transparent region by Λ, i.e.,

$$\Lambda : f \mapsto \varphi|_{\partial\Omega}, \quad H^{-1/2}(\partial\Omega) \to H^{1/2}(\partial\Omega), \tag{3.2}$$

where the photon density $\varphi \in H^1(\Omega \setminus \overline{D})$ satisfies the elliptic boundary value problem

$$\begin{aligned} \nabla \cdot \kappa\nabla\varphi - \mu_a\varphi &= 0 && \text{in } \Omega \setminus \overline{D}, \\ \nu \cdot \kappa\nabla\varphi &= f && \text{on } \partial\Omega, \\ G\varphi + \nu \cdot \kappa\nabla\varphi &= 0 && \text{on } \partial D, \end{aligned} \tag{3.3}$$

and the operators G and $\mathcal{G}$ are defined by (2.12) and (2.7), respectively, with $k = 0$ and a constant $\tilde{\mu}_a > 0$. Moreover, let Λ_0 be the Neumann-to-Dirichlet map corresponding to the same object without any nonscattering regions, meaning that Λ_0 is also defined by (3.2) but this time the first equation of (3.3) is satisfied everywhere in Ω and the inner boundary condition of (3.3) is deleted. Since μ_a and κ are assumed to be known in advance, Λ_0 can be computed and Λ—or at least a noisy incomplete version of Λ—can be measured. Notice that both Λ and Λ_0 are self-adjoint (cf. [7, 19]), i.e.,

$$\langle f_1, \Lambda f_2\rangle_{L^2(\partial\Omega)} = \overline{\langle f_2, \Lambda f_1\rangle}_{L^2(\partial\Omega)} \quad \text{and} \quad \langle f_1, \Lambda_0 f_2\rangle_{L^2(\partial\Omega)} = \overline{\langle f_2, \Lambda_0 f_1\rangle}_{L^2(\partial\Omega)}$$

for all $f_1, f_2 \in H^{-1/2}(\partial\Omega)$. Here and in what follows, we denote by $\langle f, g\rangle_{L^2(\partial\Omega)} = \int_{\partial\Omega} f\bar{g}dS$ the inner product of $L^2(\partial\Omega)$ as well as its extension to the dual system $\langle H^{-1/2}(\partial\Omega), H^{1/2}(\partial\Omega)\rangle$.

Before we can formulate the conditional characterization result, a couple of auxiliary concepts need to be introduced. Let h_y be the photon density corresponding to a point source at $y \in \Omega$, no transparent region, and a homogeneous Neumann condition on $\partial\Omega$; i.e., h_y is the solution of

$$\begin{aligned} \nabla \cdot \kappa\nabla h_y(x) - \mu_a h_y(x) &= \delta(x - y) && \text{in } \Omega, \\ \nu \cdot \kappa\nabla h_y &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{3.4}$$

where δ is the delta distribution. Furthermore, let $\Lambda_D^{-1} : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ be the Dirichlet-to-Neumann boundary map corresponding to D and the diffusion equation. To be more precise,

$$\Lambda_D^{-1} g = \nu \cdot \kappa\nabla u|_{\partial D},$$

where ν is the unit normal pointing into D and u satisfies the Dirichlet boundary value problem

$$\begin{aligned} \nabla \cdot \kappa \nabla u - \mu_a u &= 0 \qquad \text{in } D, \\ u &= g \qquad \text{on } \partial D. \end{aligned} \tag{3.5}$$

THEOREM 3.1. *Assume that $\Lambda_D^{-1} + G : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ is injective. Then, $h_y|_{\partial\Omega}$ belongs to the range of $|\Lambda - \Lambda_0|^{1/2}$ if and only if $y \in D$. Here, $\Lambda - \Lambda_0$ is interpreted as a compact operator from $L^2(\partial\Omega)$ to itself and $|\Lambda - \Lambda_0| = \{(\Lambda - \Lambda_0)^2\}^{1/2}$.*

The proof of Theorem 3.1 will be presented in the following subsection. Meanwhile, we will concentrate on the implications of the theorem itself. Since Λ can be measured and Λ_0 and $h_y|_{\partial\Omega}$ can be computed, Theorem 3.1 provides an explicit way to locate D through boundary measurements, assuming that $\Lambda_D^{-1} + G : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ is injective. This injectivity condition can be interpreted in an intuitive way: If D is strongly scattering and characterized by the optical parameters κ and μ_a, Λ_D^{-1} maps the Dirichlet boundary value of the photon density on ∂D onto the Neumann boundary value of the photon density on ∂D. On the other hand, as indicated by (2.11), $-G$ does the same job if D is a transparent region. Consequently, $\Lambda_D^{-1} + G$ is injective if and only if one can determine whether D is transparent or diffusive by any single nontrivial measurement of the Dirichlet and Neumann boundary values of the photon density on ∂D.

Since $\Lambda_D^{-1} : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ is an isomorphism [10], $G : H^{1/2}(\partial D) \to L^2(\partial D)$ is bounded, and the imbedding $L^2(\partial D) \hookrightarrow H^{-1/2}(\partial D)$ is compact [10], $\Lambda_D^{-1} + G : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ is a Fredholm operator of index zero, and so its null space is in any case finite dimensional. Hence, the condition of Theorem 3.1 does not seem totally unrealistic, although there is in general no guarantee that $\Lambda_D^{-1} + G$ is injective: One can quite easily find physically reasonable parameters μ_a, κ, and $\tilde{\mu}_a$ for which the injectivity is lost. Luckily, the odds of meeting such parameters in practice seem quite low, but unfortunately the algorithmic implementation of Theorem 3.1 may run into trouble also if the optical properties inside Ω are just close to those causing noninjectivity. These statements are clarified by the following examples and the numerical studies of section 4.

Example 3.1. Let D be an isotropic unit disc, and assume that κ and μ_a are constant within D. Solving the problem (3.5) explicitly in polar coordinates $x = r(\cos(\theta), \sin(\theta))$ yields the spectral decomposition

$$\Lambda_D^{-1} : e^{ij\theta} \mapsto -\lambda_j e^{ij\theta}, \quad j \in \mathbb{Z},$$

where λ_j is given by

$$\lambda_j = \sqrt{\mu_a \kappa} \frac{I_{j-1}\left(\sqrt{\mu_a/\kappa}\right) + I_{j+1}\left(\sqrt{\mu_a/\kappa}\right)}{2I_j\left(\sqrt{\mu_a/\kappa}\right)} \tag{3.6}$$

and I_j is the modified Bessel function of the first kind. In this simple geometry the integral operator G obeys a similar representation. Indeed, by denoting $y = (\cos(\phi), \sin(\phi))$ in (2.7), one easily sees that

$$(\mathcal{G}\Phi)(\theta) = -\frac{1}{4\sqrt{2}} \int_{-\pi}^{\pi} (1 - \cos(\phi - \theta))^{1/2} e^{-\sqrt{2}\tilde{\mu}_a (1 - \cos(\phi - \theta))^{1/2}} \Phi(\phi) d\phi. \tag{3.7}$$

By substituting $e^{ij\phi}$ into (3.7), making a suitable change of variables, bearing (2.12) in mind, and using the even parity of the kernel in (3.7), it is straightforward to deduce that

$$G : e^{ij\theta} \mapsto \frac{2}{\pi}\frac{1+\eta_j}{1-\eta_j}e^{ij\theta}, \quad j \in \mathbb{Z},$$

where $\{\eta_j\}$ are the eigenvalues of $\mathcal{G}$:

$$\eta_j = -\frac{1}{4\sqrt{2}}\int_{-\pi}^{\pi}\cos(j\upsilon)(1-\cos(\upsilon))^{1/2}e^{-\sqrt{2}\tilde{\mu}_a(1-\cos(\upsilon))^{1/2}}d\upsilon. \tag{3.8}$$

These spectral decompositions indicate that the injectivity of $\Lambda_D^{-1}+G$ is lost if and only if the identity

$$\lambda_j = \frac{2}{\pi}\frac{1+\eta_j}{1-\eta_j} \tag{3.9}$$

holds for some $j \in \mathbb{Z}$.

Let us assume that $\tilde{\mu}_a > 0$ and $\kappa > 0$ are fixed and consider how $\mu_a > 0$ should be chosen in order to make $\Lambda_D^{-1}+G$ noninjective. It is easy to check that

$$\lim_{\mu_a\to 0}\lambda_j = |j|\kappa, \quad \lim_{\mu_a\to\infty}\lambda_j = \infty, \quad \text{and} \quad \frac{\partial}{\partial\mu_a}\lambda_j > 0 \quad \text{for all } j \in \mathbb{Z} \text{ and } \mu_a > 0,$$

where the equalities follow from the basic properties of the modified Bessel functions [1] and the inequality is a consequence of the monotonicity of $\langle\Lambda_D^{-1}g, g\rangle_{L^2(\partial D)}$ with respect to μ_a for any $g \in H^{1/2}(\partial D)$ (cf. [7]). Notice also that due to Lemma 2.1, $|\eta_j| < 1$ for all $j \in \mathbb{Z}$ and η_j converges to zero as $|j|$ goes to infinity. Hence, there exists a finite $J = J(\tilde{\mu}_a, \kappa) > 0$ such that for $0 \le j \le J$, equality (3.9) is satisfied by at most one value of μ_a, whereas no μ_a solves (3.9) for any $j > J$. To put it all together, in our simplified framework there are only finitely many values of μ_a that make $\Lambda_D^{-1}+G$ noninjective for fixed κ and $\tilde{\mu}_a$. □

Example 3.2. Let us continue working with the simple framework of Example 3.1 and assume that $\tilde{\mu}_a = \mu_a \ll \min\{1, \kappa\}$. Under such circumstances, the exponential term in (3.8) can be approximated fairly well by its linearization:

$$e^{-\sqrt{2}\mu_a(1-\cos(\upsilon))^{1/2}} \approx 1 - \sqrt{2}\mu_a(1-\cos(\upsilon))^{1/2}.$$

By substituting this into (3.8) and calculating the resulting integral with the help of Lemma 3.5 of [19], one sees that

$$\frac{2}{\pi}\frac{1+\eta_0}{1-\eta_0} \approx \frac{2\mu_a}{4-\pi\mu_a} \approx \frac{\mu_a}{2}.$$

On the other hand, since $\sqrt{\mu_a/\kappa}$ is small, the modified Bessel functions in (3.6) can be approximated by the constant and linear terms of their series representations [1], yielding

$$\lambda_0 \approx \sqrt{\mu_a\kappa}\frac{\sqrt{\mu_a/\kappa}}{2} = \frac{\mu_a}{2}.$$

In consequence, (3.9) is satisfied approximately for $j = 0$.

According to the above result, it is difficult to distinguish between a nonscattering disc and a homogeneous isotropic strongly scattering disc by a single rotationally symmetric measurement if both discs are characterized by the same negligible absorption. From the point of view of the factorization method, this observation is troublesome since the algorithmic implementation of Theorem 3.1 may fail if the transparent inclusion is as absorbing as its strongly scattering surroundings. This drawback is studied numerically in section 4. □

3.1. Proof of the characterization result. Theorem 3.1 follows with some work from the following result, which is a simplified version of Theorem 3.3 in [26].

THEOREM 3.2. *Let $X \subset U \subset X^*$ be a Gelfand triple with Hilbert space U and reflexive Banach space X such that the imbeddings are dense. Furthermore, let H be another Hilbert space, $T : X^* \to H$ linear, compact, and injective with dense range, $R : X \to X^*$ linear and self-adjoint, and*

$$A = TRT^*. \tag{3.10}$$

Assume that $R : X \to X^*$ is an isomorphism that can be written as $R = E + K$, where $K : X \to X^*$ is self-adjoint and compact and $E : X \to X^*$ is self-adjoint and coercive, i.e.,

$$\langle E\psi, \psi \rangle \geq C \|\psi\|_X^2 \quad \text{for all } \psi \in X.$$

Then, the ranges of $T : X^* \to H$ and $|A|^{1/2} = (A^2)^{1/4} : H \to H$ coincide.

In order to utilize Theorem 3.2, we need to show that $\Lambda - \Lambda_0$ obeys a factorization of the type (3.10). To this end, let us introduce a few auxiliary operators. We define the mapping $L : H^{-1/2}(\partial D) \to H^{1/2}(\partial\Omega)$ through

$$L : \phi \mapsto v|_{\partial\Omega}, \tag{3.11}$$

where $v \in H^1(\Omega \setminus \overline{D})$ is the weak solution of

$$\begin{aligned} \nabla \cdot \kappa \nabla v - \mu_a v &= 0 && \text{in } \Omega \setminus \overline{D}, \\ \nu \cdot \kappa \nabla v &= 0 && \text{on } \partial\Omega, \\ \nu \cdot \kappa \nabla v &= \phi && \text{on } \partial D. \end{aligned}$$

In what follows, L plays the role of T in Theorem 3.2. The adjoint $L^* : H^{-1/2}(\partial\Omega) \to H^{1/2}(\partial D)$ is defined by (cf. [7])

$$L^* : \phi' \mapsto v'|_{\partial D}, \tag{3.12}$$

where $v' \in H^1(\Omega \setminus \overline{D})$ is the weak solution of

$$\begin{aligned} \nabla \cdot \kappa \nabla v' - \mu_a v' &= 0 && \text{in } \Omega \setminus \overline{D}, \\ \nu \cdot \kappa \nabla v' &= \phi' && \text{on } \partial\Omega, \\ \nu \cdot \kappa \nabla v' &= 0 && \text{on } \partial D. \end{aligned}$$

Next, we will introduce two mappings that constitute the intermediate operator R of Theorem 3.2. For $\psi \in H^{1/2}(\partial D)$, let $w_0 \in H^1(\Omega \setminus \partial D)$ be the weak solution of the transmission problem

$$\begin{aligned} \nabla \cdot \kappa \nabla w_0 - \mu_a w_0 &= 0 && \text{in } \Omega \setminus \partial D, \\ \nu \cdot \kappa \nabla w_0 &= 0 && \text{on } \partial\Omega, \\ w_0^+ - w_0^- &= \psi && \text{on } \partial D, \\ \nu \cdot \kappa \nabla w_0^+ - \nu \cdot \kappa \nabla w_0^- &= 0 && \text{on } \partial D. \end{aligned} \tag{3.13}$$

Here and in what follows, we denote by $w_0^{\pm}$ the trace from the exterior and from the interior of $\Omega \setminus \overline{D}$, respectively; the superscripts will often be left out if the direction of approach is clear from the context or if it does not affect the value of the trace. Moreover, assume that $w \in H^1(\Omega \setminus \overline{D})$ is the weak solution of

$$\begin{aligned} \nabla \cdot \kappa \nabla w - \mu_a w &= 0 \qquad \text{in } \Omega \setminus \overline{D}, \\ \nu \cdot \kappa \nabla w &= 0 \qquad \text{on } \partial\Omega, \\ G(w+\psi) + \nu \cdot \kappa \nabla w &= 0 \qquad \text{on } \partial D. \end{aligned} \tag{3.14}$$

We define the operators $F_0, F : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ through

$$F_0 : \psi \mapsto \nu \cdot \kappa \nabla w_0|_{\partial D}, \quad F : \psi \mapsto \nu \cdot \kappa \nabla w|_{\partial D}. \tag{3.15}$$

The following lemma lists some essential properties of the above introduced operators.

LEMMA 3.3. *The operators $L : H^{-1/2}(\partial D) \to H^{1/2}(\partial\Omega)$, $L^* : H^{-1/2}(\partial\Omega) \to H^{1/2}(\partial D)$, and $F_0, F : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ defined by* (3.11), (3.12), *and* (3.15), *respectively, are linear and bounded. Furthermore, L is injective and compact, and its range is dense in $H^{1/2}(\partial\Omega)$; $-F_0$ is a coercive and self-adjoint isomorphism, and F is compact and self-adjoint.*

Proof. The fact that $L : H^{-1/2}(\partial D) \to H^{1/2}(\partial\Omega)$ and $L^* : H^{-1/2}(\partial\Omega) \to H^{1/2}(\partial D)$ are linear and bounded follows from the standard theory of elliptic partial differential equations [10]. Furthermore, both L and L^* are injective and compact due to the unique continuation principle and the regularity theory for elliptic partial differential equations (cf. [26]). In particular, $\overline{\mathcal{R}(L)} = \mathcal{N}(L^*)^{\perp} = H^{1/2}(\partial\Omega)$; i.e., the range of L is dense in $H^{1/2}(\partial D)$.

The unique solvability of (3.13) in $H^1(\Omega \setminus \partial D)$ and the solution's continuous dependence on the data $\psi \in H^{1/2}(\partial D)$ follow, for example, from the material in [27] (see also [26]). Since $\nabla \cdot \kappa \nabla w_0 = \mu_a w_0 \in L^2(\Omega \setminus \partial D)$, $F_0 : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ is bounded according to a slight modification of [10, Lemma 1, p. 381]. Furthermore, the self-adjointness of F_0 follows by modifying the proof of Lemma 3.3 of [7] in an obvious way. In addition, $-F_0$ can be shown to be coercive by estimating as follows:

$$\begin{aligned} \langle -F_0\psi, \psi \rangle_{L^2(\partial D)} &= \langle \nu \cdot \kappa \nabla w_0, w_0^- \rangle_{L^2(\partial D)} - \langle \nu \cdot \kappa \nabla w_0, w_0^+ \rangle_{L^2(\partial D)} \\ &= \int_{\Omega \setminus \overline{D}} \left(\kappa \nabla w_0 \cdot \nabla \overline{w}_0 + \mu_a |w_0|^2 \right) dx \\ &\quad + \int_D \left(\kappa \nabla w_0 \cdot \nabla \overline{w}_0 + \mu_a |w_0|^2 \right) dx \\ &\geq C \left\{ \|w_0\|^2_{H^1(\Omega \setminus \overline{D})} + \|w_0\|^2_{H^1(D)} \right\} \\ &\geq C \left\{ \left\|w_0^-\right\|^2_{H^{1/2}(\partial D)} + \left\|w_0^+\right\|^2_{H^{1/2}(\partial D)} \right\} \\ &\geq C \left\|\psi\right\|^2_{H^{1/2}(\partial D)}, \end{aligned}$$

where we used Green's formula [10], the symmetry of $\kappa : \Omega \to \mathbb{R}^{n \times n}$, (3.1), and the trace theorem. Since $-F_0 : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ is coercive and self-adjoint, it is injective and its range is closed and dense. In other words, F_0 is an isomorphism.

To finish the proof, let us consider F. The variational formulation of (3.14) is to find $w \in H^1(\Omega \setminus \overline{D})$ such that

$$\int_{\Omega \setminus \overline{D}} \left(\kappa \nabla w \cdot \nabla \overline{v} + \mu_a w \overline{v} \right) dx + \int_{\partial D} G w \overline{v} dS = -\langle G\psi, v \rangle_{L^2(\partial D)} \tag{3.16}$$

for all $v \in H^1(\Omega \setminus \overline{D})$. Because of (3.1), Lemma 2.1, the trace theorem, and the nonnegativeness of $G : L^2(\partial D) \to L^2(\partial D)$ indicated by Theorem 3.12 of [19], the left-hand side of (3.16) defines a coercive and bounded sesquilinear form from $H^1(\Omega \setminus \overline{D}) \times H^1(\Omega \setminus \overline{D})$ to $\mathbb{C}$, and the right-hand side induces a bounded antilinear functional from $H^1(\Omega \setminus \overline{D})$ to $\mathbb{C}$. In consequence, due to the Lax–Milgram lemma, (3.16) has a unique solution $w \in H^1(\Omega \setminus \overline{D})$ which depends continuously on the data:

$$\|w\|_{H^1(\Omega\setminus\overline{D})} \leq C \|G\psi\|_{H^{-1/2}(\partial D)} \leq C \|\psi\|_{L^2(\partial D)}.$$

In particular, by the trace theorem and Lemma 2.1,

$$\|F\psi\|_{L^2(\partial D)} = \|G(w+\psi)\|_{L^2(\partial D)} \leq C \left\{ \|w\|_{L^2(\partial D)} + \|\psi\|_{L^2(\partial D)} \right\} \leq C \|\psi\|_{L^2(\partial D)};$$

i.e., F is bounded from $L^2(\partial D)$ to itself. Hence, due to the compactness of the imbeddings $H^{1/2}(\partial D) \hookrightarrow L^2(\partial D)$ and $L^2(\partial D) \hookrightarrow H^{-1/2}(\partial D)$ [10], $F : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ is bounded and compact.

In order to prove that F is self-adjoint, let $w_1, w_2 \in H^1(\Omega \setminus \overline{D})$ be the solutions of (3.14) corresponding to the inputs $\psi_1, \psi_2 \in H^{1/2}(\partial D)$, respectively. By using the self-adjointness of $G : L^2(\partial D) \to L^2(\partial D)$ and the inner boundary condition of (3.14), it follows that

$$\begin{aligned} \langle F\psi_1, \psi_2 \rangle_{L^2(\partial D)} &= -\int_{\partial D} G(\psi_1 + w_1)\overline{\psi}_2 dS \\ &= \int_{\partial D} \psi_1(\nu \cdot \kappa \nabla \overline{w}_2 + \overline{Gw_2}) dS - \int_{\partial D} w_1 \overline{G\psi}_2 dS \\ &= \overline{\langle F\psi_2, \psi_1 \rangle}_{L^2(\partial D)} + \int_{\partial D} (G\psi_1 \overline{w}_2 - w_1 \overline{G\psi}_2) dS. \end{aligned} \tag{3.17}$$

Thus, by showing that the latter term on the last line of (3.17) vanishes, the proof is complete:

$$\begin{aligned} \int_{\partial D} (G\psi_1 \overline{w}_2 - w_1 \overline{G\psi}_2) dS &= \int_{\partial D} (\nu \cdot \kappa \nabla \overline{w}_2 w_1 - \nu \cdot \kappa \nabla w_1 \overline{w}_2) dS \\ &\quad + \int_{\partial D} (w_1 \overline{Gw}_2 - G w_1 \overline{w}_2) dS \\ &= \int_{\partial\Omega} (\nu \cdot \kappa \nabla w_1 \overline{w}_2 - \nu \cdot \kappa \nabla \overline{w}_2 w_1) dS = 0, \end{aligned}$$

where we used the boundary conditions of (3.14), the self-adjointness of G, the symmetry of κ, and Green's formula. ∎

Next, we will provide the needed factorization.

LEMMA 3.4. *The difference of the operators $\Lambda, \Lambda_0 : H^{-1/2}(\partial\Omega) \to H^{1/2}(\partial\Omega)$ can be factorized as $\Lambda - \Lambda_0 = L(F - F_0)L^*$, where $L : H^{-1/2}(\partial D) \to H^{1/2}(\partial\Omega)$, $L^* : H^{-1/2}(\partial\Omega) \to H^{1/2}(\partial D)$, and $F_0, F : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ are defined by* (3.11), (3.12), *and* (3.15), *respectively.*

Proof. Let $\varphi \in H^1(\Omega\setminus\overline{D})$ be the solution of (3.3) corresponding to $f \in H^{-1/2}(\partial\Omega)$, and let $\varphi_0 \in H^1(\Omega)$ be the solution of

$$\nabla \cdot \kappa \nabla \varphi_0 - \mu_a \varphi_0 = 0 \quad \text{in } \Omega, \qquad \nu \cdot \kappa \nabla \varphi_0 = f \quad \text{on } \partial\Omega. \tag{3.18}$$

Clearly,

$$L\left(\nu \cdot \kappa \nabla(\varphi - \varphi_0)|_{\partial D}\right) = (\varphi - \varphi_0)|_{\partial\Omega} = (\Lambda - \Lambda_0) f.$$

By defining the operators

$$\begin{aligned} B: &\quad f \mapsto \nu \cdot \kappa \nabla \varphi|_{\partial D}, \quad H^{-1/2}(\partial\Omega) \to H^{-1/2}(\partial D), \\ B_0: &\quad f \mapsto \nu \cdot \kappa \nabla \varphi_0|_{\partial D}, \quad H^{-1/2}(\partial\Omega) \to H^{-1/2}(\partial D), \end{aligned}$$

we have thus far obtained the factorization

$$\Lambda - \Lambda_0 = L(B - B_0). \tag{3.19}$$

Notice that B is bounded (from $H^{-1/2}(\partial\Omega)$ to $L^2(\partial D)$) due to the trace theorem and Lemma 2.1, and the boundedness of B_0 follows from the fact that $\nabla \cdot \kappa \nabla \varphi_0 = \mu_a \varphi_0 \in L^2(\Omega)$ [10].

By following the line of reasoning used in the proof of Lemma 3.2 in [7], one easily sees that the adjoint operator of $B_0 : H^{-1/2}(\partial\Omega) \to H^{-1/2}(\partial D)$ is defined through

$$B_0^* : \psi \mapsto w_0|_{\partial\Omega}, \quad H^{1/2}(\partial D) \to H^{1/2}(\partial\Omega),$$

where $w_0 \in H^1(\Omega \setminus \partial D)$ is the solution of (3.13). Similarly, the adjoint of $B : H^{-1/2}(\partial\Omega) \to H^{-1/2}(\partial D)$ can be defined through

$$B^* : \psi \mapsto w|_{\partial\Omega}, \quad H^{1/2}(\partial D) \to H^{1/2}(\partial\Omega),$$

where $w \in H^1(\Omega \setminus \overline{D})$ is the solution of (3.14). Indeed, by using the self-adjointness of $G : L^2(\partial D) \to L^2(\partial D)$, Green's formula, and the boundary conditions of (3.3) and (3.14), we obtain that

$$\begin{aligned} \langle Bf, \psi \rangle_{L^2(\partial D)} &= -\int_{\partial D} G\varphi\overline{\psi} dS \\ &= \int_{\partial D} \varphi(\overline{Gw} + \nu \cdot \kappa \nabla \overline{w}) dS \\ &= \int_{\partial D} \nu \cdot \kappa \nabla \overline{w} \varphi dS - \int_{\partial D} \nu \cdot \kappa \nabla \varphi \overline{w} dS \\ &= \int_{\partial\Omega} \nu \cdot \kappa \nabla \varphi \overline{w} dS - \int_{\partial\Omega} \nu \cdot \kappa \nabla \overline{w} \varphi dS \\ &= \langle f, w \rangle_{L^2(\partial\Omega)}. \end{aligned}$$

Bearing in mind how $F, F_0 : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ and $L : H^{-1/2}(\partial D) \to H^{1/2}(\partial\Omega)$ were defined, it is straightforward to deduce that

$$L(F - F_0) = B^* - B_0^*. \tag{3.20}$$

By taking the adjoint of (3.20), plugging it into (3.19), and using the self-adjointness of F and F_0, we finally obtain that

$$\Lambda - \Lambda_0 = L(F - F_0)L^*,$$

which completes the proof. □

Now the proof of Theorem 3.1 follows by combining Theorem 3.2 with Lemma 3.4 and showing that $F - F_0 : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ is an isomorphism if $\Lambda_D^{-1} + G : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ is injective.

Proof of Theorem 3.1. Inspired by Theorem 3.2, let us make the choices $X = H^{1/2}(\partial D)$, $U = L^2(\partial D)$, $H = L^2(\partial\Omega)$, $R = F - F_0$, $A = (\Lambda - \Lambda_0)|_{L^2(\partial\Omega)} : L^2(\partial\Omega) \to$

$L^2(\partial\Omega)$, and $T = L$ with $\mathcal{R}(L)$ interpreted as a subspace $L^2(\partial\Omega)$. Clearly, $X \subset U \subset X^*$ is now a Gelfand triple, and T is compact and injective with dense range in H because of Lemma 3.3 and since the imbedding $H^{1/2}(\partial\Omega) \hookrightarrow L^2(\partial\Omega)$ is compact and dense [10]. Furthermore, by making the choices $E = -F_0$ and $K = F$, $R : X \to X^*$ is given as a sum of two self-adjoint operators, one of which is coercive and the other one compact. Hence, to be able to use Theorem 3.2, the only thing we still need to prove is that $F - F_0 : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ is an isomorphism.

Since $F - F_0 : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ is a sum of a compact and an isomorphic operator, it is a Fredholm operator of index zero, and so it is bijective if and only if it is injective. Assume that $\psi \in H^{1/2}(\partial D)$ belongs to the null space of $F - F_0$, meaning that the solutions of (3.14) and (3.13), namely $w \in H^1(\Omega \setminus \overline{D})$ and $w_0 \in H^1(\Omega \setminus \partial D)$, satisfy the equality

$$\nu \cdot \kappa \nabla w = \nu \cdot \kappa \nabla w_0 \quad \text{on } \partial D. \tag{3.21}$$

In consequence, w and $w_0|_{\Omega\setminus\overline{D}}$ satisfy the diffusion equation with the same Neumann boundary values in $\Omega \setminus \overline{D}$. Due to (3.1), such a Neumann boundary value problem is uniquely solvable in $H^1(\Omega\setminus\overline{D})$ [10], and so $w_0|_{\Omega\setminus\overline{D}} = w$. In particular, $w_0^-|_{\partial D} = w|_{\partial D}$.

Clearly, $w_0|_D \in H^1(D)$ satisfies (3.5) with $g = w_0^+|_{\partial D} = w|_{\partial D} + \psi$, from which it follows that

$$\Lambda_D^{-1}(w|_{\partial D} + \psi) = \nu \cdot \kappa \nabla w_0.$$

By using (3.21) and the inner boundary condition of (3.14), we deduce that

$$\left(\Lambda_D^{-1} + G\right)(w|_{\partial D} + \psi) = 0,$$

i.e., $w|_{\partial D} + \psi = 0$ since $\Lambda_D^{-1} + G$ is injective by assumption. In consequence, (3.14) transforms into a homogeneous Neumann problem, and so $w_0|_{\Omega\setminus\overline{D}} = w = 0$. In particular, $\nu \cdot \kappa \nabla w_0|_{\partial D} = 0$, or in other words, $F_0\psi = 0$. Since $F_0 : H^{1/2}(\partial D) \to H^{-1/2}(\partial D)$ is an isomorphism, we deduce that $\psi = 0$. Thus, $\mathcal{N}(F - F_0) = \{0\}$ and $F - F_0$ is a linear isomorphism.

Now Theorem 3.2 tells us that

$$\mathcal{R}\left(\left|(\Lambda - \Lambda_0)\,|_{L^2(\partial\Omega)}\right|^{1/2}\right) = \mathcal{R}(L).$$

Bearing this equality in mind, the claim finally follows by using the same line of reasoning as in the proof of Lemma 3.5 in [7]. □

4. Numerical examples. In this section, the performance of the introduced method is evaluated by simulated test cases. In subsection 4.1, we discuss the algorithmic implementation of Theorem 3.1. Subsection 4.2 considers briefly the computational methods for finding forward solutions to the radiosity-diffusion model. The reconstruction results are presented in subsection 4.3.

4.1. Algorithmic implementation. Assume that $f_y^\alpha \in L^2(\partial\Omega)$ is the unique minimizer of the Tikhonov functional

$$\left\| |\Lambda - \Lambda_0|^{1/2} f - h_y \right\|_{L^2(\partial\Omega)}^2 + \alpha \|f\|_{L^2(\partial\Omega)}^2, \quad f \in L^2(\partial\Omega), \tag{4.1}$$

where $\alpha > 0$ is a regularization parameter and $h_y \in C^\infty(\overline{\Omega} \setminus \{y\})$ is the singular solution of (3.4). Let us examine how f_y^α behaves as α goes to zero. It is well known

that f_y^α converges to the minimum norm solution of the equation

$$|\Lambda - \Lambda_0|^{1/2} f = h_y|_{\partial\Omega}$$

if $h_y|_{\partial\Omega}$ belongs to the range of $|\Lambda - \Lambda_0|^{1/2} : L^2(\partial\Omega) \to L^2(\partial\Omega)$ [24]. On the other hand, if $h_y|_{\partial\Omega} \notin \mathcal{R}(|\Lambda - \Lambda_0|^{1/2})$ and the injectivity condition of Theorem 3.1 is satisfied, in which case $\mathcal{R}(|\Lambda - \Lambda_0|^{1/2})$ is dense in $L^2(\partial\Omega)$, the $L^2(\partial\Omega)$-norm of f_y^α goes to infinity as α goes to zero. In consequence, provided that the assumptions of Theorem 3.1 are valid, y belongs to D if and only if

$$\limsup_{\alpha\to 0^+} \left\| f_y^\alpha \right\|_{L^2(\partial\Omega)} < \infty.$$

In practical computations we choose an $L^2(\partial\Omega)$-orthonormal set of input patterns $\{f_l\}_{l=-m}^m$, denote the orthogonal projector onto the span of $\{f_l\}$ by P, and replace (4.1) by

$$\left\| |P(\Lambda - \Lambda_0)P|^{1/2} f - \frac{Ph_y}{\|Ph_y\| L^2(\partial\Omega)} \right\| L^2(\partial\Omega)^2 + \alpha_\delta(y) \|f\|^2_{L^2(\partial\Omega)} . \tag{4.2}$$

Furthermore, instead of examining how the norm of the minimizer for (4.2) behaves as y moves around in Ω, we take a slightly different approach that gives better contrast. We choose $\alpha_\delta : \Omega \to \mathbb{R}_+$, $0 < \delta < 1$, in such a way that the minimizer $f_y^\delta \in L^2(\partial\Omega)$ of the Tikhonov functional (4.2) satisfies the discrepancy condition

$$\left\| |P(\Lambda - \Lambda_0)P|^{1/2} f_y^\delta - \frac{Ph_y}{\|Ph_y\| L^2(\partial\Omega)} \right\|_{L^2(\partial\Omega)} = \delta \tag{4.3}$$

for every $y \in \Omega$. If this condition cannot be met for some $y \in \Omega$, we set $\alpha_\delta(y) = 0$; not taking into account possible numerical restrictions, such a situation can occur only if the injectivity condition of Theorem 3.1 does not hold true.

Let us explain why the graph of $\alpha_\delta : \Omega \to \mathbb{R}_+$ contains information on the location of D, provided that the injectivity condition of Theorem 3.1 is satisfied. The above considerations suggest that if (4.3) is satisfied for a small $0 < \delta < 1$ and $y \in \Omega \setminus D$, the corresponding minimizer $f_y^\delta \in L^2(\partial\Omega)$ probably has a large norm. Since the norm of the minimizer of (4.2) is monotonically decreasing with respect to the regularization parameter [24], $\alpha_\delta(y)$ is likely to be small if $y \in \Omega \setminus D$. Conversely, if $y \in D$, the minimizer of (4.2) satisfying (4.3) is probably smallish in norm, and so $\alpha_\delta(y)$ is presumably large compared to the case $y \in \Omega \setminus D$. In consequence, the graph of $\alpha_\delta : \Omega \to \mathbb{R}_+$ should be flat apart from an elevation over the inclusion D.

In the practical algorithm, we have chosen to work with $Ph_y / \|Ph_y\|_{L^2(\partial\Omega)}$ instead of Ph_y because one is ultimately interested in the shape of the singular solution on $\partial\Omega$, not in its magnitude; it is trivial to check that Theorem 3.1 remains valid if $h_y|_{\partial\Omega}$ is replaced by $h_y|_{\partial\Omega} / \|h_y\|_{L^2(\partial\Omega)}$.

4.2. Numerical implementation. To be able to test the algorithm introduced in subsection 4.1 numerically, one needs to simulate $\{\Lambda f_l\}$ and $\{\Lambda_0 f_l\}$, compute $P(h_y|_{\partial\Omega})$ for numerous $y \in \Omega$, and introduce a procedure aiming at satisfying (4.3). In what follows, we will discuss each of these steps briefly.

4.2.1. Simulation of the measurement data. The numerical solution of (3.3) is based on the finite element method with piecewise linear basis functions. The

domain $\Omega \setminus \overline{D}$ is divided into T triangles joined at N vertex nodes, and the solution is sought in the form

$$\varphi^N = \sum_{k=1}^{N} a_k \phi_k,$$

where $a_k \in \mathbb{C}$ and the basis function ϕ_k is continuous, piecewise linear, smooth in each triangle, has value one at the kth node, and vanishes at all the other nodes. The coefficient vector $a = (a_1, \ldots, a_N)^T$ is solved from the matrix equation

$$(K + M + A)a = F,$$

which is obtained from the variational formulation of problem (3.3) in standard finite element style (cf. [19]):

$$\begin{aligned} K_{jk} &= \int_{\Omega\setminus\overline{D}} \kappa \nabla\phi_k \cdot \nabla\phi_j dx, \\ M_{jk} &= \int_{\Omega\setminus\overline{D}} \mu_a \phi_k \phi_j dx, \\ A_{jk} &= \int_{\partial D} G\phi_k \phi_j dS, \\ F_j &= \int_{\partial\Omega} f\phi_j dS. \end{aligned} \tag{4.4}$$

These matrix elements are approximated by integrals over the finite element triangulation; an approximative way to compute $\mathcal{G}\phi_k$, needed in the evaluation of (4.4), is given in [4]. After $a \in \mathbb{C}^N$ corresponding to the input f_l has been solved, Λf_l is approximated by the Dirichlet boundary value of φ^N on $\partial\Omega$.

The elliptic boundary value problem corresponding to Λ_0, i.e., (3.18), could be solved in a similar manner. However, in the numerical studies presented in subsection 4.3, Ω is the unit disc, μ_a and κ are constant and scalar in Ω, and $f_l = e^{il\theta}$ with θ being the polar angle. This simplifies the situation considerably as Λ_0 obeys the spectral representation (see Example 3.1)

$$\Lambda_0 : e^{il\theta} \mapsto \frac{2}{\sqrt{\mu_a \kappa}} \frac{I_l\left(\sqrt{\mu_a/\kappa}\right)}{I_{l-1}\left(\sqrt{\mu_a/\kappa}\right) + I_{l+1}\left(\sqrt{\mu_a/\kappa}\right)} e^{il\theta}, \quad l = \mathbb{Z}, \tag{4.5}$$

where I_l is the modified Bessel function of the first kind. In what follows, $\{\Lambda_0 f_l\}$ are computed using (4.5).

4.2.2. Computation of the singular photon density. Let us assume that μ_a and κ are scalar and constant in Ω. In this case, the solution of (3.4) can be computed with the help of the modified Bessel function of the second kind because

$$\tilde{h}_y = -\frac{1}{2\pi\kappa} K_0\left(\sqrt{\frac{\mu_a}{\kappa}}|x-y|\right)$$

satisfies the first equation of (3.4) [1]. Hence, the solution of (3.4) can be given as

$$h_y = \tilde{h}_y - v_y,$$

where v_y is the solution of the boundary value problem

$$\nabla \cdot \kappa \nabla v - \mu_a v = 0 \quad \text{in } \Omega, \qquad \nu \cdot \kappa \nabla v = \nu \cdot \kappa \nabla \tilde{h}_y \quad \text{on } \partial\Omega. \tag{4.6}$$

In consequence, for each probe location $y \in \Omega$ we must, in general, solve one elliptic Neumann boundary value problem using, for example, the finite element method.

The numerical studies of subsection 4.3 are conducted in the unit disc with the Fourier input patterns $\{e^{il\theta}\}$. Since only the projection $P(h_y|_{\partial\Omega})$ is explicitly needed in (4.2) and (4.3), in this simple framework the algorithm of subsection 4.1 can be implemented without having to solve (4.6). Indeed, by using the spectral decomposition (4.5), it is easy to see that

$$P\left(h_y|_{\partial\Omega}\right) = P\left(\tilde{h}_y|_{\partial\Omega}\right) - P\Lambda_0\left(\nu \cdot \kappa \nabla \tilde{h}_y|_{\partial\Omega}\right) = P\left(\tilde{h}_y|_{\partial\Omega}\right) - \Lambda_0 P\left(\nu \cdot \kappa \nabla \tilde{h}_y|_{\partial\Omega}\right),$$

which can be computed efficiently with the help of (4.5).

4.2.3. Computation of the indicator function. Let us consider finding the regularization parameter $\alpha_\delta(y)$ satisfying (4.3) for a fixed $y \in \Omega$. It is well known that the minimizer of the functional (4.2), with $\alpha_\delta(y)$ replaced by a generic regularization parameter $\alpha > 0$, is given by [24]

$$f_y^\alpha = \frac{1}{\|Ph_y\|_{L^2(\partial\Omega)}} \left\{|P(\Lambda - \Lambda_0)P| + \alpha I\right\}^{-1} |P(\Lambda - \Lambda_0)P|^{1/2} Ph_y,$$

where we have used the self-adjointness of Λ, Λ_0, and P. Furthermore, the derivative of the discrepancy function

$$e(\alpha) = \left\| |P(\Lambda - \Lambda_0)P|^{1/2} f_y^\alpha - \frac{Ph_y}{\|Ph_y\|\, L^2(\partial\Omega)} \right\|^2_{L^2(\partial\Omega)}$$

can be written as [24]

$$e'(\alpha) = 2\alpha \left\langle f_y^\alpha, \left\{|P(\Lambda - \Lambda_0)P| + \alpha I\right\}^{-1} f_y^\alpha \right\rangle_{L^2(\partial\Omega)}.$$

By using these formulae, $\alpha_\delta(y)$ can be computed efficiently by Newton's method; if the discrepancy condition (4.3) cannot be achieved within the working precision, $\alpha_\delta(y)$ is set to zero. Notice that the number of orthogonal input patterns used is typically quite low—in our studies, nine—and so computing the inverse operators needed above is relatively cheap.

4.3. Results. In this subsection, we evaluate the performance of the proposed method with four test cases. In the first test case, we try to locate a nonconvex transparent inclusion inside an object that is far less scattering and absorbing than the tissues encountered in practical applications. The idea is to set up an upper bound on the performance of the algorithm: When performing measurements with an object characterized by smaller absorption and scattering coefficients, there are more photons that travel through the inner parts of the object without being absorbed before arriving at the object boundary. As a consequence, the measurements contain more information on the inner parts of the object and the inverse problem becomes less ill-posed. The second test case considers a circular nonscattering inhomogeneity and optical parameter values that are chosen so that the injectivity condition of

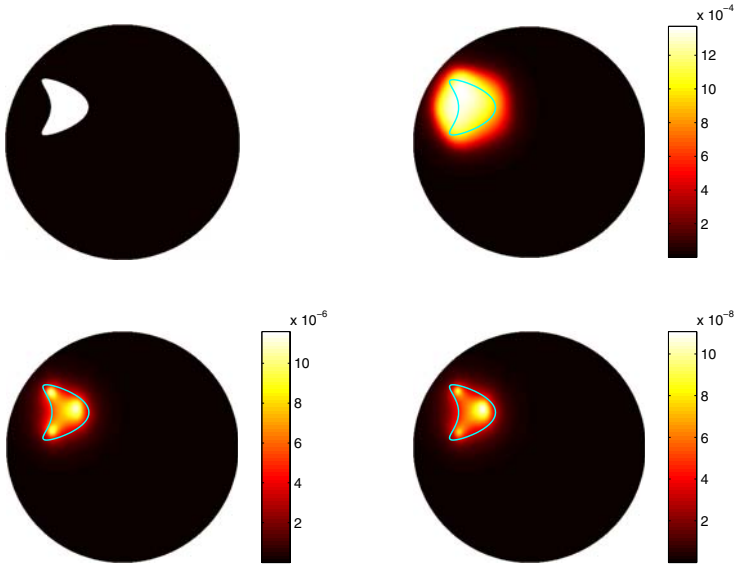


FIG. 4.1. *The first test case with the optical parameters* $\kappa = 0.5$, $\mu_a = 0.05$, *and* $\tilde{\mu}_a = 0.005$. *The graph of the indicator function* α_δ *is compared with the original kite-shaped transparent cavity for different* $\delta > 0$. *Top left: The original cavity. Top right:* $\delta = 10^{-3}$. *Bottom left:* $\delta = 10^{-5}$. *Bottom right:* $\delta = 10^{-7}$.

Theorem 3.1 is not satisfied. The third and fourth cases work with realistic optical parameters and noisy data. In the third test, the object is contaminated by a nonconvex transparent inclusion, and in the fourth one we have two circular inclusions.

The computations were conducted in two dimensions, and the object of interest Ω was an isotropic unit disc with constant background diffusion and absorption coefficients. In all studies, we worked with full aperture data and used nine L^2-orthogonal input patterns $\{e^{il\theta}\}_{l=-4}^{l=4}$, where θ is the polar angle. By following the line of reasoning introduced in [15] for electrical impedance tomography, the algorithm could also be implemented with limited aperture data, but the resulting reconstructions would be worse than those presented in this work. The reason behind the low number of inputs is twofold: First, the higher spatial frequencies penetrate the object so poorly that their outputs contain more numerical noise than information on the location of the transparent region. Second, because of the limitations of the real-life measurement setting [16], one cannot assume the object to be exposed on arbitrarily high spatial frequencies in a controlled manner. The outputs $\{(\Lambda e^{il\cdot})(\theta)\}$ and $\{(\Lambda_0 e^{il\cdot})(\theta)\}$ were simulated by the finite element method, with approximately five hundred thousand nodal points, and by formula (4.5), respectively. In the computations, we used two types of inclusion shapes:

$$D_1 = \{x \in \Omega \mid |x - x_0| < R\}, \quad x_0 \in \Omega,\ |x_0| + R < 1,$$

and D_2 is the celebrated nonconvex kite-shaped inclusion; i.e., D_2 is the interior of the region bounded by the curve

$$z(t) = x_0 + R(\cos t + 0.65 \cos 2t - 0.65, 1.5 \sin t), \quad t \in (-\pi, \pi],$$

where x_0 and R are such that Ω contains the curve.

The results for the first test case are illustrated in Figures 4.1 and 4.2. The original transparent inclusion, namely D_2 with $x_0 = (-0.45, 0.3)$ and $R = 0.16$, is shown in the

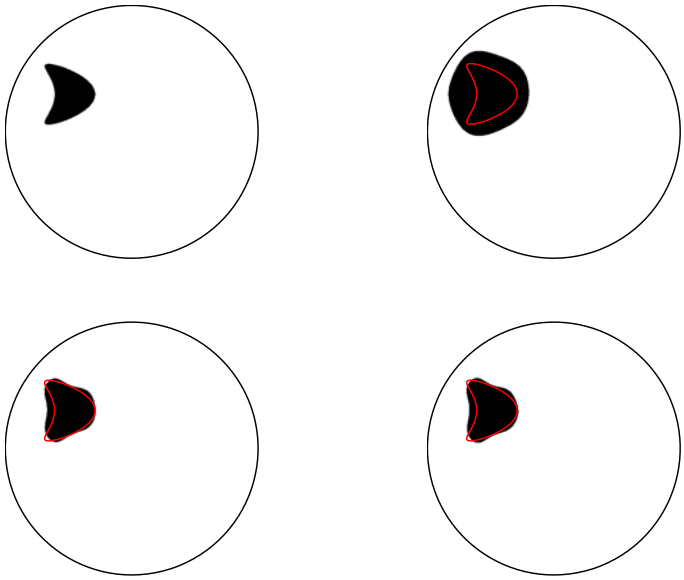


FIG. 4.2. *The first test case. The inclusion supports obtained by approximating the indicator functions in the sense of least squares are compared with the original kite-shaped transparent cavity. Top left: The original inclusion. Top right:* $\delta = 10^{-3}$. *Bottom left:* $\delta = 10^{-5}$. *Bottom right:* $\delta = 10^{-7}$.

top left image of Figure 4.1. The a priori known background optical parameters were $\kappa = 0.5$ and $\mu_a = 0.05$, and the absorption in the nonscattering region was $\tilde{\mu}_a = 0.005$; these parameter values correspond to absorption and scattering that are less than a tenth of the values that would be used if one modeled the neonatal head [2]. The other three images of Figure 4.1 show the graphs of the indicator function $\alpha_\delta : \Omega \to \mathbb{R}$ on a rectangular grid with three different discrepancy parameter values: $\delta = 10^{-3}$, 10^{-5}, and 10^{-7}. The reconstructed inclusion supports presented in Figure 4.2 were obtained by approximating the indicator functions in the sense of least squares by piecewise constant functions that take at most two distinct values. As Figures 4.1 and 4.2 demonstrate, the algorithm finds the approximate location of the inclusion with all three discrepancy parameters, whereas only the smallest two parameter values result in reconstructions that represent some characteristics of the inclusion shape. Although the reconstructed inclusion supports illustrated on the bottom row of Figure 4.2 are slightly concave on the side that corresponds to the nonconvex face of the original kite-shaped cavity, one cannot claim that the convexity properties of the inclusion are reconstructed correctly since there is also some concavity on the two originally convex sides. In fact, by testing the method with transparent cavities that are more concave than the one used here, it can be demonstrated that the proposed method is not accurate enough to capture nonconvexity reliably.

Figure 4.3 shows the findings of the second test case, where the transparent inclusion was D_1 with $x_0 = (-0.4, -0.5)$ and $R = 0.2$. In this test, we used three different sets of optical parameters: $\kappa = 0.5$, $\mu_a = 0.05$, and $\tilde{\mu}_a = 0.005$; $\kappa = 0.5$, $\mu_a = 0.05$, and $\tilde{\mu}_a = 0.05$; $\kappa = 0.254$, $\mu_a = 0.05$, and $\tilde{\mu}_a = 0.005$. By using the same line of reasoning as in Examples 3.1 and 3.2, it is easy to see that the latter two parameter sets make $\Lambda_{D_1}^{-1} + G$ nearly noninjective. To be more precise, with the second set of parameters

$$(\Lambda_{D_1}^{-1} + G)(e^{ij\theta}) \approx 0$$

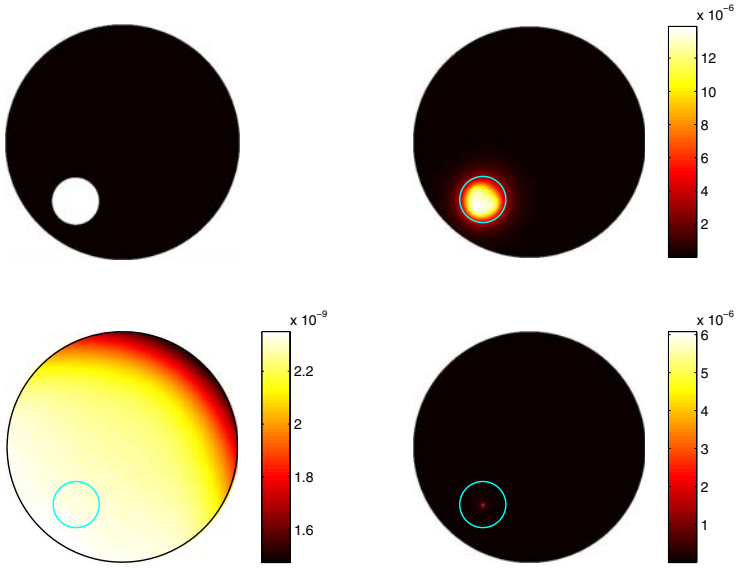


FIG. 4.3. *The second test case. The graph of the indicator function α_δ, with $\delta = 10^{-5}$, is compared with the original circular inclusion. Three different sets of optical parameters were used. Top left: The original inclusion. Top right: $\kappa = 0.5$, $\mu_a = 0.05$, and $\tilde{\mu}_a = 0.005$. Bottom left: $\kappa = 0.5$, $\mu_a = 0.05$, and $\tilde{\mu}_a = 0.05$. Bottom right: $\kappa = 0.254$, $\mu_a = 0.05$, and $\tilde{\mu}_a = 0.005$.*

for $j = 0$, and with the third set the same equation holds for $j = \pm 1$. Here, θ denotes the polar angle with respect to the center of the inclusion. The graphs of the indicator functions corresponding to the three parameter sets and $\delta = 10^{-5}$ are shown in Figure 4.3. The reconstruction corresponding to the first set of parameters is good, but the reconstructions with the latter two parameter sets have their shortcomings as expected: The indicator function corresponding to the second set of parameters contains almost no information on the location of the inclusion, whereas the indicator function corresponding to the last set of parameters finds the location correctly but contains little information on the size of the inhomogeneity.

The inclusion geometry of the third test case was the same as in the first test, as shown in the top left image of Figure 4.4. This time, the optical parameters were chosen so that the unit disc could model a neonatal head of radius 25 mm [2]: $\kappa = 0.05$, $\mu_a = 0.5$, and $\mu_a = 0.05$. The noisy measurements were simulated by adding Gaussian random noise with standard deviation $\epsilon \geq 0$ times the maximum element of $P\Lambda P \in \mathbb{R}^{9\times 9}$ to each element of the matrix $P(\Lambda - \Lambda_0)P \in \mathbb{R}^{9\times 9}$. The results are illustrated in Figure 4.4, where the top right image shows the indicator function α_δ for $\epsilon = 0$ and $\delta = 0.01$, the bottom left for $\epsilon = 10^{-4}$ and $\delta = 0.03$, and the bottom right for $\epsilon = 10^{-3}$ and $\delta = 0.1$. As Figure 4.4 demonstrates, the shape of the inclusion is not reconstructed correctly even if the data contains no noise. Compared to the first test case, this shortcoming is probably due to the higher level of ill-posedness caused by the introduction of the realistic optical parameters. However, one cannot rule out the possibility that the injectivity condition of Theorem 3.1 also plays a role in the relatively poor quality of the reconstructions. On the positive side, the graph of the indicator function contains information on the location of the inclusion with all three noise levels. The amount of additive noise used in the computations can be put into perspective by noticing that the maximal absolute value of the elements of

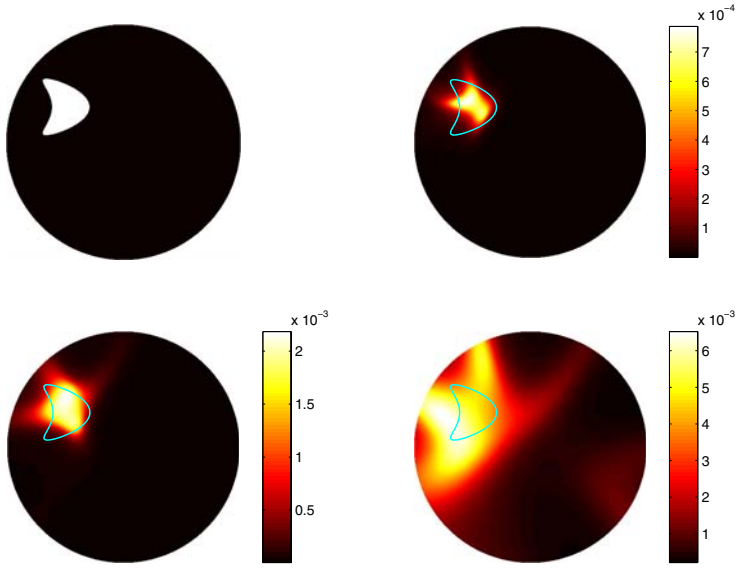


FIG. 4.4. *The third test case with the optical parameters* $\kappa = 0.05$, $\mu_a = 0.5$, *and* $\mu_a = 0.05$. *The graph of the indicator function* α_δ *is compared with the original kite-shaped transparent cavity for three different noise levels. Top left: The original inclusion. Top right:* $\epsilon = 0$ *and* $\delta = 0.01$. *Bottom left:* $\epsilon = 10^{-4}$ *and* $\delta = 0.03$. *Bottom right:* $\epsilon = 10^{-3}$ *and* $\delta = 0.1$.

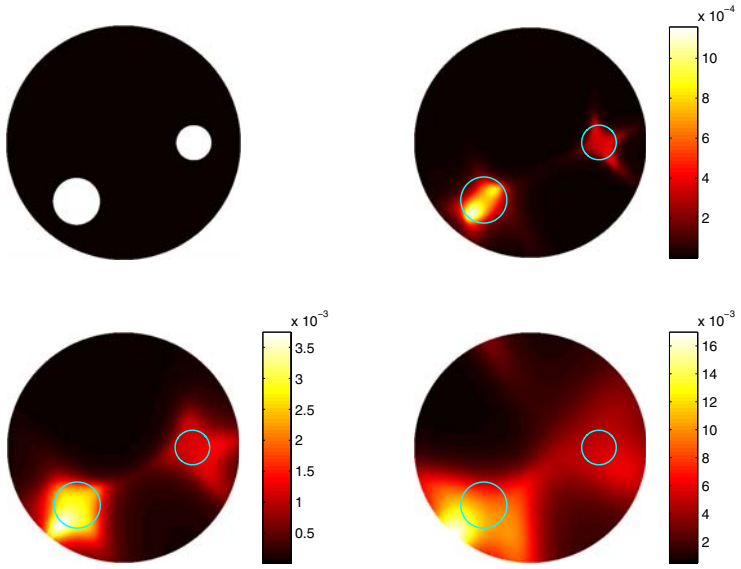


FIG. 4.5. *The fourth test case with the optical parameters* $\kappa = 0.05$, $\mu_a = 0.5$, *and* $\mu_a = 0.05$. *The graph of the indicator function* α_δ *is compared with the original circular transparent cavities for three different noise levels. Top left: The original inclusions. Top right:* $\epsilon = 0$ *and* $\delta = 0.01$. *Bottom left:* $\epsilon = 10^{-4}$ *and* $\delta = 0.03$. *Bottom right:* $\epsilon = 10^{-3}$ *and* $\delta = 0.1$.

$P\Lambda P$ is 7.7, whereas the maximal absolute value of the elements of $P(\Lambda - \Lambda_0)P$ is only 0.048.

Figure 4.5 illustrates the results of the fourth and final test case, where we used

the same optical parameters as in the third test. The top left image of Figure 4.5 shows the original nonscattering region that consists of two distinct circular inclusions of the type D_1. The top right image of Figure 4.5 shows the indicator function α_δ for $\epsilon = 0$ and $\delta = 0.01$, the bottom left for $\epsilon = 10^{-4}$ and $\delta = 0.03$, and the bottom right for $\epsilon = 10^{-3}$ and $\delta = 0.1$. Both inclusions are visible but malformed in all three reconstructions.

5. Conclusions. We have shown that, within the framework of the radiosity-diffusion model of OAST, the factorization method provides means to extract information on the transparent cavities embedded in known strongly scattering background from boundary measurements. Although the quality of the obtained reconstructions is quite sensitive to noise and depends strongly on the optical properties of the diffusive background and the absorption inside the nonscattering cavities, the results are quite promising because using sampling-type techniques, as is the factorization method, to locate the transparent regions seems straightforward compared to implementing Newton-type methods that search for the optimal inclusion shape iteratively.

In this work, we considered only the situation where the optical properties of the diffusive region surrounding the cavities are known in advance. However, in practice one is ultimately interested in locating absorbing inhomogeneities in the strongly scattering tissue. In consequence, to make the factorization method more useful for the further development of OAST, the possibility of characterizing transparent and absorbing inclusions simultaneously should be considered. This observation and the testing of the proposed method with measured data provide interesting subjects for future studies.

REFERENCES

[1] G. Arfken, *Mathematical Methods for Physicists*, Academic Press, New York, 1968.

[2] S. R. Arridge, *Optical tomography in medical imaging*, Inverse Problems, 15 (1999), pp. R41–R93.

[3] S. R. Arridge, *Diffusion tomography in dense media*, in Scattering and Inverse Scattering in Pure and Applied Science, Vol. 1, R. Pike and P. Sabatier, eds., Academic Press, San Diego, CA, 2002, pp. 920–936.

[4] S. R. Arridge, H. Dehghani, M. Schweiger, and D. T. Delpy, *The finite element model for the propagation of light in scattering media: A direct method for domains with non-scattering regions*, Medical Phys., 27 (2000), pp. 252–264.

[5] S. R. Arridge and J. C. Hebden, *Optical imaging in medicine:* II. *Modelling and reconstruction*, Phys. Med. Biol., 42 (1997), pp. 841–853.

[6] G. Bal, *Reconstructions in impedance and optical tomography with singular interfaces*, Inverse Problems, 21 (2005), pp. 113–131.

[7] M. Brühl, *Explicit characterization of inclusions in electrical impedance tomography*, SIAM J. Math. Anal., 32 (2001), pp. 1327–1341.

[8] M. Brühl and M. Hanke, *Numerical implementation of two noniterative methods for locating inclusions by impedance tomography*, Inverse Problems, 16 (2000), pp. 1029–1042.

[9] K. M. Case and P. F. Zweifel, *Linear Transport Theory*, Addison-Wesley, New York, 1967.

[10] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 2, Springer-Verlag, Berlin, 1988.

[11] H. Dehghani, S. R. Arridge, M. Schweiger, and D. T. Delpy, *Optical tomography in the presence of void regions*, J. Opt. Soc. Amer., 17 (2000), pp. 1659–1670.

[12] M. Firbank, S. R. Arridge, M. Schweiger, and D. T. Delby, *An investigation of light transport through scattering bodies with non-scattering regions*, Phys. Med. Biol., 41 (1996), pp. 767–783.

[13] B. Gebauer, *The factorization method for real elliptic problems*, J. Anal. Appl., 25 (2006), pp. 81–102.

[14] A. P. Gibson, J. C. Hebden, and S. R. Arridge, *Recent advances in diffuse optical tomography*, Phys. Med. Biol., 50 (2005), pp. R1–R43.

[15] M. Hanke and M. Brühl, *Recent progress in electrical impedance tomography*, Inverse Problems, 19 (2003), pp. S65–S90.
[16] J. C. Hebden, S. R. Arridge, and D. T. Delpy, *Optical imaging in medicine:* I. *Experimental techniques*, Phys. Med. Biol., 42 (1997), pp. 825–840.
[17] J. Heino, S. R. Arridge, J. Sikora, and E. Somersalo, *Anisotropic effects in highly scattering media*, Phys. Rev. E, 68 (2003), paper 31908.
[18] J. Heino and E. Somersalo, *Estimation of optical absorption in anisotropic background*, Inverse Problems, 18 (2002), pp. 559–573.
[19] N. Hyvönen, *Analysis of optical tomography with non-scattering regions*, Proc. Edinburgh Math. Soc., 45 (2002), pp. 257–276.
[20] N. Hyvönen, *Complete electrode model of electrical impedance tomography: Approximation properties and characterization of inclusions*, SIAM J. Appl. Math., 64 (2004), pp. 902–931.
[21] N. Hyvönen, *Characterizing inclusions in optical tomography*, Inverse Problems, 20 (2004), pp. 737–751.
[22] N. Hyvönen, *Diffusive Tomography Methods: Special Boundary Conditions and Characterization of Inclusions*, D.Sc. dissertation, Department of Engineering Physics and Mathematics, Helsinki University of Technology, Espoo, Finland, 2004; available online at http://lib.tkk.fi/Diss/2004/isbn9512270862/.
[23] N. Hyvönen, *Application of a weaker formulation of the factorization method to the characterization of absorbing inclusions in optical tomography*, Inverse Problems, 21 (2005), pp. 1331–1343.
[24] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer-Verlag, Berlin, 2004.
[25] A. Kirsch, *Characterization of the shape of a scattering obstacle using the spectral data of the far field operator*, Inverse Problems, 14 (1998), pp. 1489–1512.
[26] A. Kirsch, *The factorization method for a class of inverse elliptic problems*, Math. Nachr., 278 (2005), pp. 258–277.
[27] O. A. Ladyzhenskaya, *The Boundary Value Problems of Mathematical Physics*, Springer, New York, 1985.
[28] E. Okada, M. Firbank, M. Schweiger, S. R. Arridge, M. Cope, and D. T. Delby, *Theoretical and experimental investigation of near-infrared light propagation in a model of the adult head*, Appl. Optics, 36 (1997), pp. 21–31.
[29] J. Riley, H. Dehghani, M. Schweiger, S. Arridge, J. Ripoll, and M. Nieto-Vesperinas, 3*D optical tomography in the presence of void regions*, Opt. Express, 7 (2000), pp. 462–467.

SIAM J. Appl. Math.
Vol. 67, No. 4, pp. 1124–1146

RECONSTRUCTION OF THE SHAPE AND SURFACE IMPEDANCE FROM ACOUSTIC SCATTERING DATA FOR AN ARBITRARY CYLINDER*

J. J. LIU†, G. NAKAMURA‡, AND M. SINI§

Abstract. The inverse scattering for an obstacle $D \subset R^2$ with mixed boundary condition can be considered as a prototype for radar detection of complex obstacles with coated and noncoated parts of the boundary. We construct some indicator functions for this inverse problem using the far-field pattern directly, without the necessity of transforming the far field to the near field. Based on careful singularity analysis, these indicator functions enable us to reconstruct the shape of the obstacle and distinguish the coated from the noncoated part of the boundary. Moreover, an explicit representation formula for the surface impedance in the coated part of the boundary is also given. Our reconstruction scheme reveals that the coated part of the obstacle is less visible than the noncoated one, which corresponds to the physical fact that the coated boundary absorbs some part of the scattered wave. Numerics are presented for the reconstruction formulas, which show that both the boundary shape and the surface impedance in the coated part of the boundary can be reconstructed accurately. The theoretical reconstruction techniques proposed in this work can be applied in the practical 3-dimensional electromagnetic inverse scattering problems with promising numerical performance. Such problems are of great importance in the design of nondetectable obstacles.

Key words. inverse scattering, far field, impedance boundary, singularity analysis, numerics

AMS subject classifications. 35P25, 35R30, 78A45

DOI. 10.1137/060654220

Introduction and examples. Inverse scattering problems aim to identify some properties of an obstacle such as the boundary shape and type from the information contained in the scattered wave for given incident waves. Optimization techniques are well known for reconstructing the obstacle, up to some accuracy, by minimizing the objective functional for an unknown obstacle from given inversion input data by iteration procedures. However, it seems that a good initial guess is needed.

In recent years, some new inversion methods for the reconstruction of obstacle boundaries have been proposed. The common idea of these methods is the construction of some indicator functions from given inversion input data, which depend on some detecting point (a parameter) varying inside or outside the obstacle. When this point approaches the obstacle, these indicator functions blowup. The linear sampling method [7], the factorization method [16], and the singular sources method [21] construct the indicator functions from the far-field pattern directly, while the probe method [13, 14] constructs the indicator in terms of the near field. The near field

*Received by the editors March 13, 2006; accepted for publication (in revised form) January 17, 2007; published electronically May 14, 2007.
http://www.siam.org/journals/siap/67-4/65422.html

†Department of Mathematics, Southeast University, Nanjing, 210096, P.R. China (jjliu@seu.edu.cn). This author was supported by NSFC grant 10371018.

‡Department of Mathematics, Hokkaido University, Sapporo, 060-0810, Japan (gnaka@math.sci.hokudai.ac.jp). This author was partially supported by Grant-in-Aid for Scientific research (B)(2) (N0.14340038) of the Japan Society for Promotion of Science.

§Department of Mathematics, Yonsei University, Seoul, 120-749, Korea. Current address: Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstrasse 69, Linz, A-4040, Austria (mourad.sini@oeaw.ac.at). This author was supported by the IIRC of Kyung Hee University, Korea, via the SRC/ERC program of MOST/KOSEF (R11-2002-103).

can be obtained from the far field by some regularization procedures [21]. However, we can also state the natural version of the probe method directly from the far-field data, without reducing the far field to the near field; see [11]. For a review of these methods, the readers are referred to [22, 23], and for some relations between them, to [11, 20].

If the scattering is caused by multiple obstacles with different types of boundary or with mixed boundary condition, one should identify both the boundary shape, boundary type, and surface impedance. These kinds of problems come from some industry designs such as radar detection by electromagnetic wave scattering; see [10]. The obstacle is illuminated by an electromagnetic wave coming from an antenna. The wave is scattered by the obstacle and received by an antenna located in a different place. One of the objectives is to design the shape of the obstacle such that a reflected wave can be avoided or minimized in some directions. One possible approach to this goal is to introduce a coating on the surface of the obstacle or on some of its parts. This is motivated by the fact that reflections are minimized by applying such a surface coating. The surface coating is modeled by introducing an impedance boundary condition on a part or on the whole surface of the scatterer, which gives a relation between the electric and the magnetic field through a coefficient called surface impedance.

Due to this practical importance, the reconstruction of boundary impedance has been addressed by many authors. In [1], the authors construct the inhomogeneous boundary impedance for a cylinder obstacle with known shape using only one incident wave, assuming that the surface impedance is distributed along the whole boundary of the obstacle. In this case, the scattering of electromagnetic waves can be described by the 2-dimensional Helmholtz equation. We also refer to [17], where an optimization method is applied. After reducing the far field to the near field, a moment method is suggested in [6] to reconstruct the surface impedance approximately in the case of a completely coated obstacle, and the identification of different types of multiple obstacles is given in [5] in the case where on each obstacle we have one type of boundary condition.

The problem of whether a part of the surface of the obstacle is coated or not is important. Answering this question and reconstructing the surface impedance, in case of coating, from far-field measurements is our main object. In this work, we restrict ourselves to the acoustic wave scattering governed by a 2-dimensional Helmholtz equation, noticing that the 3-dimensional electromagnetic wave scattering in the cylinder case can be modeled by the 2-dimensional Helmholtz equation [8]. These issues were first considered in [2, 3] by the linear sampling method, where the authors simultaneously reconstruct the obstacle and compute the L^∞-norm of the surface impedance. This can be used to answer the question of existence or absence of coating and to give the value of the surface impedance in case it is known to be constant.

Motivated by these last works, our aim is to give another way to consider these issues and give further information on the obstacle. We proceed by constructing some indicator functions giving a direct link between the far-field pattern and the physical parameters of the obstacle. More precisely, we establish pointwise formulas which enable us to detect the boundary of the scatterer and distinguish and recognize the coated and the noncoated parts of the obstacle surface. In addition, on the coated part of the obstacle, the indicator functions give explicitly the pointwise values of this surface impedance as a functional of the far fields. These types of results have been initiated in [19], where the theoretical justification of these formulas in

3-dimensional acoustic scattering is given. Since we need more singularity analysis in the 2-dimensional case than in 3-dimensional case, which is due to the use of a more singular point source, we give the theoretical justification of the steps where it is necessary, and refer to [19] for the rest of the proof. We would like to emphasize that we are reconstructing the obstacle, localizing the eventual coated part and reconstructing the surface impedance in one step, i.e., simultaneously; compare with [1, 2, 3, 5, 6, 17]. Also, since the analysis is done pointwise, we can also consider multiple obstacles and give similar results.

The validity of the theoretical reconstruction formula presented in this paper is also checked by numerical tests with satisfactory performances. We would like to mention the following observations from the numerics. The coated part of an obstacle with larger impedance is less visible than the other part in terms of the value of the indicator. This explains the practical motivation for introducing the coating, i.e., to avoid or perturb the detection of an obstacle by applying an absorbing boundary layer. On the other hand, for nonconvex obstacles with mixed boundary conditions, the inversion formulas proposed in this paper also generate a satisfactory reconstruction by combining different blowing-up criterion together. These reconstruction performances are supported by our numerical implementations given in the last section of this paper.

The rest of the paper is organized as follows. In section 1, we state the problem mathematically. In section 2, we present the results, which we prove in section 3. Section 4 is devoted to the numerical tests.

1. Statement of the problem. Let D be a bounded domain of R^2 such that $R^2 \setminus \overline{D}$ is connected. We assume that its boundary ∂D is of class C^2 and has the following form:

$$\partial D = \overline{\partial D_I} \cup \overline{\partial D_D}, \quad \partial D_I \cap \partial D_D = \emptyset,$$

where ∂D_D and ∂D_I are open surfaces in ∂D.

The propagation of time-harmonic acoustic fields in homogeneous cylinder media can be modeled by the Helmholtz equation

$$\Delta u + \kappa^2 u = 0 \quad \text{in} \quad R^2 \setminus \overline{D}, \tag{1.1}$$

where $\kappa > 0$ is the wave number. At the part ∂D_I of the obstacle boundary, we assume the total field u that satisfies the impedance boundary condition, while the part ∂D_D satisfies the Dirichlet boundary condition. That is,

$$\frac{\partial u}{\partial \nu} + i\kappa\sigma u = 0 \text{ on } \partial D_I \tag{1.2}$$

with some impedance function σ and

$$u = 0 \text{ on } \partial D_D, \tag{1.3}$$

where ν is the outward unit normal of ∂D. We assume that σ is a real valued Holder continuous function of order $\beta \in (0, 1]$ and has a uniform lower bound $\sigma_- > 0$ on ∂D_I. The part ∂D_I is referred to by the coated part of ∂D, and ∂D_D is the noncoated part.

For a given incident plane wave $u^i(x, d) = e^{i\kappa d \cdot x}$ with incident direction $d \in S^1$, where S^1 is a unit circle in R^2, we look for a solution $u := u^i + u^s$ of (1.1), (1.2), and (1.3), where the scattered field u^s satisfies the Sommerfeld radiation condition

$$\lim_{r\to\infty} \sqrt{r} \left(\frac{\partial u^s}{\partial r} - i\kappa u^s \right) = 0 \tag{1.4}$$

with $r = |x|$ and the limit is uniform for all directions $\hat{x} \in S^1$.

The mixed problem (1.1)–(1.4) is well posed. More generally, for $f \in H^{\frac{1}{2}}(\partial D_D)$ and $h \in H^{-\frac{1}{2}}(\partial D_I)$, there exists a unique solution of the mixed problem

$$\begin{cases} (\Delta + \kappa^2)u = 0 \quad \text{in } R^2 \setminus \overline{D}, \\ u = f \quad \text{on } \partial D_D, \\ \frac{\partial u}{\partial \nu} + i\kappa\sigma u = h \quad \text{on } \partial D_I, \\ \lim_{r\to\infty} \sqrt{r}(\frac{\partial u}{\partial r} - i\kappa u) = 0, \end{cases} \tag{1.5}$$

and the solution satisfies

$$\|u\|_{H^1(\Omega_R \cap (R^2 \setminus \overline{D}))} \le C_R(\|f\|_{H^{1/2}(\partial D_D)} + \|h\|_{H^{-\frac{1}{2}}(\partial D_I)}), \tag{1.6}$$

where Ω_R is a disk of radius R and C_R is positive constant depending on R; see [4] for more details.

It is well known (see [8]) that the scattered wave has the asymptotic behavior

$$u^s(x,d) = \frac{e^{i\kappa r}}{\sqrt{r}} u^\infty(\hat{x},d) + O(r^{-3/2}), \quad r := |x| \to \infty, \tag{1.7}$$

where the function $u^\infty(\cdot, d)$ defined on S^1 is called the far field of the scattered wave u^s corresponding to incident direction d. We introduce a constant $\gamma_2 := \frac{e^{i\pi/4}}{\sqrt{8\pi\kappa}}$ and

$$\Phi(x,y) := \frac{i}{4} H_0^{(1)}(\kappa|x-y|), \quad x \neq y, x, y \in R^2,$$

the fundamental solution to the Helmholtz equation in R^2, where $H_0^{(1)}$ is the Hankel function of the first kind of order zero. In this paper, we will consider the following.

Inverse scattering problem for an obstacle with mixed boundary type. Given $u^\infty(\cdot,\cdot)$ on $S^1 \times S^1$ for the scattering problem (1.1)–(1.4), reconstruct the shape of obstacle D, identify ∂D_I and ∂D_D, and reconstruct the surface impedance $\sigma(x)$ on ∂D_I.

2. Presentation of the results. It is well known also (see [8]) that the scattered field associated with the Herglotz incident field $v_g^i := v_g$ defined by

$$v_g(x) := \int_{S^1} e^{i\kappa x\cdot d} g(d)\, ds(d), \quad x \in R^2, \tag{2.1}$$

with $g \in L^2(S^1)$ is given by

$$v_g^s(x) := \int_{S^1} u^s(x,d) g(d)\, ds(d), \quad x \in R^2 \setminus D, \tag{2.2}$$

and its far field is

$$v_g^\infty(\hat{x}) := \int_{S^1} u^\infty(\hat{x},d) g(d)\, ds(d), \quad \hat{x} \in S^1. \tag{2.3}$$

We will need the identity (see [8])

$$u^\infty(\hat{x},d) = -\gamma_2 \int_{\partial D} \left\{ \frac{\partial u^s(y,d)}{\partial \nu} e^{-i\kappa\hat{x}\cdot y} - \frac{\partial e^{-i\kappa\hat{x}\cdot y}}{\partial \nu} u^s(y,d) \right\} ds(y) \tag{2.4}$$

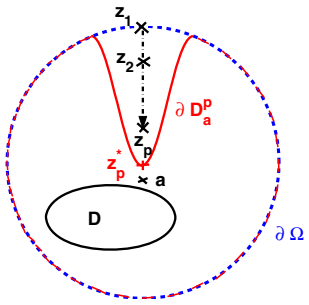


FIG. 2.1. *Geometric configuration.*

and the representation formula for the scattered wave $\Phi^s(\cdot, z)$ in $R^2 \setminus \overline{D}$ for the point source $\Phi(\cdot, z)$,

(2.5)
$$\Phi^s(x,z) = -\int_{\partial D} \left\{ \frac{\partial \Phi^s(y,z)}{\partial \nu(y)} \Phi(x,y) - \Phi^s(y,z) \frac{\partial \Phi(x,y)}{\partial \nu(y)} \right\} ds(y), \quad x, z \in R^2 \setminus \overline{D}.$$

Assume that $\overline{D} \subset\subset \Omega$ for some known Ω with smooth boundary. For $a \in \Omega \setminus D$, denote by $\{z_p\} \subset \Omega \setminus \overline{D}$ a sequence tending to a. For any z_p, set D_a^p a C^2-regular domain such that $\overline{D} \subset D_a^p$ with $z_q \in \Omega \setminus \overline{D_a^p}$ for every $q = 1, 2, \ldots, p$ and such that the Dirichlet interior problem on D_a^p for the Helmholtz equation is uniquely solvable; see Figure 2.1 for the configuration. In this case, the Herglotz wave operator $\mathbb{H}$ defined from $L^2(S^1)$ to $L^2(\partial D_a^p)$ by

$$\mathbb{H}[g](x) := v_g(x) = \int_{S^1} e^{i\kappa x \cdot d} g(d)\, ds(d) \tag{2.6}$$

is injective compact with dense range; see [8]. Let z_p^* be a point on ∂D_a^p near z_p such that $z_p^* \to a$ as $z_p \to a$, as chosen in Figure 2.1. Denote by $\nu(z_p^*)$ the outward normal of ∂D_a^p at z_p^*. Now we consider the sequence of point sources $\Phi(\cdot, z_p)$. For every p fixed, we construct two density sequences $\{g_n^p\}$ and $\{f_m^p\}$ in $L^2(S^1)$ by the Tikhonov regularization such that

$$\|v_{g_n^p} - \Phi(\cdot, z_p)\|_{L^2(\partial D_a^p)} \to 0, \quad n \to \infty, \tag{2.7}$$

$$\left\| v_{f_m^p} - \frac{\partial}{\partial \nu(z_p^*)} \Phi(\cdot, z_p) \right\|_{L^2(\partial D_a^p)} \to 0, \quad m \to \infty, \tag{2.8}$$

where $\partial_{\nu(z_p^*)} \Phi(\cdot, z_p) := \nabla_x \Phi(x, z_p) \cdot \nu(z_p^*)$. Since both $v_{g_n^p}$ and $\Phi(\cdot, z_p)$ satisfy the same Helmholtz equation in D_a^p, (2.7) implies that

$$\|v_{g_n^p} - \Phi(\cdot, z_p)\|_{H^{\frac{1}{2}}(\partial D)} \to 0, \quad n \to \infty, \tag{2.9}$$

and

$$\left\| \frac{\partial}{\partial \nu} v_{g_n^p} - \frac{\partial}{\partial \nu} \Phi(\cdot, z_p) \right\|_{H^{-\frac{1}{2}}(\partial D)} \to 0, \quad n \to \infty. \tag{2.10}$$

Similarly, it follows from (2.8) that

$$\left\| v_{f_m^p} - \frac{\partial}{\partial \nu(z_p^*)} \Phi(\cdot, z_p) \right\|_{H^{\frac{1}{2}}(\partial D)} \to 0, \quad m \to \infty, \tag{2.11}$$

and

$$\left\| \frac{\partial}{\partial \nu} v_{f_m^p} - \frac{\partial}{\partial \nu} \left(\frac{\partial}{\partial \nu(z_p^*)} \Phi(\cdot, z_p) \right) \right\|_{H^{-\frac{1}{2}}(\partial D)} \to 0, \quad m \to \infty. \tag{2.12}$$

Multiplying (2.4) by $f_m^p(d) g_n^p(\hat{x})$ and integrating over $S^1 \times S^1$, we have

$$\begin{aligned} & - \int_{S^1} \int_{S^1} u^\infty(-\hat{x}, d) f_m^p(d) g_n^p(\hat{x}) \, ds(\hat{x}) ds(d) \\ &= \gamma_2 \int_{\partial D} \left\{ \int_{S^1} \frac{\partial u^s(y,d)}{\partial \nu} f_m^p(d) \, ds(d) \cdot \int_{S^1} e^{i\kappa \hat{x} \cdot y} g_n^p(\hat{x}) \, ds(\hat{x}) \right. \\ &\qquad \left. - \int_{S^1} \frac{\partial e^{i\kappa \hat{x} \cdot y}}{\partial \nu} g_n^p(\hat{x}) \, ds(\hat{x}) \cdot \int_{S^1} u^s(y,d) f_m^p(d) \, ds(d) \right\} ds(y) \\ &= \gamma_2 \int_{\partial D} \left\{ \frac{\partial v^s_{f_m^p}}{\partial \nu}(y) v^i_{g_n^p}(y) - \frac{\partial v^i_{g_n^p}}{\partial \nu}(y) v^s_{f_m^p}(y) \right\} ds(y). \end{aligned} \tag{2.13}$$

From (2.9), (2.10), and (2.13), we have

$$\begin{aligned} & \lim_{n \to \infty} \int_{S^1} \int_{S^1} u^\infty(-\hat{x}, d) \, f_m^p(d) \, g_n^p(\hat{x}) \, ds(\hat{x}) ds(d) \\ &= \gamma_2 \int_{\partial D} \left\{ v^s_{f_m^p} \frac{\partial \Phi(y, z_p)}{\partial \nu(y)} - \frac{\partial v^s_{f_m^p}}{\partial \nu(y)} \Phi(y, z_p) \right\} ds(y) \\ &= \gamma_2 v^s_{f_m^p}(z_p) \end{aligned} \tag{2.14}$$

from the Green formula, where $v^s_{f_m^p}(\cdot)$ is the scattered wave corresponding to incident wave $v^i_{f_m^p}(x) = \mathbb{H}[f_m^p](x)$.

Denote by $E^s(x, z_p)$ the scattered wave corresponding to the incident wave $\frac{\partial \Phi(x, z_p)}{\partial \nu(z_p^*)}$, which is well defined for every $x \in R^2 \setminus \overline{D}$. Then it follows from (2.11), (2.12), the well-posedness of the direct scattering problem, and the use of interior estimate that

$$E^s(x, z_p) = \lim_{m \to \infty} v^s_{f_m^p}(x), \quad x \in R^2 \setminus \overline{D}. \tag{2.15}$$

Finally, it follows from (2.14) that

$$\lim_{m \to \infty} \lim_{n \to \infty} \int_{S^1} \int_{S^1} u^\infty(-\hat{x}, d) \, f_m^p(d) \, g_n^p(\hat{x}) \, ds(\hat{x}) ds(d) = \gamma_2 E^s(z_p, z_p). \tag{2.16}$$

The reconstruction of ∂D as well as its surface impedance in the coating part can be established based on (2.16). For this purpose, an analysis of $E^s(x, z)$ near ∂D is the key point. We need the natural C^2 smoothness assumption on the regularity of ∂D. Precisely, for every point $a \in \partial D$, there exists a rigid transformation of coordinates under which the image of a is $\mathbf{0}$ and a function $f \in C^2(-r, r)$ such that

$$f(0) = \frac{df}{dx}(0) = 0, \quad D \cap B(\mathbf{0}, r) = \{(x, y) \in B(\mathbf{0}, r); y > f(x)\} \tag{2.17}$$

in terms of the new coordinates, where $B(\mathbf{0}, r)$ is the 2-dimensional ball of center $\mathbf{0}$ with radius r.

For the points $a \in \partial D$, we choose the sequence $\{z_p\}_{p \in \mathbb{N}}$ included in $C_{a,\theta}$, where $C_{a,\theta}$ is a cone with center a, angle $\theta \in [0, \frac{\pi}{2})$, and axis $\nu(a)$. The main theoretical result of this paper is as follows.

THEOREM 2.1. *Assume that the boundary ∂D is of class C^2 and that σ is a real valued Holder continuous function with positive lower bound. Then the boundary properties of the obstacle D can be identified by the following indicator functions:*

1. *The obstacle boundary ∂D can be constructed from the following property:*

$$\lim_{p\to\infty}\lim_{m,n\to\infty}\left|\mathrm{Re}\left[\gamma_2^{-1}\int_{S^1}\int_{S^1}u^\infty(-\hat{x},d)f_m^p(d)g_n^p(\hat{x})\,ds(\hat{x})ds(d)\right]\right| = \begin{cases} +\infty, a\in\partial D,\\ <+\infty, \qquad a\in\Omega\setminus\overline{D}.\end{cases} \tag{2.18}$$

Precisely, we have the blowup rate

$$\lim_{m,n\to\infty}\mathrm{Re}\left[\gamma_2^{-1}\int_{S^1}\int_{S^1}u^\infty(-\hat{x},d)f_m^p(d)g_n^p(\hat{x})\,ds(\hat{x})ds(d)\right] = \frac{\pm 1}{4\pi|(z_p-a)\cdot\nu(a)|} + O(|\ln|z_p-a||^2), \tag{2.19}$$

where $z_p := (z_{p,1}, z_{p,2})$ and $a = (a_1, a_2)$. The sign $(+)$ is for $a\in\partial D_D$, while the sign $(-)$ is for $a\in\partial D_I$.

2. *The coated and the noncoated parts of ∂D can also be distinguished from the following properties*

$$\lim_{p\to\infty}\lim_{m,n\to\infty}\frac{\mathrm{Im}\left[\gamma_2^{-1}\int_{S^1}\int_{S^1}u^\infty(-\hat{x},d)f_m^p(d)g_n^p(\hat{x})\,ds(\hat{x})ds(d)\right]}{|\ln|(z_p-a)\cdot\nu(a)||^s} = \begin{cases}+\infty, & a\in\partial D_I,\\ 0, & a\in\partial D_D,\end{cases} \tag{2.20}$$

by choosing any fixed $s\in(0,1)$.

3. *The impedance coefficient on ∂D_I can be detected by the following formula:*

$$\lim_{p\to\infty}\lim_{m,n\to\infty}\frac{\mathrm{Im}\left[\gamma_2^{-1}\int_{S^1}\int_{S^1}u^\infty(-\hat{x},d)f_m^p(d)g_n^p(\hat{x})\,ds(\hat{x})ds(d)\right]}{|\ln|(z_p-a)\cdot\nu(a)||} = \frac{\kappa}{\pi}\sigma(a), \quad a\in\partial D_I. \tag{2.21}$$

REMARK 2.2. *The formula* (2.18) *is also true if f_m^p is replaced by g_n^p. That is, the singularity of $\Phi(x,z_p)$ is theoretically enough for identifying ∂D. However, as the blowup rate in this 2-dimensional case is of logarithmic order, it is not suitable to localize the obstacle clearly in the numerical experiments. For this reason, we introduced the density f_m^p, which is related to a stronger singularity $\frac{\partial}{\partial\nu(z_p^*)}\Phi(\cdot,z_p)$, to get a blowup rate of order $|z_p-a|^{-1}$. For the formulas* (2.20) *and* (2.21)*, the stronger singularity of $\frac{\partial}{\partial\nu(z_p^*)}\Phi(\cdot,z_p)$ is necessary. Moreover, we can in fact use $\nu(a)$ instead of $\nu(z_p^*)$, since ∂D has been determined in terms of* (2.18)*. The formula* (2.19) *can also be used to distinguish the coated part ∂D_I from the noncoated ∂D_I.*

REMARK 2.3. *Theorem* 2.1 *is stated for the case of a single obstacle. However, these results are still true for the multiple obstacle case with coated and noncoated parts.*

REMARK 2.4. *If $a\in D$, then the limit in* (2.18) *is conjectured to be ∞; see* [11]*. However, up to now, we do not have the full answer. The approach in* [15] *can be used to justify it in the case where the frequency κ is small enough.*

3. Proof of Theorem 2.1. For any given point $a\in\partial D$, we first take the rotation R_a and the translation M_a such that

$$R_a(\nu(a)) = (0,1), \quad R_a(a) + M_a = \mathbf{0}$$

in the new coordinate system $\tilde{x}$. Under the transform $\tilde{x} := \mathbb{T}(x) := R_a(x) + M_a$, it follows that

$$\mathbb{T}(\nu(a)) = (0,1), \quad \mathbb{T}(a) = \mathbf{0}.$$

Define $\tilde{\sigma}(\tilde{x}) := \sigma(x)$ and consider the following two problems in the coordinate $\tilde{x} = (\tilde{x}_1, \tilde{x}_2)$ for any given $\tilde{z} = (\tilde{z}_1, \tilde{z}_2) \in R^2_+$. We set $\tilde{w}^+_{\tilde{\sigma}(\mathbf{0})}(\tilde{x}, \tilde{z})$ and $\tilde{w}^+_D(\tilde{x}, \tilde{z})$ to be two functions satisfying

$$\text{(3.1)} \quad \begin{cases} \Delta \tilde{w}^+_{\tilde{\sigma}(\mathbf{0})} = 0, \quad \tilde{x} \in R^2_+, \\ (\frac{\partial}{\partial \tilde{x}_2} \tilde{w}^+_{\tilde{\sigma}(\mathbf{0})} + i\kappa\tilde{\sigma}(\mathbf{0}) \tilde{w}^+_{\tilde{\sigma}(\mathbf{0})})(\tilde{x}, \tilde{z})|_{\tilde{x}_2=0} = -(\frac{\partial}{\partial \tilde{x}_2} + i\kappa\tilde{\sigma}(\mathbf{0})) \frac{\partial}{\partial x_2} \Gamma(\tilde{x}, \tilde{z})|_{\tilde{x}_2=0}, \end{cases}$$

$$\text{(3.2)} \quad \begin{cases} \Delta \tilde{w}^+_D = 0, \quad \tilde{x} \in R^2_+, \\ \tilde{w}^+_D(\tilde{x}, \tilde{z})|_{\tilde{x}_2=0} = -\frac{\partial}{\partial x_2} \Gamma(\tilde{x}, \tilde{z})|_{\tilde{x}_2=0}, \end{cases}$$

respectively, where $\Gamma(\tilde{x}, \tilde{z}) = \frac{1}{2\pi} \ln \frac{1}{|\tilde{x} - \tilde{z}|}$ and the subscript D in $\tilde{w}^+_D(\tilde{x}, \tilde{z})$ refers to the Dirichlet boundary condition in (3.2).

We give explicit solutions to these two problems in the following proposition.

PROPOSITION 3.1. *We have the explicit form of* $w^+_{\tilde{\sigma}(\mathbf{0})}(\tilde{x}, \tilde{z})$,

$$\text{(3.3)} \quad \tilde{w}^+_{\tilde{\sigma}(\mathbf{0})}(\tilde{x}, \tilde{z}) = \frac{1}{4\pi} \int_R e^{i(\tilde{x}_1 - \tilde{z}_1)\xi_1} e^{-(\tilde{x}_2 + \tilde{z}_2)|\xi_1|} \frac{|\xi_1| + i\kappa\tilde{\sigma}(\mathbf{0})}{|\xi_1| - i\kappa\tilde{\sigma}(\mathbf{0})} d\xi_1,$$

while $\tilde{w}^+_D(\tilde{x}, \tilde{z})$ *has the form*

$$\text{(3.4)} \quad \tilde{w}^+_D(\tilde{x}, \tilde{z}) = -\frac{1}{4\pi} \int_R e^{i(\tilde{x}_1 - \tilde{z}_1)\xi_1} e^{-(\tilde{x}_2 + \tilde{z}_2)|\xi_1|} d\xi_1.$$

This proposition can be proven by expressing

$$\tilde{w}^+_{\tilde{\sigma}(\mathbf{0})}(\tilde{x}, \tilde{z}) = (U_+[\tilde{x}_2]\phi_+)(\tilde{x}_1), \quad \tilde{w}^+_D(\tilde{x}, \tilde{z}) = (U_+[\tilde{x}_2]\phi_-)(\tilde{x}_1)$$

in R^2_+ with $(U_\pm[\tilde{x}_2]\phi)(\tilde{x}_1) := \frac{1}{2\pi} \int_R e^{i\tilde{x}_1\xi_1 \mp \tilde{x}_2|\xi_1|} \hat{\phi}(\xi_1, \tilde{z}) d\xi_1$ and computing the density functions $\phi_\pm$ from the boundary value problems (3.1), (3.2), where $\hat{\phi}$ is the 1-dimensional Fourier transform of ϕ; see [19] for explicit computations.

Define

$$w^+_{\sigma(a)}(x, z) = \tilde{w}^+_{\tilde{\sigma}(\mathbf{0})}(\mathbb{T}x, \mathbb{T}z), \quad w^+_D(x, z) = \tilde{w}^+_D(\mathbb{T}x, \mathbb{T}z)$$

for $x, z \in R^2 \setminus \overline{D}$ near a, which is well defined by the definition of $\mathbb{T}$.

The next proposition gives the relation between $E^s(x, z)$ and $w^+_{\sigma(a)}(x, z), w^+_D(x, z)$ near the point a.

PROPOSITION 3.2. *If* $a \in \partial D_I$, *then there exist* $\delta(a) > 0$ *and* $C > 0$ *such that*

$$\text{(3.5)} \quad |\text{Im } E^s(x, z) - \text{Im } w^+_{\sigma(a)}(x, z)| \le C \qquad \textit{for } (x, z) \in B_+(a, \delta(a)) \cap C_{a,\theta},$$

$$\text{(3.6)} \quad |\text{Re } E^s(x, z) - \text{Re } w^+_{\sigma(a)}(x, z)| \le C |\ln |x - a|| \cdot |\ln |z_p - a|| \qquad \textit{for } (x, z) \in B_+(a, \delta(a)) \cap C_{a,\theta},$$

where $B_+(a, \delta(a)) := B(a, \delta(a)) \cap (R^2 \setminus D)$ *and* $B(a, \delta(a))$ *is the ball of center* a *and radius* $\delta(a)$.

Similarly, if $a \in \partial D_D$, we obtain (3.5) *and* (3.6) *by replacing $w^+_{\sigma(a)}$ by w^+_D.*

REMARK 3.3. *The estimate of* $|\text{Re } E^s(x,z) - \text{Re } w^+_{\sigma(a)}(x,z)|$ *is not optimal. We do not need the term* $|\ln|x-a||$. *The upper bound in* (3.6) *can be replaced by* $C_\alpha|z-a|^{-\alpha}$ *for any* $\alpha > 0$, *where* C_α *depends on* α. *But to prove Theorem* 2.1 *the estimate given in* (3.6) *is enough.*

Now we can prove Theorem 2.1 based on these propositions.

Proof of Theorem 2.1.

Step A: It follows from Proposition 3.2 that

$$|\text{Re } E^s(x,z_p) - \text{Re } w^+(x,z_p)| \le C \ln \frac{1}{|z_p - a|} \tag{3.7}$$

uniformly for all x, z_p near any fixed point $a \in \partial D$, where $w^+(z_p,z_p)$ may be $w^+_{\sigma(a)}(z_p,z_p)$ or $w^+_D(z_p,z_p)$, depending on the position of a. For $w^+(z_p,z_p) = w^+_D(z_p,z_p)$, it follows from (3.4) that

$$\text{Re } w^+(z_p,z_p) = -\frac{1}{4\pi}\int_R e^{-2|z_{p,2}-a_2||\xi_1|}d\xi_1 = \frac{1}{4\pi|z_{p,2}-a_2|},$$

while for $w^+(z_p,z_p) = w^+_{\sigma(a)}(z_p,z_p)$ it holds from (3.3) that

$$\text{Re } w^+(z_p,z_p) = \frac{1}{4\pi}\int_R e^{-2|z_{p,2}-a_2||\xi_1|}\frac{|\xi_1|^2 - \kappa^2\sigma^2(a)}{|\xi_1|^2+\kappa^2\sigma^2(a)}d\xi_1 = -\frac{1}{4\pi|z_{p,2}-a_2|} + O(1),$$

where $z_p = (z_{p,1}, z_{p,2}) \to a = (a_1,a_2) \in \partial D$ as $p \to \infty$. The application of the above relations in (3.7) leads to (2.19) and then $\lim_{p\to\infty}|\text{Re } E^s(z_p,z_p)| = +\infty$. Now (2.18) is proven for $a \in \partial D$.

Suppose that a is outside $\overline{D}$. We can construct z_p^*, z_p tending to a as we did for $a \in \partial D$. Recall that $E^s(x,z_p)$ satisfies

$$\begin{cases} (\Delta+\kappa^2)E^s(x,z_p) = 0 \text{ in } R^2 \setminus \overline{D}, \\ E^s(\cdot,z_p) = -\frac{\partial\Phi}{\partial\nu(z_p^*)}(x,z_p) \text{ on } \partial D_D, \\ (\frac{\partial}{\partial\nu} + i\kappa\sigma(x))E^s(x,z_p) = -(\partial_\nu + i\kappa\sigma(x))\frac{\partial\Phi}{\partial\nu(z_p^*)}(x,z_p) \text{ on } \partial D_I, \\ E^s(\cdot,z_p) \text{ satisfies the Sommerfeld radiation conditions,} \end{cases}$$

where $\nu(z_p^*)$ is the unit outward normal on ∂D_a^p at the point z_p^*. Hence the boundary condition is bounded with respect to x in $H^{1/2}(\partial D_D)$ and $H^{-\frac{1}{2}}(\partial D_I)$, respectively, for z_p^*, z_p near a. It follows from the well-posedness of the direct problem and interior estimates near a (i.e., away from ∂D) that $E^s(x,z_p)$, and then $E^s(z_p,z_p)$, is bounded.

Step B. Let $a \in \partial D_I$. From (3.3) we have

$$\tilde{w}^+_{\tilde{\sigma}(\mathbf{0})}(\tilde{z},\tilde{z}) = \frac{1}{4\pi}\int_R e^{-2\tilde{z}_2|\xi_1|}\frac{|\xi_1| + i\kappa\tilde{\sigma}(\mathbf{0})}{|\xi_1| - i\kappa\tilde{\sigma}(\mathbf{0})}d\xi_1. \tag{3.8}$$

By taking the imaginary part and setting $\tilde{z} = (\tilde{z}_1,\tilde{z}_2) = R_a(z) + M_a$ for $z \in C(a,\theta)$, we get

$$\begin{aligned} \text{Im}(4\pi w^+_{\sigma(a)}(z,z)) &= 4\kappa\sigma(a)\int_0^{+\infty}\frac{e^{-2(z-a)\cdot\nu(a)r}}{r^2+\kappa^2\sigma(a)^2}r\,dr \\ &= 4\kappa\sigma(a)\Big[-\ln(\kappa\sigma(a)) - \ln((z-a)\cdot\nu(a)) \\ &\qquad + 2\int_0^{+\infty}\ln(r^2+\kappa^2|(z-a)\cdot\nu(a)|^2\sigma^2(a))e^{-2r}dr\Big], \end{aligned} \tag{3.9}$$

which leads to the first relation in (2.20) by dividing by $|\ln((z-a)\cdot\nu(a))|^s$ for $0<(z-a)\cdot\nu(a)<1$ with $0<s<1$ using (3.5) and (2.16). The representation (2.21) for $\sigma(a)$ can be gotten from the above relation by dividing by $|\ln((z-a)\cdot\nu(a))|$ for $0<(z-a)\cdot\nu(a)<1$.

Step C. Let $a\in\partial D_D$. Proposition 3.2 for $w_D^+(x,z)$ and (2.14) imply the second relation in (2.20), noticing the fact that $\mathrm{Im}w_D^+(z,z)=0$. □

The rest of this section is devoted to the proof of Proposition 3.2. As we said in the introduction, in the 2-dimensional case, we need more singularity analysis than in [19]. This is due to the use of the more singular point source $\frac{\partial}{\partial\nu(z_p^*)}\Phi(\cdot,z_p)$. We give the detailed analysis and refer to [19] for the steps which do not need important changes.

3.1. Proof of Proposition 3.2. We give the proof for $a\in\partial D_I$. The proof for $a\in\partial D_D$ is similar.

Let $\tilde{E}^s(x,z_p)$ be the solution of

$$\text{(3.10)}\qquad \begin{cases} (\Delta+\kappa^2)\tilde{E}^s(x,z_p)=0 \quad \text{in } R^2\setminus\overline{D},\\ (\frac{\partial}{\partial\nu}+i\kappa\sigma(x))\tilde{E}^s(x,z_p)=-(\partial_\nu+i\sigma(x))\frac{\partial}{\partial\nu(z_p^*)}\Phi(x,z_p) \quad \text{on } \partial D,\\ \tilde{E}^s(\cdot,z) \text{ satisfies the Sommerfeld radiation condition.}\end{cases}$$

Hence $(E^s-\tilde{E}^s)(x,z_p)$ satisfies

$$\text{(3.11)}\qquad \begin{cases} (\Delta+\kappa^2)(E^s-\tilde{E}^s)(x,z_p)=0 \quad \text{in } R^2\setminus\overline{D},\\ (\frac{\partial}{\partial\nu}+i\kappa\sigma(x))(E^s-\tilde{E}^s)(x,z_p)=0 \quad \text{on } \partial D_I,\\ (E^s-\tilde{E}^s)(\cdot,z_p)=-\frac{\partial}{\partial\nu(z_p^*)}\Phi(x,z_p)-\tilde{E}^s \quad \text{on } \partial D_D,\\ (E^s-\tilde{E}^s)(\cdot,z) \text{ satisfies the Sommerfeld radiation condition.}\end{cases}$$

We state $H_\sigma(x,z):=\tilde{E}(x,z)+\partial_{\nu(z_p^*)}\Phi(x,z)$. Hence H satisfies

$$\text{(3.12)}\qquad \begin{cases} (\Delta+\kappa^2)H_\sigma(x,z)=-\nabla\delta(x,z)\cdot\nu(z_p^*) \quad \text{in } R^2\setminus\overline{D},\\ (\frac{\partial}{\partial\nu}+i\kappa\sigma(x))H_\sigma(x,z)=0 \quad \text{on } \partial D,\\ H_\sigma(\cdot,z) \text{ satisfies the Sommerfeld radiation condition.}\end{cases}$$

We have the following estimates:

$$\text{(3.13)}\qquad \begin{cases} |G_\sigma(x,z)|\le c|\ln|x-z||\\ |\nabla G_\sigma(x,z)|\le c|x-z|^{-1}\\ |H_\sigma(x,z)|\le c|x-z|^{-1}\\ |\nabla H_\sigma(x,z)|\le c|x-z|^{-2}\end{cases} \quad \text{in } R^2\setminus D, \text{ where } c \text{ is a positive constant.}$$

The justification of these properties can be derived following, for instance, the approach of [24] and [25] since an explicit form of a local fundamental solution for the half-space case can be derived as we did in Proposition 3.1. See also [18] and [12] for the case of elliptic problems with rough coefficients.

From these estimates, we deduce that $\tilde{E}(\cdot,z_p)$ and its derivatives are bounded for $x\in\partial D_D$ and z_p near $a\in\partial D_I$. The well-posedness of (3.11) implies that $(E^s-\tilde{E}^s)(\cdot,z_p)$ is bounded in $H^1_{loc}(R^2\setminus\overline{D})$ for z_p near a. Introducing a cutoff function near the point a and using (3.11), we deduce that $(E^s-\tilde{E}^s)(\cdot,z_p)$ is bounded for x near ∂D_I and z_p near a.

This means that we can replace E^s by $\tilde{E}^s$ in Proposition 3.2.

We introduce $w^s_\sigma(\cdot, z_p)$ as the solution of

$$\begin{cases} (\Delta + \kappa^2) w^s_\sigma(x, z_p) = 0 \quad \text{in } R^2 \setminus \overline{D}, \\ (\frac{\partial}{\partial \nu} + i\kappa\sigma(x)) w^s_\sigma(x, z_p) = -(\partial_\nu + i\sigma(x)) \frac{\partial}{\partial \nu(a)} \Phi(\cdot, z_p) \quad \text{on } \partial D, \\ w^s_\sigma(\cdot, z) \text{ satisfies the Sommerfeld radiation condition,} \end{cases} \tag{3.14}$$

and denote by $w^s_{\sigma(a)}(\cdot, z_p)$ the solution of (3.14) replacing $\sigma(x)$ by $\sigma(a)$. Then we have the following result.

LEMMA 3.4. *There exist $\delta(a) > 0$ and $C(R) > 0$ such that*

$$|(\tilde{E}^s - w^s_\sigma)(x, z_p)| \le C(R) |\ln d(x, \partial D)| \cdot |\ln d(z_p, \partial D)|,$$

$$|\mathrm{Im}(\tilde{E}^s - w^s_\sigma)(x, z_p)| \le C(R), \quad |(w^s_\sigma - w^s_{\sigma(a)})(x, z_p)| \le C(R)$$

for $z_p \in B(a, \delta(a)) \cap C_{a,\theta}$ and $x \in (R^2 \setminus D) \cap B(0, R)$, for any $R > 0$ fixed.

Let $w^s_{\sigma(a),\Phi}(\cdot, z)$ be the solution of

$$\begin{cases} (\Delta + \kappa^2) w^s_{\sigma(a),\Phi}(x, z) = 0 \quad \text{in } \Omega \setminus \overline{D}, \\ (\frac{\partial}{\partial \nu} + i\kappa\sigma(a)) w^s_{\sigma(a),\Phi}(x, z) = -(\frac{\partial}{\partial \nu} + i\kappa\sigma(a)) \frac{\partial}{\partial \nu(a)} \Phi(x, z_p) \quad \text{on } \partial D, \\ w^s_{\sigma(a),\Phi}(\cdot, z) = -\frac{\partial}{\partial \nu(a)} \Phi(x, z_p) \quad \text{on } \partial\Omega \end{cases} \tag{3.15}$$

and $w^s_{\sigma(a),\Gamma}(\cdot, z)$ be the solution of (3.15) replacing Φ by Γ. Then we have the following claim.

LEMMA 3.5. *There exists $C > 0$ such that*

$$|(w^s_{\sigma(a)} - w^s_{\sigma(a),\Phi})(x, z)| \le C, \quad |(w^s_{\sigma(a),\Phi} - w^s_{\sigma(a),\Gamma})(x, z)| \le C$$

for $z \in \Omega \setminus D$ near D and $x \in \Omega \setminus D$.

We define $w^{s,0}_{\sigma(a)}$ to be the solution of (3.15) replacing Φ by Γ and the Helmholtz equation by the Laplace equation. Then we have the next claim.

LEMMA 3.6. *There exists $C > 0$ such that $|(w^s_{\sigma(a),\Gamma} - w^{s,0}_{\sigma(a)})(x, z)| \le C$, for $z \in \Omega \setminus D$ near D and $x \in \Omega \setminus D$.*

Finally, we have the next claim.

LEMMA 3.7. *There exist $C > 0$, $\delta(a) > 0$ such that*

1. $|(\mathrm{Im}\, w^{s,0}_{\sigma(a)} - \mathrm{Im}\, w^+_{\sigma(a)})(x, z)| \le C$ *for* $(x, z) \in B(a, \delta(a)) \cap C_{a,\theta}$;
2. $|(\mathrm{Re} w^{s,0}_{\sigma(a)} - \mathrm{Re} w^+_{\sigma(a)})(x, z)| \le C |\ln d(z_p, \partial D)|$ *for* $(x, z) \in B(a, \delta(a)) \cap C_{a,\theta}$.

By combining all the lemmas stated above, we end the proof of Proposition 3.2. □

In the proofs of these lemmas we do not, in general, specify the interdependency of the constants appearing in the estimates. However, we distinguish the constants that do or do not depend on the angle θ.

Proof of Lemma 3.4. The function $\tilde{E}^s(x, z_p) - w^s_\sigma(x, z_p)$ satisfies

$$\begin{cases} (\Delta + \kappa^2)(\tilde{E}^s - w^s_\sigma)(x, z_p) = 0 \quad \text{in } R^2 \setminus \overline{D}, \\ (\frac{\partial}{\partial \nu} + i\kappa\sigma)(\tilde{E}^s - w^s_\sigma)(x, z_p) = -(\frac{\partial}{\partial \nu} + i\kappa\sigma) \nabla\Phi \cdot [\nu(z^*_p) - \nu(a)] \quad \text{on } \partial D, \\ (\tilde{E}^s - w^s_\sigma)(\cdot, z_p) \text{ satisfies the Sommerfeld radiation condition.} \end{cases} \tag{3.16}$$

We need the following lemma.

LEMMA 3.8. *We have the estimate*

$$|\nu(z^*_p) - \nu(a)| \le C |z^*_p - a|$$

for z_p^* near a, where C is a positive constant.

Proof of Lemma 3.8. Take a point $b \in \partial\Omega$ and connect it to a by a C^3 smooth curve l such that $l(0) = a$, $l(1) = b$, $l(s) \in \Omega \setminus \overline{D}$ $(s \in (0,1))$. By Theorem 7.1 of [11], there is a C^2 strict deformation family $\{D_a^{l(s)}\}$ of a and $\partial\Omega$. That is, each $\partial D_a^{l(s)}$ is C^2 diffeomorphic to the unit circle and $\{D_a^{l(s)}\}$ satisfies the following properties:

(i) $a \in \partial D_a^{l(0)}$, $D \subset D_a^{l(0)} \subset \Omega$.

(ii) $\partial D_a^{l(1)} = \partial\Omega$, $\overline{D_a^{l(s)}} \subset D_a^{l(s')}$ $(0 \le s < s' \le 1)$.

(iii) l intersects $\partial D_a^{l(s)}$ at $l(s)$.

(iv) $\partial D_a^{l(s)}$ depends C^2 smoothly on $s \in [0,1]$.

For every $s \in [0,1]$ we define $\nu(l(s))$ to be the unit normal of $\partial D_a^{l(s)}$ at $l(s)$. From (iv), the map $s \in [0,1] \to \nu(l(s))$ is C^1. Choosing l to be a one-to-one curve near a, we deduce that the map $l(s) \in l([0,1]) \to \nu(l(s))$ is also C^1 near $l(0)$. Now let $\{s_p\} \subset [0,1]$ be such that $s_p \to 0$ $(p \to \infty)$ and let l be a C^3 smooth curve such that $z_p^* := l(s_p)$ and $\{z_p\}_{p\in\mathbb{N}} \subset l([0,1])$. Since for every p, z_p and z_p^* are in $\Omega \setminus \overline{D}$, then we can always choose l such that $l(s) \in \Omega \setminus \overline{D}$, $(s \in [0,1])$. We set $D_a^p := D_a^{l(s_p)}$. Hence the sequence $\nu(z_p^*)$ satisfies Lemma 3.8. □

The function $\mathrm{Im}(\tilde{E}^s - w_\sigma^s)$ satisfies

(3.17)
$$\begin{cases} (\Delta + \kappa^2)\mathrm{Im}(\tilde{E}^s - w_\sigma^s)(x,z) = 0 \quad \text{in } R^2 \setminus \overline{D}, \\ \frac{\partial}{\partial\nu}\mathrm{Im}(\tilde{E}^- w_\sigma^s)(x,z) = [\kappa\sigma\mathrm{Re}(\tilde{E}^s - w_\sigma^s) - [\frac{\partial}{\partial\nu}\mathrm{Im}\nabla\Phi + \kappa\sigma\mathrm{Re}\nabla\Phi] \cdot [\nu(z_p^*) - \nu(a)]](x,z) \\ \qquad \text{on } \partial D. \end{cases}$$

Hence we have

$$-\mathrm{Im}(\tilde{E}^s - w_\sigma^s)(x,z_p) = \int_{\partial D} F(y,z_p) G_N(y,x) ds(y) + \int_{\partial B_R} \mathrm{Im}(\tilde{E}^s - w_\sigma^s)(y,z_p) \frac{\partial}{\partial\nu} G_N(y,x) ds(y), \tag{3.18}$$

where

$$F(x,z_p) := \kappa\sigma(x)\mathrm{Re}(\tilde{E}^s - w_\sigma^s)(x,z) + \left[-\frac{\partial}{\partial\nu}\mathrm{Im}\nabla\Phi - \kappa\sigma(x)\mathrm{Re}\nabla\Phi\right](x,z) \cdot [\nu(z_p^*) - \nu(a)],$$

and $G_N(x,y)$ is the Green's function of the problem given by the Helmholtz equation in $B_R \setminus \overline{D}$ with Neumann boundary condition on ∂D and Dirichlet boundary condition on ∂B_R. The normal ν is oriented outside $B \setminus \overline{D}$.

From (3.16), we have

$$-(\tilde{E}^s - w_\sigma^s)(x,z_p) = \int_{\partial D} G_\sigma(x,y) \left(\frac{\partial}{\partial\nu} + i\kappa\sigma(y)\right) [\nabla\Phi \cdot (\nu(z_p^*) - \nu(a))](y,z_p) ds(y).$$

Hence

$$|(\tilde{E}^s - w_\sigma^s)(x,z_p)| \le c \int_{\partial D} (\ln|x-y|)|y-z_p|^{-2}|z_p^* - a| ds(y).$$

For $y \in \partial D$ and $z_p \in C_{a,\theta}$ we have the inequality $|z_p - a| \le C(\theta)|z_p - y|$. Applying this inequality to z_p^*, enlarging θ if necessary, we have

$$|(\tilde{E}^s - w_\sigma^s)(x,z_p)| \le c \int_{\partial D} (\ln|x-y|)|y-z_p|^{-1} ds(y) \tag{3.19}$$

since $|z_p^* - y| \le |z_p - y|$, for $y \in \partial D$ and z_p near a. Hence for every $\alpha > 0$ there exists $C_\alpha > 0$ such that

$$|(\tilde{E}^s - w_\sigma^s)(x, z_p)| \le C_\alpha |x - z_p|^{-\alpha}.$$

From the explicit form of Φ, we have $|\frac{\partial}{\partial \nu} \mathrm{Im} \nabla \Phi(x, z_p)| \cdot |\nu(z_p^*) - \nu(a)| \le c|x - z_p|^{-1} \cdot |z_p^* - a|$ and $\mathrm{Re} \nabla \Phi(x, z_p)| \cdot |\nu(z_p^*) - \nu(a)| \le c|x - z_p|^{-1} |z_p^* - a|$. Hence $F(y, z_p) \le C_\alpha |y - z_p|^{-\alpha}$ for $y \in \partial D$. Using the estimate $|G_N(x, y)| \le C_\alpha \ln |x - y|$ and choosing $\alpha < 1$, we deduce from (3.18) that

$$|\mathrm{Im}(\tilde{E}^s - w_\sigma^s)(x, z_p)| \le C_\alpha.$$

We have the first estimate of Lemma 3.4 from (3.19), i.e.,

$$|(\tilde{E}^s - w_\sigma^s)(x, z_p)| \le C \ln d(x, \partial D)| \cdot |\ln d(z_p, \partial D)|.$$

Now consider the third estimate of Lemma 3.4. We set $R(x, z) := w_\sigma^s(x, z) - w_{\sigma(a)}^s(x, z)$. Then it satisfies

(3.20)
$$\begin{cases} (\Delta + \kappa^2) R(x, z) = 0 \quad \text{in } R^2 \setminus \overline{D}, \\ \frac{\partial R(x,z)}{\partial \nu} + i\kappa\sigma(a) R(x, z) = -i\kappa(\sigma(x) - \sigma(a))(w_\lambda^s(x, z) + \frac{\partial}{\partial \nu(a)} \Phi(x, z)) \quad \text{on } \partial D, \\ R(\cdot, z) \text{ satisfies the Sommerfeld radiation condition.} \end{cases}$$

From (3.20), we have the representation

$$R(x, z) = -\int_{\partial D} i\kappa(\sigma(y) - \sigma(a)) G_{\sigma(a)}(y, x) \left(w_\sigma^s + \frac{\partial}{\partial \nu(a)} \Phi \right)(y, z) ds(y) \quad \text{for } (x, z) \in R^2 \setminus \overline{D}. \tag{3.21}$$

We define $K(x, z) := -(\frac{\partial}{\partial \nu} + i\kappa\sigma(x)) \frac{\partial}{\partial \nu(a)} \Phi(x, z)$. From (3.14) we have the representation

$$w_\sigma^s(x, z) = \int_{\partial D} G_{\sigma(a)}(y, x) K(y, z) ds(y);$$

hence, due to the estimates of the Green's function $G_{\sigma(a)}(x, y)$ and $\Phi(x, y)$, we have

$$|w_\sigma^s(x, z)| \le \int_{\partial D} |\ln(|y - x|)||z - y|^{-2} ds(y) \le \frac{c}{|x - z|}.$$

From (3.21) and the Holder regularity of $\sigma(x)$, we deduce that

$$|R(x, z)| \le c \int_{\partial D} |y - a|^\beta \ln(|y - x|)||z - y|^{-1} ds(y).$$

From the inequality $|y - a| \le c(\theta)|y - z|$ for $y \in \partial D$ and $z \in C_{a,\theta} \cap B(a, \delta(a))$ we have

$$\frac{|y - a|^\beta}{|y - z|} \le \frac{c(\theta)^\beta C}{|y - z|^{1-\beta}},$$

which implies

$$|R(x, z)| \le \int_{\partial D} \frac{c(\theta)^\beta C |\ln |y - x||}{|y - z|^{1-\beta}} dy$$

and therefore $|R(x,z)| = O(1)$ for $x \in R^2 \setminus D$ and $z \in C_{a,\theta} \cap B(0,R)$. ∎

Proof of Lemmas 3.5 *and* 3.6. Similarly to the proof of Lemma 3.4, these proofs are based on the use of integral representations and the pointwise estimates of the Green's functions of the corresponding problems. We omit the details. ∎

Proof of Lemma 3.7. Since $w^{s,0}_{\sigma(a)}$ satisfies

$$\begin{cases} \Delta w^{s,0}_{\sigma(a)}(x,z) = 0 \quad \text{in } \Omega \setminus \overline{D}, \\ (\frac{\partial}{\partial \nu} + i\kappa\sigma(a))(w^{s,0}_{\sigma(a)}(\cdot,z)) = -(\frac{\partial}{\partial \nu} + i\kappa\sigma(a))\frac{\partial}{\partial \nu(a)}\Gamma \quad \text{on } \partial D, \\ w^{s,0}_{\sigma(a)}(\cdot,z) = -\frac{\partial}{\partial \nu(a)}(\Gamma) \quad \text{on } \partial\Omega, \end{cases} \tag{3.22}$$

then it is clear that $G^0_{\sigma(a)} := w^{s,0}_{\sigma(a)}(x,y) + \partial_{\nu(a)}\Gamma(x,y)$ satisfies

$$\begin{cases} \Delta(G^0_{\sigma(a)})(x,z) = -\frac{\partial}{\partial \nu(a)}\delta(x-y) \quad \text{in } \Omega \setminus \overline{D}, \\ (\frac{\partial}{\partial \nu} + i\kappa\sigma(a))(G^0_{\sigma(a)})(\cdot,z) = 0 \quad \text{on } \partial D, \\ (G^0_{\sigma(a)})(\cdot,z) = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{3.23}$$

We can assume without loss of generality that $a = (0,0)$ and $\nu(a) = (0,1)$ by using the rigid transformation of coordinates $[R_a(\nu(a)) + M_a]$. Let $\xi = F(x)$ be the local change of variables

$$\xi_1 = x_1, \quad \xi_2 = x_2 - f(x_1), \tag{3.24}$$

where f is the function defined in the introduction. We have the following properties:

$$\begin{cases} c_1|x-z| \le |F(x) - F(z)| \le c_2|x-z|, \\ |F(x) - x| \le c_3|x|^2, \\ |DF(x) - I| \le c_4|x| \end{cases} \tag{3.25}$$

for x, z near the point a, where c_i $(i = 1,2,3,4)$ are positive constants, which is due to hypothesis on the regularity of ∂D.

Let x, z be points near a. From (3.23), we deduce that $\tilde{G}^0_{\sigma(a)}(\xi,\eta) = G^0_{\sigma(a)}(x,z)$ satisfies

$$\begin{cases} \nabla_\xi \cdot B(\xi)\nabla_\xi \tilde{G}^0_{\sigma(a)} = -J^T(\xi)\nabla_\xi\delta(\xi-\eta)\cdot(0,1) \quad \text{near } F(a), \\ |J^{-T}\nu|B(\xi)\nabla_\xi\tilde{G}^0_{\sigma(a)}\cdot\tilde{\nu} + i\kappa\sigma(a)\tilde{G}^0_{\sigma(a)} = 0 \quad \text{on } \partial R^2_+ \text{ near } F(a), \end{cases} \tag{3.26}$$

where $\xi := F(x)$ and $\eta := F(z)$, $B := JJ^T$ and $J := \frac{\partial \xi}{\partial x}(F^{-1}(\xi))$, and $\tilde{\nu} := (0,1)$ is the unit normal to ∂R^2_+. We denoted by J^{-T} the adjoint of J^{-1}. We have from (3.25) that

$$|J^T(\xi) - J^T(0)| \le c|\xi|, \quad |B(\xi) - B(0)| \le c|\xi|$$

and $J(0) = B(0) = I$.

We set $\Gamma_{\sigma(a)}(x,z) := (w^+_{\sigma(a)} + \frac{\partial}{\partial x_2}\Gamma)(x,z)$ and write $\tilde{R}(\xi,\eta) := \tilde{G}^0_{\sigma(a)}(\xi,\eta) - \Gamma_{\sigma(a)}(\xi,\eta)$. Then the function $\tilde{R}(\cdot,\eta)$ satisfies

$$\begin{cases} \nabla_\xi \cdot B(\xi)\nabla_\xi\tilde{R} = \nabla_\xi\cdot(I-B)\nabla_\xi\Gamma_{\sigma(a)} + (I - J^T(\xi))\nabla_\xi\delta(\xi-\eta)\cdot(0,1), \\ B(\xi)\nabla_\xi\tilde{R}\cdot\tilde{\nu} + i\kappa\sigma(a)\tilde{R} = (I-B)\nabla_\xi\Gamma_{\sigma(a)}\cdot\tilde{\nu} + i\kappa\sigma(a)(1 - |J^{-T}\nu|^{-1})\tilde{R}, \end{cases} \tag{3.27}$$

where the first relation holds in R^2_+ near $F(a)$, while the second one is satisfied on ∂R^2_+ near $F(a)$.

We remark that $(I - J^T(\xi))$ is equal to the matrix given by the two line vectors $(0, -f(\xi))$ and $(0, 0)$. Hence we have

$$(I - J^T(\xi))\nabla_\xi \mathrm{Im}\Gamma_{\sigma(a)}(\xi, \eta) \cdot (0, 1) = 0.$$

With this remark, the problem (3.27) is exactly the one studied in [19]. We write $\partial B_r^+ = S_r \cup S_r^c$ with $S_r := \partial B_r^+ \cap \partial F(D)$. Arguing as in [19], we obtain the estimate

$$|\mathrm{Im}\ \tilde{R}(\xi, \eta)| < c \quad \text{for } \xi \in S_r \text{ and } \eta \in C_{F(a),\theta}$$

and

$$|\mathrm{Re}\ \tilde{R}(\xi, \eta)| < c|\ln|\xi - \eta|| \quad \text{for } \xi \in S_r \text{ and } \eta \in C_{F(a),\theta}.$$

We go back to $R(x, z) := G^0_{\sigma(a)}(x, z) - \Gamma_{\sigma(a)}(x, z)$. We have

$$R(x, z) = G^0_{\sigma(a)}(x, z) - \Gamma_{\sigma(a)}(F(x), F(z)) + \Gamma_{\sigma(a)}(F(x), F(z)) - \Gamma_{\sigma(a)}(x, z),$$

which can be rewritten as

$$\begin{aligned} R(x, z) = \tilde{R}(F(x), F(z)) &+ [\Gamma_{\sigma(a)}(F(x), F(z)) - \Gamma_{\sigma(a)}(F(x), z)] \\ &+ [\Gamma_{\sigma(a)}(F(x), z) - \Gamma_{\sigma(a)}(x, z)]. \end{aligned} \tag{3.28}$$

By the same argument as in [19], we end up with the estimate

$$|\mathrm{Im}\ R(x, z)| \le c(\theta) \tag{3.29}$$

for $x \in B(a, \delta(a))$ such that $F(x) \in S_r$ and $z \in C_{a,\theta} \cap B(a, \delta(a))$.

For $z \in C_{a,\theta} \cap B(a, \frac{\delta(a)}{2})$ and $x \in [\partial B(a, \delta(a))] \cap R^2 \setminus \overline{D}$, we have

$$|\mathrm{Im}\ R(x, z)| \le c \tag{3.30}$$

with some positive constant c, because $C_{a,\theta} \cap B(a, \frac{\delta(a)}{2})$ and $[\partial B(a, \delta(a))] \cap R^2 \setminus \overline{D}$ are separated sets. Since in $B(a, \delta(a)) \cap (R^2 \setminus \overline{D})$, we have $\Delta_x \mathrm{Im}\ R(x, z) = 0$; then, using (3.29) and (3.30), we have $|\mathrm{Im}\ R(x, z)| \le c(\theta)$ for $x \in [R^2 \setminus D] \cap B(a, \delta(a))$ and $z \in C_{a,\theta} \cap B(a, \frac{\delta(a)}{2})$, by the maximum principle.

Similarly we have $|\mathrm{Re}R(x, z)| \le C|\ln|x - z||$ for $x \in B(a, \delta(a))$ such that $F(x) \in S_r$ and $z \in C_{a,\theta} \cap B(a, \delta(a))$. Hence $|\mathrm{Re}R(x, z)| \le C|\ln d(z, \partial D)|$ for $x \in B(a, \delta(a))$ such that $F(x) \in S_r$ and $z \in C_{a,\theta} \cap B(a, \delta(a))$, which is the counterpart of (3.30) for $\mathrm{Re}\tilde{R}$. The rest of the proof is the same as that for the imaginary part. ∎

4. Numerical tests. In this section, we consider two reconstruction model problems for numerical tests based on Theorem 2.1. In the first model we take the obstacle to be a disc, and on its whole boundary we impose the impedance boundary condition. The purpose of considering such a model is to show the influence of wave number κ and singularity strength on the inversion scheme. In the second model, we consider a nonconvex obstacle with mixed boundary conditions. We check our inversion formulas fully, especially for the identification of a different type of boundary. Also we show the effect of the nonconvex part on the inversion performance.

In the reconstruction scheme, the approximation of $\Phi(x, z_p)$ and $\Phi_{\nu(a)}(x, z_p)$ by the Herglotz wave function plays a key role. To do this, we need to construct the approximate domain D_a^p in the way mentioned in section 2. Then its density functions

can be determined by the standard argument of a minimum norm solution of the integral equation of the first kind. In this way the numerators in (2.18), (2.20), and (2.21) can be computed for every sequence of points $(z_p)_{p\in\mathbb{N}}$ approaching ∂D in terms of the far-field pattern.

In testing our inversion scheme, we simulate the inversion input data (far-field pattern) by solving the direct problem using the combined single- and double-layer potential method; see [8] and [9].

In subsections 4.1 and 4.2, we consider the first model, while subsection 4.3 is dedicated to a nonconvex obstacle with mixed boundary conditions, from which we can test our theoretical results.

4.1. Reconstruction of ∂D. For the numerical test we take $\partial D = \{x = (x_1(t), x_2(t)) = 1.2(\cos t, \sin t) : t \in [0, 2\pi]\}$. We can construct D_a^p in a special way by

$$\partial D_a^p = \{(\tilde{x}_1(t), \tilde{x}_2(t)) = (1.2 + \delta_0(p))(\cos t, \sin t) : t \in [0, 2\pi]\}$$

and take $z_p = (\tilde{x}_1(t_0), \tilde{x}_2(t_0)) + \delta_1(p)(\cos t_0, \sin t_0)$ outside of $\overline{D}_a^p$ for $a = 1.2(\cos t_0, \sin t_0)$ in ∂D. The smallness of $\delta_1(p)$ determines the singularity of $\Phi(x, z_p), \Phi_{\nu(a)}(x, z_p)$ on ∂D_a^p near z_p.

Example 1. Construction of ∂D with unknown impedance $\sigma(x)$ for given far-field data.

It follows from the first point of Theorem 2.1 that the boundary ∂D can be reconstructed from the blowup behavior of the approximate values of the indicator function:

$$I_{m,n}(z_p) :\approx \left(\frac{\pi}{N}\right)^2 \left| \sum_{i,j=1}^{2N-1} \mathrm{Re}\left(\gamma_2^{-1} U^\infty(i,j) F_m^p(j) G_m^p(i)\right) \right| \tag{4.1}$$

for large m, n, where $G_n^p(i) := g_n^p(\hat{x}_i)$, $F_m^p(j) := f_m^p(\hat{d}_j)$, and $U^\infty(i,j)$ represents $u^\infty(-\hat{x}_i, d_j)$.

Therefore, we can test the numerical performance of this formula by taking z_p approaching ∂D. If $I_{m,n}(z_p)$ is greater than some a priori given large value, for fixed m, n large enough, we consider z_p to be almost in ∂D. Notice that only the singularity in $\Phi_{\nu(\tilde{z}_p)}(x, z_p)$ is needed in this procedure; we do not need the boundary ∂D.

In our numerical implementations, we take $\kappa = 0.6$, $N = 32$ and keep the singularity of $\Phi(\cdot, z_p)$ unchanged near ∂D_a^p by fixing $\delta_1 = 0.015$. When z_p approaches ∂D by decreasing δ_0 from different directions t_0, we get different values of $I_{m,n}(z_p)$. We choose the same constant CB as the criterion for the blowing-up of the indicator $I_{m,n}(z_p(t_0))$ at different direction t_0. That is, we choose $z_p(t_0)$ as the approximation of the point $a \in \partial D$ in the direction t_0 when $I_{m,n}(z_p(t_0))$ is larger than CB. Then the constructed approximate position of ∂D is given by interpolating those points. In this way, we can draw the approximate shape of ∂D by choosing all directions t_0 using the same given CB. As obtained in the theoretical result, the larger the value CB is, the better the approximation of ∂D should be.

Consider the boundary reconstruction problem with nonconstant impedance

$$\sigma(x) = \frac{2 + x_1 x_2}{(3 + x_2)^2}, \quad x \in \partial D.$$

Under this configuration, the reconstructions for different blowup criteria are shown in Figure 4.1(left), while the distribution of indicator value is given in the right panel.

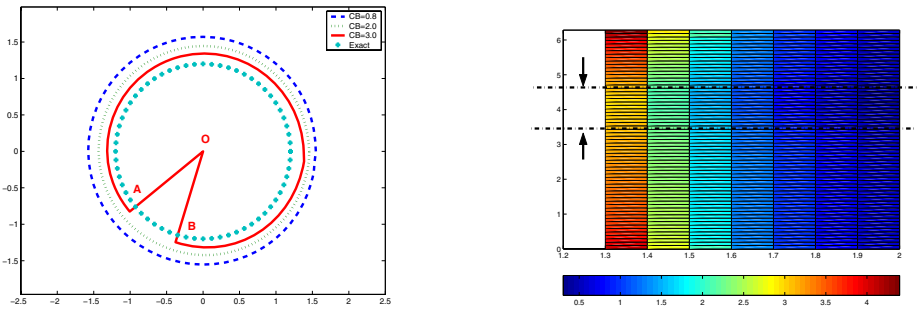


FIG. 4.1. *Reconstruction of ∂D for variable impedance with $\kappa = 0.6$ with different blowup constants CB; the values of the indicator are small in the directions corresponding to large impedance.*

There is an interesting phenomenon in the numerics. For a reasonable indicator value $CB = 2$, we can see the whole rough shape of ∂D. However, for larger values, i.e., $CB = 3$, most of the part of ∂D can be seen with more satisfactory accuracy, but some part, i.e., $(\widehat{AB})$, is not visible. In this part the indicator value is less than 3 (but, of course, bigger than 2). This numerical performance is closely related to the impedance distribution in ∂D. It can be seen from the right-hand part of Figure 4.1 that the indicator value is obviously smaller in the narrow domain at each radius layer. Therefore as r decreases, the part of ∂D related to these angles cannot be detected with the same accuracy. Considering the distribution of boundary impedance, this part corresponds to $\sigma(x)$ with large value, so it cannot be seen as clearly as the other part by using the same criterion values CB. This may be explained by the fact that the scattered wave along these directions will be much absorbed. Another, but related, reason is the property (2.19), i.e.,

$$\lim_{m,n\to\infty} \mathrm{Re}\left[\gamma_2^{-1}\int_{S^1}\int_{S^1} u^\infty(-\hat{x},d) f_m^p(d) g_n^p(\hat{x})\, ds(\hat{x}) ds(d)\right] = \frac{\pm 1}{4\pi|(z_p-a)\cdot\nu(a)|} + O(|\ln|z_p-a||),$$

where $O(|\ln|z_p-a||)$ may be large if σ is large near the point a. Hence numerically the second term can weaken the blowup of the first term.

Physically, this is the reason why we introduce the coated part of an obstacle, i.e., to avoid or perturb the detection of the obstacle.

The other special property of this example is that the whole boundary shape can be reconstructed well using only one blowup criterion CB. This comes from the special geometric shape and the fact that we have a complete impedance boundary condition. However, in the case of general problems, as for a nonconvex obstacle, with a mixed boundary condition we need to use multiple blowup values to get a satisfactory reconstruction; see subsection 4.3.

4.2. Reconstruction of $\sigma(x)$ for known ∂D. For a given $a \in \partial D$, we take $z_p \in R^2 \setminus \overline{D}_a^p$ with $D \subset\subset D_a^p$.

By the theoretical result given in (2.21), the approximate formula for $\sigma(a)$ is

$$\frac{\kappa}{\pi}\sigma(a) \approx \frac{1}{|\ln((z_p-a)\cdot\nu(a))|}\left(\frac{\pi}{n}\right)^2 \sum_{i,j=1}^{2n-1} \mathrm{Im}\,\left(\gamma_2^{-1} U^\infty(i,j) F_m^p(j) G_n^p(i)\right) \tag{4.2}$$

for large m, n and small $|z_p - a|$.

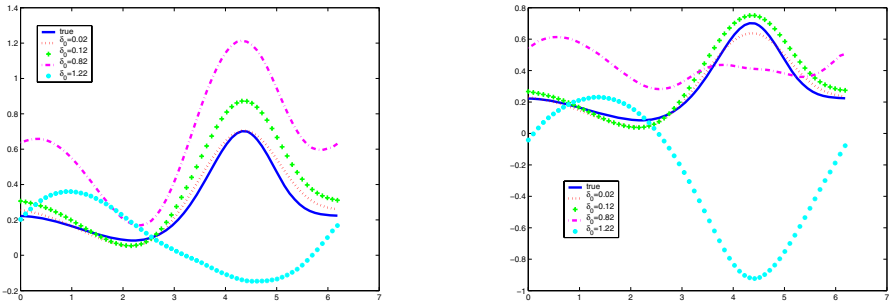


FIG. 4.2. *Reconstruction of* $\sigma(x) = \frac{2+x_1x_2}{(3+x_2)^2}$ *for* $\delta_0 = 1.22, 0.82, 0.12, 0.02$ *with two different wave numbers* $\kappa = 0.6$ *(left) and* $\kappa = 0.7$ *(right), where we fix* $\delta_1 = 0.015$.

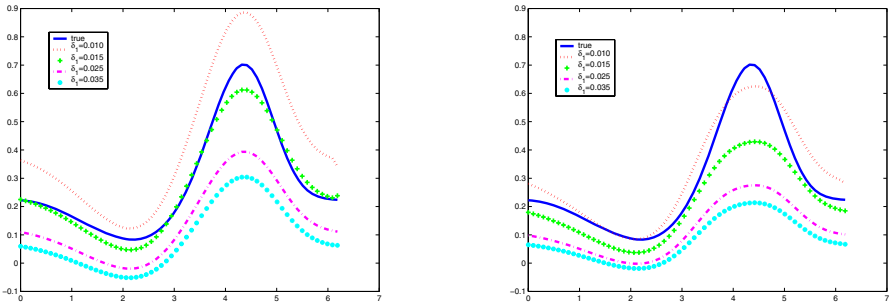


FIG. 4.3. *Reconstruction of* $\sigma(x) = \frac{2+x_1x_2}{(3+x_2)^2}$ *for* $\delta_1 = 0.010, 0.015, 0.025, 0.035$ *with two different wave numbers* $\kappa = 0.6$ *(left) and* $\kappa = 1.0$ *(right), where we fix* $\delta_0 = 0.002$.

Example 2. Consider the variable impedance distribution given in Example 1.

We take $n = 32$. First, let us keep the singularity unchanged by fixing δ_1 and shrink the radius of ∂D_a^p such that $z_p \to a \in \partial D$. The computed values of $\sigma(x(t))$ with fixed $\delta_1 = 0.015$ and different $\delta_0 = 1.22, 0.82, 0.12, 0.02$ for two wave numbers $\kappa = 0.6, \kappa = 0.7$ are shown in Figure 4.2. It can be seen from this figure that when $\delta_0 \to 0$ ($\delta_0 = 0.02$), the reconstruction results are satisfactory for both wave numbers.

It is interesting to see that when δ_0 is large enough ($\delta_0 = 1.22$), the reconstruction is invalid, even if here we use a strong singularity $\delta_1 = 0.015$. This is reasonable since the approximation to the strong singularity of the fundamental solution contains much error.

Finally, we fix ∂D_a^p near ∂D and take z_p tending to ∂D_a^p. In our configuration, this means $\delta_1 \to 0$ for fixed small δ_0. The tests with $\delta_0 = 0.002$ for two wave numbers $\kappa = 0.6, 1.0$ at different $\delta_1 = 0.025, 0.020, 0.015, 0.010$ are given in Figure 4.3, which shows the influence of the singularity. It can be seen that the approximation is sensitive to the wave number κ. The reason is that we ignore the remaining term $C/|\ln|(z_p - a)\cdot\nu(a)||$, where the constant C comes from (3.5). Theoretically, this term tends to 0 as $z_p \to a$. However, this procedure causes a difficulty in approximating the fundamental solution. We expect that the constant C becomes large as κ becomes small. This is naturally related to the following relation, in the 2-dimensional case, between the fundamental solution to the Helmholtz and Laplace equations:

$$H_0^{(1)}(\kappa|x-y|) = -\frac{1}{2\pi}\ln(|x-y|) - \ln\kappa + O(1)$$

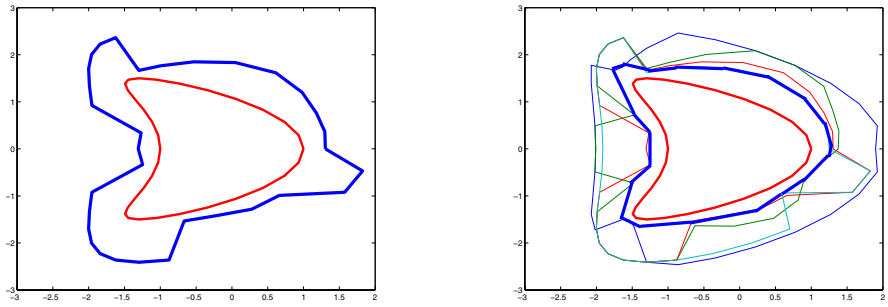


FIG. 4.4. *Reconstruction of ∂D with $CB = 3.0$ (left) and four CBs (right) by concave-hull.*

locally for $x, y \in R^2$ and κ small enough; see [8]. This implies that the difference between the fundamental solution behaves as $\ln \kappa$. So the constant C should have a similar behavior with respect to κ. These remarks on the dependency on wave number κ are also observed in the tests for detecting ∂D. However, we think that this is just a 2-dimensional phenomenon.

4.3. Reconstruction of an obstacle with mixed-type boundary. Since the main advantage of our inversion method is its ability to identify the full complex obstacle simultaneously, here we consider the numerical behavior of our inversion method acting on a nonconvex obstacle with mixed boundary condition.

Example 3. Consider a nonconvex obstacle D with the boundary

$$\partial D = \{x : x(t) = (x_1(t), x_2(t)) = (\cos t + 0.65 \cos 2t - 0.65, 1.5 \sin t), t \in [0, 2\pi]\}.$$

Let ∂D be composed of sound-soft part ∂D_D for $t \in [0, 1.42\pi]$, and impedance part ∂D_I for $t \in [1.42\pi, 2\pi]$. In ∂D_I, we assume the impedance coefficient $\sigma(x(t)) \equiv 3$.

We also choose ∂D_a^p and z_p in a special way. For two small constants $\delta_0, \delta_1 > 0$, we take

$$\begin{cases} \partial D_a^p = \{y(t) := x(t) + \delta_0 \times (\cos t, \sin t) : x(t) \in \partial D, t \in [0, 2\pi]\}, \\ z_p(t) = y(t) + \delta_1 \times (\cos t, \sin t) \quad \text{for } t \in [0, 2\pi]. \end{cases} \tag{4.3}$$

In terms of Theorem 2.1, the inversion schemes contain the following three steps:

Step 1. Identify the location of ∂D using (2.18);

Step 2. Distinguish the different parts of ∂D in terms of (2.20);

Step 3. Reconstruct $\sigma(x)$ in ∂D_I from (2.21).

We present the numerical results with fixed wave number $\kappa = 0.9$ and $\delta_1 = 0.01$.

Step 1. we take $n = 16$ and decrease $\delta_0 = l \times 0.05$ by taking l from 20 to 2. The indicator values in (2.18) for ∂D are computed for $z_p(t)$ and ∂D_a^p specified here for every direction t_j and different δ_0. Then we draw the contour line to obtain an approximation of ∂D. As for Examples 1 and 2, we choose some appropriate value CB for the stopping rule of l. In the case when the indicator is always less than CB in some direction, we take z_p for the initial guess (the largest l) as an approximate location of points in ∂D.

In this case, the situation is different from the examples given in the previous subsections; the reconstruction of the boundary using only one blowup value CB is not sufficient. See the left-hand picture in Figure 4.4 for the reconstruction of $CB = 3.0$, where the kite-shape in red color is the exact obstacle. Enlarging CB

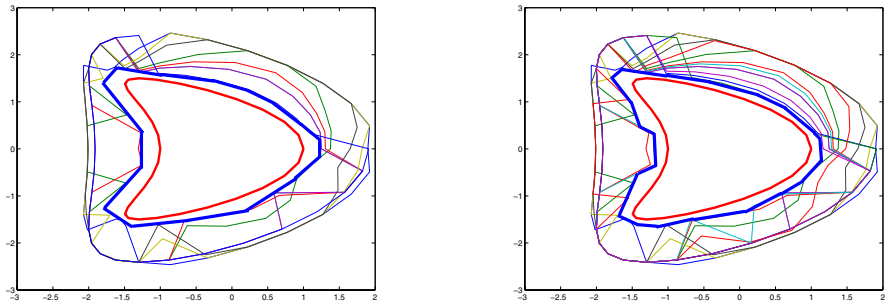


FIG. 4.5. *Reconstruction of ∂D by concave-hull using eight CBs (left) and twelve CBs (right).*

can improve the reconstruction along some directions, but the approximation to the whole boundary is still not satisfactory. The reason for this phenomenon is that our theoretical results do not guarantee the uniform blowup property for all directions.

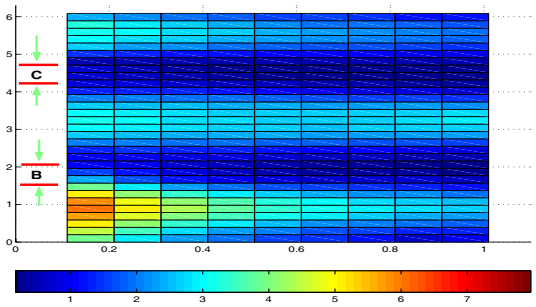


FIG. 4.6. *The indicator values for boundary shape.*

To overcome this difficulty and apply our reconstruction formula, we propose to combine the reconstruction results for different CB together and take the concave hull. Then ∂D can be approximated very well; see Figure 4.4, where we compare the shape given by using one CB only and the one obtained by using four CB values. In Figure 4.5, we show the reconstruction results by using eight and twelve CB values. Precisely, the reconstructions in Figures 4.4 and 4.5 correspond to the following CB values:

Figure 4.4 (right): four values–$CB = 0.8, 2.5, 3.0, 3.5$;

Figure 4.5 (left): eight values–$CB = 0.8, 2.5, 3.0, 3.5, 1.2, 1.5, 5.5, 6.5$;

Figure 4.5 (right): twelve values–$CB = 0.8, 2.5, 3.0, 3.5, 1.2, 1.5, 5.5, , 6.5, 2.0, 3.2, 4.5, 6.0$.

It can be seen that the reconstruction is satisfactory. This means that, using the technique of combining several CB's, the theoretical formulas provide good reconstructions.

We give the indicator value distribution in Figure 4.6 for all directions t with different l at each direction. We can see how the indicator near $t = 0.58\pi, \pi, 1.42\pi$ has some special property, which explains the difficulty of reconstructing these parts shown in Figures 4.4 and 4.5.

Next we consider Steps 2 and 3 by using (2.20) and (2.21), that is, we test the numerical behavior in distinguishing the boundary type and impedance in ∂D_I. Different from the formula (2.18), these two formulas need the boundary shape ∂D,

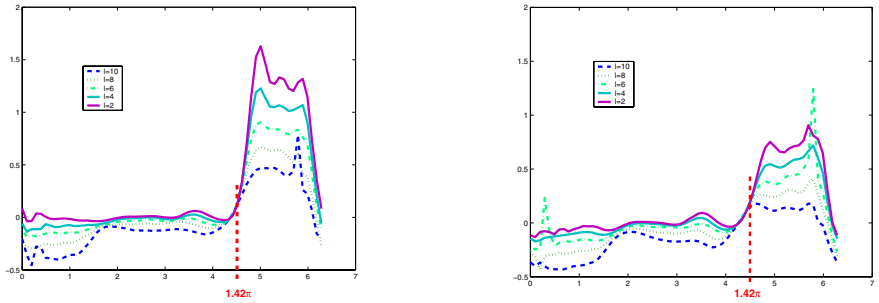


FIG. 4.7. *Indicator values for different parts of the boundary using exact ∂D (left) and $\partial\tilde{D}$ with $\delta^* = 0.1$ (right).*

which is theoretically obtained from (2.18). Since we can get some approximation to ∂D only numerically in terms of (2.18), it is necessary to check the approximate versions of (2.20) and (2.21) for distinguishing the boundary type and recovering $\sigma(x)$ on the coating part.

Step 2. We express the quantitative behavior for distinguishing ∂D_D and ∂D_I by giving the indicator distribution. As explained in Step 1, by using different blowup criteria in the shape reconstruction, we can get a good approximation to ∂D. In this step, we specify $\partial\tilde{D} \approx \partial D$ with an explicit expression given by

$$\partial\tilde{D} = \begin{pmatrix} \cos\delta^* & -\sin\delta^* \\ \sin\delta^* & \cos\delta^* \end{pmatrix} [\partial D + \{\delta^* \times (\cos t, \sin t) : t \in [0, 2\pi]\}] \tag{4.4}$$

with small constant $\delta^* > 0$, for simplicity. In this way, $\tilde{D}$ is no longer symmetric with respect to x_1, and the location for the corner part also differs from that of ∂D. To check the effect of this domain approximation on (2.20) and (2.21), we also evaluate them using the exact ∂D.

We generate $(\partial\tilde{D}_a^p, \partial D_a^p)$ from $(\partial\tilde{D}, \partial D)$ and therefore the sequences $(\{\tilde{z}_p\}, \{z_p\})$ as in (4.3). In this way, we have $\tilde{z}_p \to \tilde{a} \in \partial\tilde{D}$, $z_p \to a \in \partial D$, and $\delta_0, \delta_1 \to 0$. In the computation, we take $n = 32$ and decrease $\delta_0 = l \times 0.05 \to 0$ by taking l from 10 to 2. The indicator behavior using the same far-field pattern for ∂D and $\partial\tilde{D}$ with $\delta^* = 0.05$ is given in Figure 4.7.

Noticing the fact that the sound-soft part corresponds to the parameter $t \in [0, 1.42\pi]$, while the impedance part is related to $t \in [1.42\pi, 2\pi]$, the above numerical behavior supports (2.21) strongly with a large difference in $[0, 1.42\pi]$ and $[1.42\pi, 2\pi]$; that is, we can distinguish the boundary type in terms of the obvious difference of indicator values when z_p approaches the boundary (for small l), even if the boundary shape is known with a relative error.

Step 3. We compute the impedance coefficient on ∂D_I by applying the formulas for ∂D and also for its approximation $\partial\tilde{D}$. The reconstruction behavior for exact ∂D as well as its approximation $\partial\tilde{D}$ with $\delta^* = 0.05$ is shown in Figure 4.8. It can be seen that, for a given exact boundary shape ∂D, the theoretical result (2.21) for the impedance is valid (left figure), except near the end points of ∂D_I. The rough approximation in this part is reasonable, since this part is near to the sound-soft boundary. Using the approximate domain $\partial\tilde{D}$, the impedance can still be captured (right figure). Of course, the reconstruction is less accurate. Due to the nonconvex property of the obstacle and the mixed boundary condition, we think these results are satisfactory.

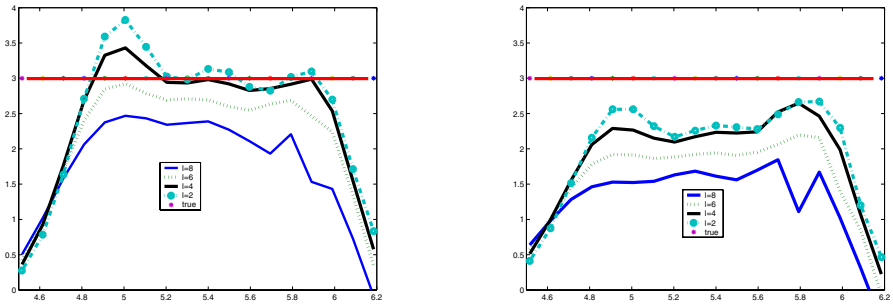


FIG. 4.8. *Recovery of σ using exact ∂D (left) and approximate $\partial\tilde{D}$ with $\delta^* = 0.05$ (right).*

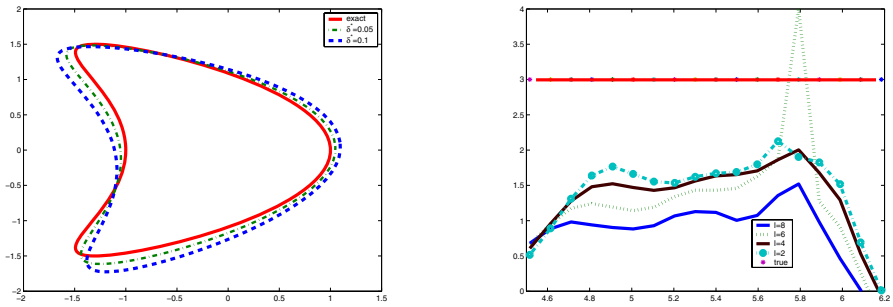


FIG. 4.9. *Approximate $\partial\tilde{D}$ with $\delta^* = 0.05, 0.1$ (left) and reconstruction of σ for different l from approximate $\partial\tilde{D}$ with $\delta^* = 0.1$ (right).*

However, if the perturbation for ∂D is very large, then the reconstruction of $\sigma(x(t))$ is much contaminated. This is due to the sensitivity of the approximate reconstruction to the boundary shape, especially for the corner part of the nonconvex domain, noticing that in the formula (2.21), the normal direction $\nu(a)$ appears. The perturbation scheme (4.4) moves the position of the corner part by rotation. An inversion result for the impedance with $\delta^* = 0.1$ is shown in Figure 4.9. From the left picture of this figure, we see how the corner part of $\partial\tilde{D}$ with $\delta^* = 0.1$ has been much moved from that of ∂D with a relative error almost 10%.

We conclude that the theoretical results (2.20) and (2.21) are well supported by our numerical tests even for nonconvex domains. If the approximate domain $\partial\tilde{D}$ is used in these two formulas, then we can still distinguish the boundary type in terms of the obvious blowup property of the indicator. However, the quantitative identification of impedance depends on the error level of ∂D.

Acknowledgments. The authors thank the anonymous referees for their valuable comments and suggestions, which led to a modified version of this paper.

REFERENCES

[1] I. AKDUMAN AND R. KRESS, *Direct and inverse scattering problems for inhomogeneous impedance cylinders of arbitrary shape*, Radio Sci., 38 (2003), pp. 1055–1064.

[2] F. CAKONI AND D. COLTON, *The determination of the surface impedance of a partially coated obstacle from far field data*, SIAM. J. Appl. Math., 64 (2004), pp. 709–723.

[3] F. CAKONI, D. COLTON, AND P. MONK, *The determination of the surface conductivity of a partially coated dielectric*, SIAM J. Appl. Math., 65 (2005), pp. 767–789.

[4] F. Cakoni and D. Colton, *Qualitative Methods in Inverse Scattering Theory*, Interaction of Mechanics and Mathematics, Springer, New York, 2006.
[5] J. Cheng, J. J. Liu, G. Nakamura, and S. Z. Wang, *Recovery of boundaries and types for multiple obstacles from the far-field pattern*, Quart. Appl. Math., to appear.
[6] J. Cheng, J. J. Liu, and G. Nakamura, *Recovery of the shape of an obstacle and the boundary impedance from the far-field pattern*, J. Math. Kyoto U., 43 (2003), pp. 165–186.
[7] D. Colton and A. Kirsch, *A simple method for solving inverse scattering problems in the resonance region*, Inverse Problems, 12 (1996), pp. 383–393.
[8] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer, Berlin, 1998.
[9] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.
[10] D. J. Hoppe and Y. Rahmat-Samii, *Impedance Boundary Condition in Electromagnetism*, Taylor and Francis, Oxford, UK, 1995.
[11] N. Honda, G. Nakamura, R. Potthast, and M. Sini, *The no-response approach and its relation to non-iterative methods for the inverse scattering*, Ann. Math. Pur. Appl., appeared online (DOI 10.1007/s10231-006-0030-1).
[12] M. Gruter and K. O. Widman, *The Green function for uniformly elliptic equations*, Manuscripta Math., 37 (1982), pp. 303–342.
[13] M. Ikehata, *Reconstruction of the shape of the inclusion by boundary measurements*, Comm. Partial Differential Equations, 23 (1998), pp. 1459–1474.
[14] M. Ikehata, *Reconstruction of obstacles from boundary measurements*, Wave Motion, 3 (1999), pp. 205–223.
[15] M. Ikehata, *A new formulation of the probe method and related problems*, Inverse Problems, 21 (2005), pp. 413–426.
[16] A. Kirsch, *Characterization of the shape of a scattering obstacle using the spectral data of the far field operator*, Inverse Problems, 14 (1998), pp. 1489–1512.
[17] R. Kress and W. Rundell, *Inverse scattering for shape and impedance*, Inverse Problems, 17 (2001), pp. 1075–1085.
[18] W. Littman, G. Stanpacchia, and H. F. Weinberger, *Regular points for elliptic equations with discontinuous coefficients*, Ann. Scuola Norm. Sup. Pisa (III), 17 (1963), pp. 43–77.
[19] G. Nakamura and M. Sini, *Obstacle and boundary determination from scattering data*, SIAM J. Math. Anal., to appear.
[20] G. Nakamura, R. Potthast, and M. Sini, *Unification of the probe and singular sources methods for the inverse boundary value problem by the no-response test*, Comm. Partial Differential Equations, 31 (2006), pp. 1505–1528.
[21] R. Potthast, *Point Sources and Multipoles in Inverse Scattering Theory*, Res. Notes Math. 427, Chapman-Hall/CRC, Boca Raton, FL, 2001.
[22] R. Potthast, *Sampling and probe methods—An algorithmical view*, Computing, 75 (2005), pp. 215–235.
[23] R. Potthast, *A survey on sampling and probe methods for inverse problems*, Inverse Problems, 22 (2006), pp. R1–R47.
[24] V. A. Solonnikov, *On Green's matrices for elliptic boundary value problems* (I), Proc. Steklov. Inst. Math., 110 (1970), pp. 123–170.
[25] V. A. Solonnikov, *The Green's matrices for elliptic boundary value problems* (II), in Boundary Value Problems of Mathematical Physics, Proceedings of the Steklov Institute of Mathematics 116, 1971, pp. 181–216; translated by the AMS, Providence, RI, 1973.

SIAM J. APPL. MATH.
Vol. 67, No. 4, pp. 1147–1165

DIFFUSION LIMITED REACTIONS*

BYRON GOLDSTEIN†, HAROLD LEVINE‡, AND DAVID TORNEY†§

Abstract. Changes to the relative separation of molecules or other interacting species on account of diffusion accompany their associative or dissociative reaction. The molecules are symbolized, for two distinct types, A, B, by the relations $A + B \quad AB$, and, if $[A]$, $[B]$, and $[AB]$ denote the corresponding densities, the equation $\frac{d}{dt}[AB] = k_+[A][B]$ specifies an associative process with *forward rate constant* k_+. An approximate version of the preceding takes the form of a linear differential equation, which can be employed to obtain significant estimates for both k_+ and the flux function $d[AB]/dt$. Such estimates are presented in different circumstances, including the localization of A, B on a common planar surface or their distribution in space, and also when the domain of A is a half space whereas that of B is a bounding planar surface. It proves advantageous to reformulate the last, a mixed boundary value problem, in terms of a linear integral equation. Biological applications are discussed, including the mechanism for the observed phosphorylation of proteins in resting cells and the incipience of phototransduction in rod photoreceptors.

Key words. Beltrami, Bessel function, cell membrane, cell signalling, chemical reaction, diffusion controlled, elliptic, equilibrium, integral equation, kinetics, ODE, PDE, rat basophilic leukemia (RBL) cell, rhodopsin, Smoluchowski, surface reaction, transducin, virial expansion

AMS subject classifications. Primary, 35G15, 80A30; Secondary, 76R50, 82B24, 92C05, 92C45

DOI. 10.1137/060655018

1. Introduction.

1.1. Diffusion limited reaction. Recall that diffusion limited reactions are those in which diffusion transports the reactants to "reaction range," where reaction takes place instantaneously. Diffusion and reaction introduce spatial correlations between molecules, which influence the rate of reaction. Nonequilibrium thermodynamics has been applied to the irreversible diffusion limited kinetics $A + B \to AB$ (cf. [8]); here the latter is encompassed as a special case of the following reversible reaction:

$$A + B \underset{k_-}{\overset{k_+}{\rightleftharpoons}} AB,$$

where k_+ and k_- denote the intrinsic *forward* and *reverse rate constants* associated, respectively, with molecular association, for which $d[AB]/dt$ equals $k_+[A][B]$, with reaction occurring at range $r = a$, and molecular dissociation, for which $d[AB]/dt = -k_-[AB]$, with dissociation occurring at range $r = d$ $(> a)$. (The $[A]$, $[B]$, etc., notations denote respective macroscopic concentrations, i.e., expected numbers of molecules inside a "unit measure.")

*Received by the editors March 23, 2006; accepted for publication (in revised form) February 5, 2007; published electronically May 14, 2007. This work was supported by the U.S.D.O.E. under contract W-7405-ENG-36 to the University of California.

http://www.siam.org/journals/siap/67-4/65501.html

†Theoretical Biology and Biophysics Group, Theoretical Division, Los Alamos National Laboratory, Mailstop K710, Los Alamos, NM 87545 (bxg@lanl.gov, dtorney@valornet.com). The first author's work was supported by NIH grant R37-GM35556.

‡Department of Mathematics, Stanford University, Stanford, CA 94305 (Hmathx@aol.com).

§Corresponding author.

Note that at equilibrium, $k_+[A][B] = k_-[AB]$. The obvious advantage of considering a system at equilibrium is that time is expunged from consideration. The cost is the introduction of the parameter d, but the latter will be seen to be advantageous in applications.

k_+ is expressible primarily in terms of the range parameters a and d and the diffusion coefficients, denoted D_A and D_B, respectively, with the sum $D_A + D_B$ hereafter denoted D. As will be seen, k_+ is obtainable from a pair density function, $\varrho(r)$: the density of A times the density of B at an arbitrary pair of respective points, lying at range r from one another. Differential equations for $\varrho(r)$ comprise triple-density functions [21]. However, using the independent-pairs approximation for triple densities, these equations become effectively linear in $\varrho(r)$, with dissociation modeled by a Dirac delta-function source.

$\varrho(r)$ is obtained for reactions occurring in $\mathbb{R}^2$ and in $\mathbb{R}^3$ (see sections 2 and 3). The inclusion of triple densities affords improvement upon the Smoluchowski theory [13]. This pioneering theory is based on an oversimplified model, but its shortcomings obtrude only when the theory is applied in challenging settings. The theory yields, for instance, the quandary that $k_+ = 0$ for reactions in $\mathbb{R}^2$. Our approach affords direct resolution thereof (cf. [8]) and, furthermore, yields two (or more) terms of asymptotic virial expansions for k_+. Section 4 addresses an interfacial reaction, combining aspects of reactions in $\mathbb{R}^2$ and $\mathbb{R}^3$. Biological applications of our results are described in section 5, and section 6 collects the k_+'s derived herein.

1.2. Biological motivations. Mobile receptors diffusing over the surface of a biological cell allow the cell to sense its environment and to respond to it. For many types of receptors, including growth-factor receptors, cytokine receptors, and immune recognition receptors, the aggregation of these receptors is required for the turning on or off of a cellular response. Therefore, for these receptors, the binding of the ligand to the receptor is insufficient to trigger a cellular response. Instead, the ligand induces the receptors to aggregate, holding the receptors in proximity for times long enough for various chemical modifications to occur on their cytoplasmic domains.

For example, growth-factor receptors are protein tyrosine kinases (PTKs), enzymes that can add a phosphate group to tyrosine residues. Growth-factor receptors also contain tyrosines in their cytoplasmic domains—whose phosphorylation induces specific functions, such as controlling the PTK's enzymatic activity and determining which cytoplasmic proteins (adapters and other participating enzymes) it can associate with. When a growth factor induces two receptors to aggregate, a change occurs inside the cell: the cytoplasmic domains of the receptors become phosphorylated, creating binding sites for proteins participating in the chemical cascade that leads to a cellular response. This is the mechanism for conveying to the inside of the cell the information that a hormone is present outside the cell, initiating what is referred to as cell signalling.

Quantitative insights into the kinetics of the primary events in receptor-initiated cell signalling require estimates for the forward and reverse rate constants for the steps that lead to receptor aggregation. Here, by means of a straightforward theory, we obtain the diffusion limited forward rate constant for a reaction between two species diffusing in a plane: a model applicable to reactions in cell membranes. This constant, which is related to the mean first passage time for two reactants to approach within a given distance, puts an upper bound on the rate at which (irreversible) receptor aggregation can occur. Furthermore, in section 5.1 we point out the importance of this rate constant, and of the corresponding reverse rate constant, for understanding

the background phosphorylation of receptors that is observed in the absence of any ligand-induced receptor aggregation.

Many cellular responses to hormones, e.g., growth factors, are mediated by cyclic nucleotides. Somewhat analogously, in photoreceptors, photons substitute for hormones as the effectors. Diffusion limited reactions taking place on phospholipid membranes—the discs of photoreceptors—are involved in phototransduction, and, in section 5.2, we quantitatively analyze the kinetics of its first reaction, the activation of the protein transducin, via the theoretical rate constant.

Although our main motivation is reactions on cell surfaces, we also present results for diffusion limited reactions in three space to verify that our theory reproduces an expression previously obtained by other means [3] and to intimate the range of applicability of our approach. Also, in section 4, we generalize a result concerning the diffusion limited forward rate constant for an interfacial reaction with ligands in solution binding to stationary receptors on a planar surface [2]. By means of a PDE with mixed boundary conditions, the receptors are also allowed to diffuse; we obtain the dependence of the diffusion limited forward rate constant upon both the diffusion coefficient of ligands in solution and the diffusion coefficient of receptors in the plane.

1.3. Fundamentals. For the present aims, the desideratum is the "reaction flux," Φ, whose dimensions are those of a concentration divided by time, and whose interpretation is the rate of creation of $[AB]$; whence, from the foregoing kinetic formulae,

$$\Phi = k_+[A][B].$$

As will be seen, Φ is obtained from the A B pair-density $\varrho(r)$, $r = a$. The latter function satisfies an equation comprising triple-density functions $\tau_\alpha(r, s, \theta)$ and $\tau_\beta(r, s, \theta)$ [21], having dimensions of the cube of a concentration. These subscripts distinguish two types of triples—B A B and A B A, respectively. These triples are viewed as being spatially centered on the sandwiched molecule, with their subscript connoting the latter, i.e., α for A and β for B. Furthermore, r and s denote the distances between the (centers of the) first and middle and between the middle and last molecules of the foregoing triples, and θ denotes the angle, at the central molecule, between the displacement vectors to the first and last, $0 \le \theta \le \pi$.

Similarly, an equation for τ includes quadruple-density functions, etc. However, as will be seen, the (unformulated) hierarchy may be pruned back by using the "independent-pairs" approximation for triple densities, yielding an equation involving only $\varrho(r)$ [21]. The additional natural assumption of constant pair densities for like-molecule pairs yields the following expressions for the τ's as a product of two ϱ's divided by a concentration:

$$\tau_\alpha(r, s, \theta) \doteq \varrho(r)\varrho(s)/[A], \quad \tau_\beta(r, s, \theta) \doteq \varrho(r)\varrho(s)/[B], \quad a \le r, s. \tag{1}$$

These constitute our *independent-pairs* approximations.

2. Reactions in planar surfaces. Here molecular trajectories are restricted to a plane. $\varrho(r)$ has the dimensions of L^{-4} and satisfies the following equation. For $a \le r < \infty$,

$$\frac{D}{r}\frac{d}{dr}r\frac{d}{dr}\varrho(r) + \left[2Da\int_0^\pi \left\{\frac{\partial}{\partial s}\tau_\alpha(R, s, \theta)\bigg|_{s=a} + \frac{\partial}{\partial s}\tau_\beta(R, s, \theta)\bigg|_{s=a}\right\} d\theta\right]\bigg|_{R=r}^{R=\infty} = -\frac{\Phi\delta(r-d)}{2\pi d}, \tag{2}$$

with boundary conditions $\varrho(a) = 0$ and $\varrho(\infty) = [A][B]$, and where Φ, with the dimensions of $L^{-2}T^{-1}$, is given by

$$\Phi = 2\pi a D \left.\frac{d\varrho(r)}{dr}\right|_{r=a}. \tag{3}$$

The foregoing equation is exact, which may be seen as follows.

The rate at which A B pairs react, at $R = a$, given in (3), is balanced by a dissociative flux, at $R = b$, which is the rate of creation of such pairs at range d, implementing microscopic balance. (For a physical interpretation of delta-function sources, see [11, p. 6]). Triples may be used to model the reaction of either molecule of such pairs with a suitable third molecule [21]. Thus, triple densities contribute an inhomogeneous term: the one contained in square brackets. The $R = r$ term reproduces the (net) rate of diminution of $\varrho(r)$ due to reaction of either member of the corresponding pair with appropriate third molecules, and the corresponding $R = \infty$ term is a constant, included so that the Laplacian of $\varrho(r)$ vanishes as $r \to \infty$, to reproduce the Smoluchowski theory at long range.

Thus, spatial correlations among triples of molecules influence $\varrho(r)$ and thereby Φ. Substituting the independent-pairs approximations (1) and performing the integrals in (2), using (3), gives

$$\frac{D}{r}\frac{d}{dr}r\frac{d}{dr}\varrho(r) + \Phi([A][B] - \varrho(r))\left(\frac{1}{[A]} + \frac{1}{[B]}\right) = -\frac{\Phi\delta(r-d)}{2\pi d}. \tag{4}$$

The leftmost term is the cylindrically symmetric Laplacian. The boundary conditions are as above. The weak nonlinearity of (4) poses no impediment to its solution.

Let $\chi(r) = [A][B] - \varrho(r)$; let $\bar{a} = a/\ell$, $\bar{d} = d/\ell$, and $\bar{r} = r/\ell$, where

$$\ell \stackrel{\text{def}}{=} \left[\frac{\Phi}{D}\left(\frac{1}{[A]} + \frac{1}{[B]}\right)\right]^{-1/2}. \tag{5}$$

Adopting these notations, (4) becomes

$$\frac{1}{\bar{r}}\frac{d}{d\bar{r}}\bar{r}\frac{d}{d\bar{r}}\chi(\bar{r}) - \chi(\bar{r}) = \frac{\Phi\delta(\bar{r}-\bar{d})}{2\pi D\bar{d}}, \tag{6}$$

with $\chi(\bar{a}) = [A][B]$ and $\chi(\infty) = 0$. Thus,

$$\chi(\bar{r}) = \begin{cases} C_1 I_0(\bar{r}) + C_2 K_0(\bar{r}), & \bar{a} \le \bar{r} \le \bar{d}, \\ C_3 K_0(\bar{r}), & \bar{d} \le \bar{r}, \end{cases}$$

where I_0 and K_0 denote modified Bessel functions of index zero. The constants C_1, C_2, and C_3 are to be obtained using three conditions, namely, $\chi(\bar{a}) = [A][B]$, the continuity of χ at $\bar{r} = \bar{d}$, and the jump condition required by (6), namely,

$$\bar{r}\left.\frac{d\chi}{d\bar{r}}\right|_{\bar{d}-0}^{\bar{d}+0} = \frac{\Phi}{2\pi D}.$$

These conditions yield

$$C_1 = -\frac{\Phi K_0(\bar{d})}{2\pi D},$$

$$C_2 = \frac{[A][B]}{K_0(\bar{a})} + \frac{\Phi I_0(\bar{a})K_0(\bar{d})}{2\pi D K_0(\bar{a})}, \qquad \text{and}$$

$$C_3 = \frac{[A][B]}{K_0(\bar{a})} + \frac{\Phi(I_0(\bar{a})K_0(\bar{d}) - I_0(\bar{d})K_0(\bar{a}))}{2\pi D K_0(\bar{a})}.$$

Thus, from (3),

$$\Phi = -2\pi D\bar{a}\,\frac{d\chi(\bar{r})}{d\bar{r}}\bigg|_{\bar{r}=\bar{a}} = 2\pi D\left\{\frac{\bar{a}K_1(\bar{a})[A][B]}{K_0(\bar{a})} + \frac{\Phi K_0(\bar{d})}{2\pi D K_0(\bar{a})}\right\}$$

or

$$\Phi = \frac{2\pi D\bar{a}K_1(\bar{a})[A][B]}{K_0(\bar{a}) - K_0(\bar{d})}. \tag{7}$$

With reference to (5), the foregoing may be regarded as implicit equations in Φ, but it simplifies matters and suffices for the present aims to focus on two limiting cases, (i) $\bar{a} < \bar{d} \to 0$ and (ii) $\bar{a} \to 0$, $\bar{d} \to \infty$, germane to irreversible reactions. Successive approximations yield, for (i),

$$\Phi \sim \frac{2\pi D[A][B]}{\log d/a},$$

where the omitted terms are $\mathcal{O}(\bar{a}^2 \log \bar{a} + \bar{d}^2 \log \bar{d})$, establishing that corrections to Φ are substantially smaller than the given term. Similarly, for (ii), these approximations yield

$$\Phi \sim \frac{2\pi D[A][B]}{-\log\sqrt{\psi}}\left(1 + \frac{\log\sqrt{-\log\sqrt{\psi}} + \frac{1}{2}\log 2 - \gamma}{\log\sqrt{\psi}} + \cdots\right), \tag{8}$$

where ψ is the expected number of A molecules plus B molecules within a circle of radius a, $\psi = \pi a^2([A] + [B]) \ll 1$, and where $\gamma \doteq 0.577\ldots$ denotes the Euler–Mascheroni constant. One may note, to leading order, that the functional dependence of k_+ upon a (via ψ) was previously obtained from nonequilibrium thermodynamics [8, (121)–(122)].

3. Bulk reactions. Here the molecular trajectories are in Euclidean three space, and the notations and definitions of section 2 are also adopted. In analogy to (4), the differential equation is

$$\frac{D}{r^2}\frac{d}{dr}r^2\frac{d}{dr}\varrho(r) + \Phi([A][B] - \varrho(r))\left(\frac{1}{[A]} + \frac{1}{[B]}\right) = -\frac{\Phi\delta(r-d)}{4\pi d^2}, \tag{9}$$

with the previous boundary conditions. This $\varrho(r)$ has the dimensions of L^{-6}. The left-hand term of (9) is the spherically symmetric Laplacian. Here Φ, with the dimensions of $L^{-3}T^{-1}$, is given by

$$\Phi = 4\pi a^2 D\,\frac{d\varrho(r)}{dr}\bigg|_{r=a}. \tag{10}$$

As above, (9) may be written

$$\frac{1}{\bar{r}^2}\frac{d}{d\bar{r}}\bar{r}^2\frac{d}{d\bar{r}}\chi(\bar{r}) - \chi(\bar{r}) = \frac{\Phi\delta(\bar{r}-\bar{d})}{4\pi D\bar{d}^2\ell},$$

with $\chi(\bar{a}) = [A][B]$ and $\chi(\infty) = 0$. Thus,

$$\chi(\bar{r}) = \begin{cases} C_1 \dfrac{e^{\bar{r}}}{\bar{r}} + C_2 \dfrac{e^{-\bar{r}}}{\bar{r}}, & \bar{a} \le \bar{r} \le \bar{d}, \\ C_3 \dfrac{e^{-\bar{r}}}{\bar{r}}, & \bar{d} \le \bar{r}. \end{cases}$$

C_1, C_2, and C_3 are obtainable as before; here the jump condition is

$$\bar{r}^2 \left. \frac{d\chi}{d\bar{r}} \right|_{\bar{d}-0}^{\bar{d}+0} = \frac{\Phi}{4\pi D\ell} \stackrel{\text{def}}{=} \Phi^{3/2}\Gamma,$$

where reference to (5) yields

$$\Gamma = \frac{1}{4\pi D\ell\sqrt{\Phi}} = \frac{1}{4\pi D^{3/2}} \sqrt{\frac{1}{[A]} + \frac{1}{[B]}}. \tag{11}$$

It follows that

$$C_1 = -\frac{\Phi^{3/2}\Gamma e^{-\bar{d}}}{2\bar{d}},$$

$$C_2 = \bar{a}e^{\bar{a}}[A][B] + \frac{\Phi^{3/2}\Gamma e^{2\bar{a}-\bar{d}}}{2\bar{d}}, \qquad \text{and}$$

$$C_3 = \bar{a}e^{\bar{a}}[A][B] + \frac{\Phi^{3/2}\Gamma(e^{2\bar{a}-\bar{d}} - e^{\bar{d}})}{2\bar{d}}.$$

Thus, from (10), using (11),

$$\Phi = -\frac{4\pi Da^2}{\ell} \left. \frac{d\chi(\bar{r})}{d\bar{r}} \right|_{\bar{r}=\bar{a}} = 4\pi Da \left\{ [A][B](1+\bar{a}) + \frac{\Phi e^{\bar{a}-\bar{d}}}{4\pi Dd} \right\}.$$

As before, we address the limits (i) $\bar{a} < \bar{d} \to 0$ and (ii) $\bar{a} \to 0$, $\bar{d} \to \infty$. For (i),

$$\Phi \sim \frac{4\pi Da[A][B](1 + o(\sqrt{\phi}))}{1 - a/d};$$

i.e., the first correction is $\mathcal{O}(\phi)$, where ϕ is the expected number of A molecules plus B molecules within a sphere of radius a: $\phi = \frac{4}{3}\pi a^3([A] + [B]) \ll 1$. Similarly, for (ii),

$$\Phi \sim 4\pi Da[A][B](1 + \sqrt{3\phi} + \cdots).$$

The leading-order behavior reproduces the Smoluchowski rate, and the first-order correction agrees, for the case $[A] \ll [B]$, with that of [3, (1.3)].

4. Reaction on an interface. Here the B's and the AB's diffuse in the plane $z = 0$, whereas the A's diffuse in the half space $z \ge 0$. Alternative interfacial diffusion limited reactions have been considered (cf. [2, 5]).

4.1. Partial differential equation. The A B pair density is denoted $\varrho(r,z)$, with $r=\sqrt{x^2+y^2}$ and with r and z denoting the magnitudes of perpendicular components of displacement. This density is the product of the concentrations of A and B. Thus, $\varrho(r,z)$ has the dimensions of L^{-5}. Here it is taken to be the solution of the Smoluchowski differential equation

$$\left(\frac{D}{r}\frac{\partial}{\partial r}r\frac{\partial}{\partial r}+\frac{D_A\partial^2}{\partial z^2}\right)\varrho(r,z)=-\frac{\Phi\delta(z)\delta(r-d)}{2\pi d}, \tag{12}$$

with boundary conditions $\varrho(r,0)=0$; $0<r<a$, $\partial\varrho(r,0)/\partial z=0$; $a<r<\infty$, $\varrho(r,z)\to[A][B]$ as $r^2+z^2\to\infty$. Here

$$\Phi=2\pi D_A\int_0^a\frac{\partial\varrho(r,0)}{\partial z}rdr. \tag{13}$$

Note that Φ has the dimensions of $L^{-2}T^{-1}$: the rate of generation of $[AB]$, in the plane $z=0$. As before, the source term of (12) is distinguished from one which characterizes a fixed-source Green's function of potential theory by the appearance of the factor Φ, a quantity related to $\varrho(r,z)$ via (13).

The simplifications leading to (12) include planar sources and sinks, furnishing a half-space problem with mixed boundary conditions. Furthermore, because the current aim is exploratory, the term derived from triple densities and (1),

$$\Phi([A][B]-\varrho(r,z))\left(\frac{\delta(z)}{[A]}+\frac{1}{[B]}\right), \tag{14}$$

is omitted from the left-hand side of (12). We conjecture that the k_+ obtained from the solution of (12) is nevertheless correct to leading order (in the concentrations). (Note that (14) might require modification, for instance, at $r=a$, $z=0$, for well-posedness.)

Let

$$r=\sqrt{D}R,\qquad z=\sqrt{D_A}Z; \tag{15}$$

then (12) takes the form

$$\left(\frac{1}{R}\frac{\partial}{\partial R}R\frac{\partial}{\partial R}+\frac{\partial^2}{\partial Z^2}\right)\varrho(R,Z)=-\frac{\Phi\delta(Z)\delta(R-d/\sqrt{D})}{2\pi d\sqrt{D_AD}}.$$

Let

$$\varrho(R,Z)=[A][B]-\beta(R,Z),$$

and the system defining $\beta(R,Z)$ comprises

$$\nabla^2\beta(R,Z)=\left(\frac{1}{R}\frac{\partial}{\partial R}R\frac{\partial}{\partial R}+\frac{\partial^2}{\partial Z^2}\right)\beta(R,Z)=\frac{\Phi\delta(Z)\delta(R-d/\sqrt{D})}{2\pi d\sqrt{D_AD}}, \tag{16}$$

$$\beta(R,0)=[A][B],\qquad 0<R\le a/\sqrt{D}, \tag{17}$$

$$\frac{\partial}{\partial Z}\beta(R,0)=0,\qquad a/\sqrt{D}<R, \tag{18}$$

$$\beta\to 0,\qquad R^2+Z^2\to\infty. \tag{19}$$

4.2. An associated integral equation. To recast the boundary value problem defined by (16)–(19) in terms of an integral equation, introduce a Green's function $G(R, Z; R', Z')$ specified by

$$\nabla^2 G = \left(\frac{1}{R}\frac{\partial}{\partial R} R \frac{\partial}{\partial R} + \frac{\partial^2}{\partial Z^2}\right) G = -\frac{\delta(R-R')\delta(Z-Z')}{2\pi R}, \qquad 0 < R, R', Z, Z' < \infty, \tag{20}$$

$$\frac{\partial}{\partial Z} G = 0, \quad Z = 0, \quad R > 0, \quad \text{and} \quad G \to 0, \quad R^2 + Z^2 \to \infty.$$

If

$$\bar{G}(\zeta, Z; R', Z') = \int_0^\infty R J_0(\zeta R) G(R, Z; R', Z') dR, \tag{21}$$

where J_0 denotes the Bessel function of the first kind of order zero, then it follows from (20) that

$$\left(\frac{\partial^2}{\partial Z^2} - \zeta^2\right) \bar{G}(\zeta, Z; R', Z') = -\frac{J_0(\zeta R')\delta(Z-Z')}{2\pi},$$

and explicitly

$$\bar{G}(\zeta, Z; R', Z') = \frac{J_0(\zeta R')}{2\pi|\zeta|} \cosh \zeta Z_< \exp^{-|\zeta| Z_>},$$

with $Z_<$ and $Z_>$ denoting the smaller and larger of Z and Z', respectively. The inverse of the transform (21) thus yields the Green's function representation

$$G(R, Z; R', Z') = \int_0^\infty \zeta J_0(\zeta R) \bar{G}(\zeta, Z; R', Z') d\zeta = \frac{1}{2\pi} \int_0^\infty J_0(\zeta R) J_0(\zeta R') \cosh \zeta Z_< \exp^{-\zeta Z_>} d\zeta. \tag{22}$$

Apply Green's integral relation

$$\int (G \nabla^2 \beta - \beta \nabla^2 G) dV = \int \left(G \frac{\partial}{\partial n}\beta - \beta \frac{\partial}{\partial n} G\right) dS$$

in the half space $Z \geq 0$, where the surface integral consists of a hemispherical portion and a plane circular domain where $Z = 0$. Invoking the various conditions obeyed by β and G, it follows that

$$\frac{\Phi}{D\sqrt{D_A}} G(d/\sqrt{D}, 0; R', Z') + \beta(R', Z') = \int_0^{a/\sqrt{D}} G(R, 0; R', Z') 2\pi R \sigma(R) dR, \tag{23}$$

with

$$\sigma(R) = -\left.\frac{\partial \beta}{\partial Z}\right|_{Z=0}, \qquad 0 < R < a/\sqrt{D}. \tag{24}$$

After changing variables in (23) (with regard for the argument symmetry of G), one obtains, for $Z \geq 0$,

$$\begin{aligned}\beta(R,Z) &= -\frac{\Phi}{D\sqrt{D_A}}G(R,Z;d/\sqrt{D},0) + 2\pi\int_0^{a/\sqrt{D}} G(R,Z;R',0)\sigma(R')R'dR' \\ &= -\frac{\Phi}{2\pi D\sqrt{D_A}}\int_0^\infty J_0(\zeta R)J_0(\zeta d/\sqrt{D})\exp^{-\zeta Z}\,d\zeta \\ &\quad + \int_0^{a/\sqrt{D}} R'\sigma(R')dR'\int_0^\infty J_0(\zeta R)J_0(\zeta R')\exp^{-\zeta Z}\,d\zeta. \qquad (25)\end{aligned}$$

The representation (25) for $\beta(R,Z)$ implies [11, (6.3.62)] that

$$\frac{\partial}{\partial Z}\beta(R,Z)\bigg|_{Z=0} = -\sigma(R), \qquad R < a/\sqrt{D}, \qquad (26)$$

and

$$\frac{\partial}{\partial Z}\beta(R,Z)\bigg|_{Z=0} = 0, \qquad a/\sqrt{D} < R \ \ (\neq d/\sqrt{D}),$$

in accord with (24).

An integral equation for $\sigma(R)$ is next secured on imposing the condition (17), namely, for $R < a/\sqrt{D}$,

$$\begin{aligned}[A][B] &= -\frac{\Phi}{2\pi D\sqrt{D_A}}\int_0^\infty J_0(\zeta R)J_0(\zeta d/\sqrt{D})d\zeta \\ &\quad + \int_0^{a/\sqrt{D}} \sigma(R')R'dR'\int_0^\infty J_0(\zeta R)J_0(\zeta R')d\zeta. \qquad (27)\end{aligned}$$

From (13), (15), and (26),

$$\Phi = 2\pi D\sqrt{D_A}F, \qquad F \stackrel{\text{def}}{=} \int_0^{a/\sqrt{D}} \sigma(R)RdR, \qquad (28)$$

enabling (27) to be rewritten in the forms

$$\begin{aligned}&[A][B] + F\int_0^\infty J_0(\zeta R)J_0(\zeta d/\sqrt{D})d\zeta \\ &\quad = \int_0^{a/\sqrt{D}} \sigma(R')R'dR'\int_0^\infty J_0(\zeta R)J_0(\zeta R')d\zeta, \qquad R \leq a/\sqrt{D}, \qquad (29)\end{aligned}$$

or

$$\begin{aligned}V(R) &= \int_0^{a/\sqrt{D}} \sigma(R')R'dR'\int_0^\infty J_0(\zeta R)J_0(\zeta R')d\zeta \\ &= \int_0^\infty J_0(\zeta R)\varphi(\zeta)d\zeta, \qquad R < a/\sqrt{D},\end{aligned}$$

where

$$V(R) = [A][B] + F\int_0^\infty J_0(\eta R)J_0(\eta d/\sqrt{D})d\eta \qquad (30)$$

and

$$\varphi(\zeta) = \int_0^\infty R'\sigma(R')J_0(\zeta R')dR',$$

$$\sigma(R) = 0, \qquad R > a/\sqrt{D},$$

$$\sigma(R) = \int_0^\infty \zeta\varphi(\zeta)J_0(\zeta R)d\zeta. \tag{31}$$

Beltrami established the representation

$$\varphi(\zeta) = \frac{2}{\pi}V(0)\frac{\sin\zeta a/\sqrt{D}}{\zeta} + \frac{2}{\pi}\int_0^{a/\sqrt{D}} \bar{R}\cos\zeta\bar{R}d\bar{R}\int_0^{\bar{R}} \frac{V'(R')}{\sqrt{\bar{R}^2 - R'^2}}dR' \tag{32}$$

(cf. [1, 12]). In accordance with (30),

$$\frac{dV}{dR} = V'(R) = -F\int_0^\infty \eta J_1(\eta R)J_0(\eta d/\sqrt{D})d\eta,$$

whence

$$\begin{aligned}\int_0^{\bar{R}} \frac{V'(R')}{\sqrt{\bar{R}^2 - R'^2}}dR' &= -F\int_0^\infty \eta J_0(\eta d/\sqrt{D})d\eta \int_0^{\bar{R}} \frac{J_1(\eta R')}{\sqrt{\bar{R}^2 - R'^2}}dR' \\ &= -\frac{F}{\bar{R}}\int_0^\infty J_0(\eta d/\sqrt{D})(1 - \cos\eta\bar{R})d\eta \\ &= -\frac{F}{\bar{R}}\left\{\frac{\sqrt{D}}{d} - \frac{1}{\sqrt{d^2/D - \bar{R}^2}}\right\}, \qquad \bar{R} < d/\sqrt{D}. \end{aligned} \tag{33}$$

Also,

$$V(0) = [A][B] + F\frac{\sqrt{D}}{d}. \tag{34}$$

Thus, inasmuch as $\varphi(0) = F$, (32) implies

$$\begin{aligned} F &= \frac{2a}{\pi\sqrt{D}}\left[[A][B] + F\frac{\sqrt{D}}{d}\right] - \frac{2F}{\pi}\int_0^{a/\sqrt{D}} \left(\frac{\sqrt{D}}{d} - \frac{1}{\sqrt{d^2/D - \bar{R}^2}}\right)d\bar{R} \\ &= \frac{2a}{\pi}\frac{[A][B]}{\sqrt{D}} + \frac{2F}{\pi}\sin^{-1}\frac{a}{d} \end{aligned}$$

or

$$F\left\{1 - \frac{2}{\pi}\sin^{-1}\frac{a}{d}\right\} = \frac{2a}{\pi}\frac{[A][B]}{\sqrt{D}}. \tag{35}$$

From (28) and (35), taken together, the specification

$$\Phi = \frac{4a\sqrt{DD_A}[A][B]}{1 - \frac{2}{\pi}\sin^{-1}\frac{a}{d}} \tag{36}$$

follows.

4.3. Explicit form for $\sigma(R)$. From (32), via (31), there follows

$$\sigma(R) = \frac{2}{\pi}\int_0^\infty \zeta J_0(\zeta R)\left\{V(0)\frac{\sin\zeta a/\sqrt{D}}{\zeta} + \int_0^{a/\sqrt{D}} \bar{R}\cos\zeta\bar{R}d\bar{R}\int_0^{\bar{R}}\frac{V'(R')}{\sqrt{\bar{R}^2-R'^2}}dR'\right\}d\zeta.$$

Using (33) and (34), it is found that

$$\begin{aligned}\int_0^\infty \zeta J_0(\zeta R)\frac{2}{\pi}V(0)\frac{\sin\zeta a/\sqrt{D}}{\zeta}d\zeta &= \frac{2}{\pi}\left([A][B]+F\frac{\sqrt{D}}{d}\right)\int_0^\infty J_0(\zeta R)\sin\zeta a/\sqrt{D}d\zeta \\ &= \frac{2}{\pi}\left([A][B]+F\frac{\sqrt{D}}{d}\right)\frac{1}{\sqrt{a^2/D-R^2}}, \qquad R<a/\sqrt{D}.\end{aligned}$$

Further,

$$\begin{aligned}&\frac{2}{\pi}\int_0^{a/\sqrt{D}}\bar{R}\cos\zeta\bar{R}\int_0^{\bar{R}}\frac{V'(R')}{\sqrt{\bar{R}^2-R'^2}}dR'd\bar{R} \\ &\quad= -\frac{2F}{\pi}\int_0^{a/\sqrt{D}}\cos\zeta\bar{R}\left\{\frac{\sqrt{D}}{d}-\frac{1}{\sqrt{d^2/D-\bar{R}^2}}\right\}d\bar{R} \\ &\quad= \frac{2F}{\pi}\left\{-\frac{\sqrt{D}}{d}\frac{\sin\zeta a/\sqrt{D}}{\zeta}+\int_0^{a/\sqrt{D}}\frac{\cos\zeta\bar{R}}{\sqrt{d^2/D-\bar{R}^2}}d\bar{R}\right\},\end{aligned}$$

and thus

$$\begin{aligned}\sigma(R) = \frac{2}{\pi}\int_0^\infty \zeta J_0(\zeta R)\Bigg\{&\left([A][B]+F\frac{\sqrt{D}}{d}\right)\frac{\sin\zeta a/\sqrt{D}}{\zeta}-F\frac{\sqrt{D}}{d}\frac{\sin\zeta a/\sqrt{D}}{\zeta} \\ &+F\int_0^{a/\sqrt{D}}\frac{\cos\zeta\bar{R}}{\sqrt{d^2/D-\bar{R}^2}}d\bar{R}\Bigg\}d\zeta \\ = \frac{2}{\pi}\frac{[A][B]}{\sqrt{a^2/D-R^2}}&+\frac{2F}{\pi}\int_0^\infty\zeta J_0(\zeta R)\int_0^{a/\sqrt{D}}\frac{\cos\zeta\bar{R}}{\sqrt{d^2/D-\bar{R}^2}}d\bar{R}d\zeta, \qquad R<a/\sqrt{D}.\end{aligned}$$

Write

$$\begin{aligned}\int_0^{a/\sqrt{D}}\frac{\cos\zeta\bar{R}}{(d^2/D-\bar{R}^2)^{1/2}}d\bar{R} &= \int_0^{a/\sqrt{D}}\frac{1}{(d^2/D-\bar{R}^2)^{1/2}}d_{\bar{R}}\left(\frac{\sin\zeta\bar{R}}{\zeta}\right) \\ &= \frac{\sin\zeta a/\sqrt{D}}{\zeta\sqrt{(d^2-a^2)/D}}-\frac{1}{\zeta}\int_0^{a/\sqrt{D}}\frac{\bar{R}\sin\zeta\bar{R}}{(d^2/D-\bar{R}^2)^{3/2}}d\bar{R},\end{aligned}$$

and then

(37)
$$\sigma(R) = \frac{2}{\pi}\frac{[A][B]}{\sqrt{a^2/D - R^2}} + \frac{2F}{\pi}\left\{\frac{1}{\sqrt{(d^2-a^2)/D}\sqrt{a^2/D-R^2}} - \int_R^{a/\sqrt{D}} \frac{\bar{R}d\bar{R}}{(d^2/D-\bar{R}^2)^{3/2}(\bar{R}^2-R^2)^{1/2}}\right\}.$$

In accordance with the appendix,

(38)
$$\int_R^{a/\sqrt{D}} \frac{\bar{R}d\bar{R}}{(d^2/D-\bar{R}^2)^{3/2}(\bar{R}^2-R^2)^{1/2}} = \frac{1}{(d^2/D-R^2)}\sqrt{\frac{a^2-DR^2}{d^2-a^2}}$$

(a result which can be checked by considering $R = 0$), it follows from (37) that

$$\sigma(R) = \frac{2}{\pi}\frac{[A][B]}{\sqrt{a^2/D - R^2}} + \frac{2F}{\pi}\left\{\frac{1}{\sqrt{(d^2-a^2)/D}\sqrt{a^2/D-R^2}} - \frac{1}{(d^2/D-R^2)}\sqrt{\frac{a^2-DR^2}{d^2-a^2}}\right\},$$

whence the explicit determination

(39)
$$\sigma(R) = \frac{2}{\pi}\frac{[A][B]}{\sqrt{a^2/D-R^2}} + \frac{2F}{\pi\sqrt{D}}\frac{\sqrt{d^2-a^2}}{\sqrt{a^2/D-R^2}}\frac{1}{(d^2/D-R^2)}, \qquad 0 < R < a/\sqrt{D} < d/\sqrt{D},$$

follows.

A final check is called for: multiply by R in (39) and integrate over $0 < R < a/\sqrt{D}$, which yields

$$F = \frac{2}{\pi}[A][B]\frac{a}{\sqrt{D}} + \frac{2F}{\pi}\sqrt{\frac{d^2-a^2}{D}}\int_0^{a/\sqrt{D}} \frac{RdR}{\sqrt{a^2/D-R^2}(d^2/D-R^2)}.$$

Since

$$\int_0^{a/\sqrt{D}} \frac{RdR}{\sqrt{a^2/D-R^2}(d^2/D-R^2)} = \frac{a}{2\sqrt{D}}\int_0^1 \frac{dx}{\sqrt{1-x}(d^2/D-a^2x/D)}$$
$$= \frac{a\sqrt{D}}{2d^2}\int_0^1 \frac{dx}{\sqrt{1-x}(1-\frac{a^2}{d^2}x)} = \frac{a\sqrt{D}}{2d^2}\frac{2}{\frac{a}{d}\sqrt{1-\frac{a^2}{d^2}}}\tan^{-1}\left(\frac{-\frac{a^2}{d^2}\sqrt{1-x}}{\frac{a}{d}\sqrt{1-\frac{a^2}{d^2}}}\right)\Bigg|_{x=0}^{x=1}$$
$$= \frac{\sqrt{D}}{\sqrt{d^2-a^2}}\tan^{-1}\left(\frac{a}{d}\frac{1}{\sqrt{1-\frac{a^2}{d^2}}}\right) = \frac{\sqrt{D}}{\sqrt{d^2-a^2}}\sin^{-1}\frac{a}{d},$$

equation (35) follows.

4.4. Irreversible reaction. One may directly obtain the limit $d \to \infty$ of (36) by omitting the "source" term on the right-hand side of (12). This implies the setting aside of the first term on the left-hand side of (27), yielding the integral equation

$$[A][B] = \int_0^{a/\sqrt{D}} \sigma(R')R'dR' \int_0^\infty J_0(\zeta R)J_0(\zeta R')d\zeta, \qquad R < a/\sqrt{D}.$$

This equation applies to a classical mixed boundary value problem for an axially symmetric potential/harmonic function in the half space $Z > 0$. Write

$$[A][B] = \int_0^\infty J_0(\zeta R)\varphi(\zeta)d\zeta, \qquad R < a/\sqrt{D}, \tag{40}$$

where

$$\varphi(\zeta) = \int_0^\infty R'\sigma(R')J_0(\zeta R')dR',$$

$$\sigma(R) = 0, \qquad R > a/\sqrt{D};$$

then,

$$\sigma(R) = \int_0^\infty \zeta\varphi(\zeta)J_0(\zeta R)d\zeta = 0, \qquad R > a/\sqrt{D}. \tag{41}$$

The dual equations (40), (41) are satisfied with the specification

$$\varphi(\zeta) = \frac{2[A][B]}{\pi}\frac{\sin\zeta a/\sqrt{D}}{\zeta}$$

because

$$\int_0^\infty J_0(\zeta R)\frac{\sin\zeta a/\sqrt{D}}{\zeta}d\zeta = \frac{\pi}{2}, \qquad R < a/\sqrt{D},$$

and

$$\sigma(R) \propto \int_0^\infty J_0(\zeta R)\sin(\zeta a/\sqrt{D})d\zeta = 0, \qquad R > a/\sqrt{D}.$$

Furthermore,

$$\sigma(R) = \frac{2}{\pi}[A][B]\int_0^\infty J_0(\zeta R)\sin(\zeta a/\sqrt{D})d\zeta = \frac{2}{\pi}\frac{[A][B]}{\sqrt{a^2/D - R^2}}, \qquad R < a/\sqrt{D},$$

the limit $d \to \infty$ of (39). From (28), one obtains

$$\Phi = 4a\sqrt{DD_A}[A][B],$$

the corresponding limit of (36).

5. Applications to biological cell signalling.

5.1. Basal level of receptor phosphorylation. As described in the introduction, for a number of receptor families, ligand-induced receptor aggregation is rapidly followed by phosphorylation of sites on cytoplasmic and membrane-associated proteins. Even in the absence of an outside stimulus, cells show basal levels of protein phosphorylation (e.g., [15, 18]). In the cell there is a constant competition between enzymes that phosphorylate their substrates (kinases) and enzymes that reverse this phosphorylation and dephosphorylate their substrates (phosphatases). In the absence of a ligand which binds to receptors and induces them to aggregate, diffusion

will nonetheless bring them into proximity, although these aggregates will be short-lived, as diffusion will also lead to their separation. If these receptors are kinases (growth-factor receptors) or are associated with kinases (immune recognition and cytokine receptors), then they may phosphorylate their neighbor, leading to the observed basal phosphorylation.

These adjacent pairs, referred to as dimers, are modeled by the homogeneous reaction $A + A \rightleftharpoons A_2$. Linearization of a corresponding formula suggests that, subject to the following proviso, the forward rate constants derived for the heterogeneous reaction analyzed herein are also applicable to the homogeneous reaction (at least to leading order): based on combinatorial considerations, the forward rate constant for the homogeneous reaction is taken to be half that of the heterogeneous reaction, given by (8).

Given k_+, the easiest way to obtain an expression for k_-, the reverse rate constant governing dissociation of these dimers, is from the diffusion limited equilibrium constant

$$K = \frac{k_+}{k_-}.$$

For the situation of interest, the dilute limit where the typical distance between receptors is much larger than a, $K \doteq \pi a^2$, provided that the receptors are treated as point particles [4]. When the receptors are treated instead as hard discs of diameter b, so that their centers must be at least a distance b apart, then [22]

$$K = \pi a^2(1 - (b/a)^2). \tag{42}$$

To derive (42), focus on a "reference receptor." Because of the assumption of diluteness, the centers of other receptors are distributed approximately as a spatial Poisson process with density ρ_T outside a disc of diameter b centered on the reference receptor. Recall that the "distance between receptors" is the distance between their centers. Thus, for another receptor to be within a distance a of the reference receptor, its center must lie within an annulus of inner radius b and outer radius a. From the Poisson distribution for the number of points in a given area in a spatial Poisson process, the probability P_0 that the reference receptor has no other receptors within a, and the probability P_1 that there is exactly one other receptor, are given by

$$P_0 = e^{-\rho_T \pi (a^2 - b^2)} = e^{-\psi(1-(b/a)^2)},$$

$$P_1 = \rho_T \pi (a^2 - b^2) e^{-\rho_T \pi (a^2 - b^2)} = \rho_T \pi a^2 (1 - (b/a)^2)) e^{-\psi(1-(b/a)^2)}.$$

Multiplying the above equations by ρ_T gives us the equilibrium concentrations of isolated receptors, $[A]$, and dimers, $[D]$. When these equations are expanded, to leading order in ψ

$$[D] \approx [A]^2 \pi a^2 (1 - (b/a)^2), \tag{43}$$

and because $K \approx [D]/[A]^2$, (42) follows. Thus

$$k_- = \frac{k_+}{\pi a^2 (1 - (b/a)^2)}. \tag{44}$$

5.1.1. Estimates for k_+ and k_-. We estimate the values of the rate constants for a well-studied receptor system: the high affinity receptor, FcεRI, that binds IgE antibody on the surface of rat basophilic leukemia (RBL) cells. This receptor plays a central role in allergic reactions, as its activation leads to the release of histamine and other mediators of allergic reactions.

An RBL cell has approximately 3×10^5 FcεRIs on its cell surface and has a surface area of about 5×10^{-6} cm^2, so that $\rho_T = 6 \times 10^{10}$ cm^{-2}. Unlike, for example, growth-factor receptors, FcεRI is not an intrinsic kinase. However, it associates with a tyrosine kinase called Lyn, and when Lyn is bound to a receptor it phosphorylates other receptors in its proximity.

We will take a to be the maximum distance that a receptor associated with Lyn can be from its neighbor and still phosphorylate it. Experiments from another receptor system suggest $a \approx 40$ Å [14]. With this value of a, $\psi = \pi a^2 \rho_T = 0.03$. A reasonable range for b is 10–20 Å. Taking $b = 10$ Å, from (42), $K = 5.4 \times 10^{-13}$ cm^2. The diffusion coefficient of FcεRI on RBL, $D_A = 2$–3×10^{-10} cm^2/s (see [10, 17]). The last parameter requiring estimation is d: the separation distance which two receptors, initially a dimer, must achieve before they are deemed dissociated. We take d to be the average nearest-neighbor distance so that $d = \sqrt{\rho_T}/2 = 2 \times 10^{-6}$ cm. Rather than solving the implicit equation (7) for Φ, since ψ is small, (8) may be used to obtain an estimate for k_+. For the parameters given, $k_+ = 8.7 \times 10^{-10}$ cm^2/s, and from (44), $k_- = 1.85 \times 10^3$ s^{-1}. Thus, the average lifetime of an FcεRI dimer on an RBL cell, which aggregates and dissociates purely by diffusion, is $1/k_- = 5.4 \times 10^{-4}$ s.

The obvious question this result poses is, For such short dimer lifetimes can phosphorylation occur at the levels seen? By using these rate constants in the FcεRI signalling model of [6], this question may be further investigated.

5.2. Phototransduction. Diffusion limited reactions in $\mathbb{R}^2$ are thought to occur in the primary amplification steps of phototransduction, which accrue high levels of photosensitivity in the dark-adapted rod photoreceptor. When a photon is absorbed by the *retinal* moiety of a rhodopsin molecule, R, photoisomerization ensues, inducing a conformational change in rhodopsin: $R \to R^*$. Then R^*, diffusing in the phospholipid membrane of "discs," catalytically activates the G-protein transducin, T. This reaction,

$$R^* + T \to R^* + T^*,$$

is one which is thought to be diffusion limited. In the time scale of interest, the decay $R^* \to R$ may be ignored, and R^* effectively acts as a catalyst. The method of section 2 is readily adapted to encompass this reaction. As before, the reagents, R^* and T, may not coexist at a separation a. The novelty lies in the triple-density terms. Instead of including two terms, there is only the one pertaining to $R^*\,T\,R^*$, because it is not possible for two T's to interfere with one another in reacting with one R^*: this reaction is catalytic. This results in the substitution of $1/[T]$ for the coefficient $(1/[R^*] + 1/[T])$, obtaining a k_+ in the form (8) but with a $\psi = \pi a^2 [R^*]$.

We estimate the parameter a of this k_+, using the stochastic-simulation data of [9], whose parameters arise from experimental data from amphibian rods. We have $D_{R^*} = 0.7 \times 10^{-8}$cm^2/s, $D_T = 1.2 \times 10^{-8}$cm^2/s, and the (initial) concentrations are $[R^*] = 2.5 \times 10^7$/cm^2 and $[T] = 2.5 \times 10^{11}$/cm^2 (molecules/cm^2) [9, Figure 3]. Using $k_+ \sim -2\pi D / \log \sqrt{\psi}$ and the initial rate of concentration change $d[T^*]/dt = k_+[R^*][T] = 1.75 \times 10^{11}$/cm^2 s [9], it follows that

$$\log \sqrt{\psi} \doteq -4\pi \times 6.25/17.5 \doteq -4.5.$$

Whence, from $\psi = \pi a^2[R^*] \doteq e^{-9} \doteq 1.2 \times 10^{-4}$, $a \doteq 100$ Å. Because this is a plausible distance for interaction between the reactants, this result corroborates the hypothesis of diffusion limitation.

This a enters into the pair densities of this reaction. πa^2 is interpretable as a "cross section" for reaction. However, the length scale of this reaction, analogous to ℓ (see (5)) with $1/[T]$ for the parenthetic term, is approximately equal to 2×10^{-4} cm. Because this length is comparable to the linear dimension of a disc, the detailed shape of the latter could influence these kinetics—as would also be expected to be the case when a disc contains a single R^*.

A related analysis could be applied to the subsequent reaction, in which T^* activates cGMP phosphodiesterase, the enzyme which cleaves cGMP. This requires an appropriate treatment for the nonuniform distribution of T^*, which traces the path of R^*s. The latter may best be comprised via Monte Carlo simulations [9].

6. k_+'s. In this section, to facilitate the application of our theoretical results, we collect the rate constants derived herein, derived to be used with "concentrations" given as numbers of molecules per unit volume (or area). In practice, one may employ molar concentrations, obtained by dividing the number of molecules per unit volume (or area) by Avogadro's number $N_0 \doteq 6.02 \times 10^{23}$. Therefore, to obtain a k_+ appropriate for molar concentrations, multiply the following k_+'s by N_0. For reactions in $\mathbb{R}^2$,

$$k_+ \sim \frac{2\pi D}{\log d/a},$$

and correspondingly, in $\mathbb{R}^3$,

$$k_+ \sim \frac{4\pi D a(1 + o(\sqrt{\phi}))}{1 - a/d},$$

where ϕ denotes the expected number of A molecules plus B molecules within a sphere of radius a. For an interfacial reaction, with A's diffusing in a half space and B's (and AB's) diffusing in a plane bounding the half space,

$$k_+ \sim \frac{4a\sqrt{DD_A}}{1 - \frac{2}{\pi}\sin^{-1}\frac{a}{d}}.$$

Letting $d \to \infty$ in the differential equation corresponds to irreversible reaction because, when $a \ll d$, $\varrho(r)$ becomes uninfluenced by dissociation. The related k_+ is therefore obtainable by omitting the "source" term modeling dissociation from the equation (see section 4.4). For irreversible reactions $d[AB]/dt = -d[A]/dt = -d[B]dt = k_+[A][B]$, and, in $\mathbb{R}^2$,

$$k_+ \sim \frac{2\pi D}{-\log\sqrt{\psi}}\left(1 + \frac{\log\sqrt{-\log\sqrt{\psi}} + \frac{1}{2}\log 2 - \gamma}{\log\sqrt{\psi}} + \cdots\right),$$

recalling that $\psi = \pi a^2([A] + [B])$ ($\ll 1$), and where γ denotes Euler's constant. Recall from section 5.2 that for the catalytic reaction $A + B \to A + C$, one needs only substitute $\psi = \pi a^2[A]$ in the foregoing formula for k_+ (with an analogous substitution for spatial reactions). Correspondingly, in $\mathbb{R}^3$,

$$k_+ \sim 4\pi a D(1 + \sqrt{3\phi} + \cdots),$$

recalling that $\phi = \frac{4}{3}\pi a^3([A]+[B])$ ($\ll 1$). Agreement with previous results [3, (1.3)], [8, (121)–(122)] suggests that the independent-pair approximation for the triple density imparts negligible errors to these terms. Furthermore, for irreversible interfacial reactions,

$$k_+ \sim 4a\sqrt{DD_A}.$$

For the latter, when a protein A in solution reacts with a protein B in a cell membrane, $D_A \gg D_B$, and $D \approx D_A$, the above equation reduces to that obtained in [2], under the assumption that $D_B = 0$. Even when a protein in the cytosol, where the viscosity is 1.5 times larger than in solution, reacts with a lipid diffusing in a cell membrane, the required correction to the result of [2] will be negligible for most proteins. In detail, lipids have diffusion coefficients $\sim 10^{-8}\,\mathrm{cm}^2/\mathrm{s}$, two orders of magnitude larger than the coefficient of FCϵR1 on the cell surface. Furthermore, as the enzyme ribonuclease has a diffusion coefficient $\sim 10^{-6}\,\mathrm{cm}^2/\mathrm{s}$ in solution, only for massive proteins would our enhancement be nonnegligible.

7. Discussion. As may be surmised from section 5, the present results are practicable; they constitute extensions to the Smoluchowski theory. The applications presented above required adaptation of our theory, perhaps unsurprisingly, but they kindle the expectation that our approach can facilitate the modeling of a cornucopia of diffusion limited reactions, including irreversible reactions—exclusive of kinetic aberrations such as those attributable to reaction-induced concentration "archipelagoes" of A and B [16, 19, 20]. The forward reaction rate constant may be obtained from abridged solution of the respective differential equation, fostering the analysis of complex reactions [7].

It follows that the types of criticisms that can be leveled at the Smoluchowski theory may also apply to ours. We have sidestepped the complicated problem of estimating the error of our theory. For example, for $A+B \to AB$, although the use of the independent-pairs approximation for triple densities (1) did not, apparently, compromise the accuracy of the given results, the determination of the first decommissioned term (in the virial expansion) remains to be investigated, as do alternative reactions, such as $A + A \to A_2$, for which the independent-pairs approximation is likely to be less propitious. Thus, many challenges lie ahead.

Appendix. Consider the integral

$$\begin{aligned} I(\alpha,\beta) &= \int_R^\alpha \frac{\bar{R}d\bar{R}}{(\beta^2-\bar{R}^2)^{1/2}(\bar{R}^2-R^2)^{1/2}} \\ &= \frac{R}{2}\int_1^{\alpha^2/R^2} \frac{dx}{\sqrt{\beta^2-R^2x}\sqrt{x-1}} = \frac{R}{2}\int_1^{\alpha^2/R^2} \frac{dx}{\sqrt{-\beta^2+(\beta^2+R^2)x-R^2x^2}} \\ &= \frac{R}{2}\left(-\frac{1}{R}\right)\sin^{-1}\frac{-2R^2x+\beta^2+R^2}{\beta^2-R^2}\Bigg|_{x=1}^{x=\alpha^2/R^2} = \frac{\pi}{4}-\frac{1}{2}\sin^{-1}\frac{\beta^2-2\alpha^2+R^2}{\beta^2-R^2}, \\ &\qquad\qquad \alpha<\beta,\ \bar{R}<\alpha, \end{aligned}$$

Now

$$\frac{dI}{d\beta} = -\beta\int_R^\alpha \frac{\bar{R}d\bar{R}}{(\beta^2-\bar{R}^2)^{3/2}(\bar{R}^2-R^2)^{1/2}}$$

$$= -\frac{1}{2}\frac{1}{\sqrt{1-\left(\frac{\beta^2-2\alpha^2+R^2}{\beta^2-R^2}\right)^2}}\left\{\frac{2\beta}{\beta^2-R^2} - \frac{2\beta(\beta^2-2\alpha^2+R^2)}{(\beta^2-R^2)^2}\right\}$$

$$= -\frac{\beta}{\beta^2-R^2}\sqrt{\frac{1-\lambda}{1+\lambda}}, \qquad \lambda = \frac{\beta^2-\alpha^2-(\alpha^2-R^2)}{\beta^2-R^2},$$

$$1-\lambda = 2\frac{\alpha^2-R^2}{\beta^2-R^2}, \quad 1+\lambda = 2\frac{\beta^2-\alpha^2}{\beta^2-R^2}, \quad \sqrt{\frac{1-\lambda}{1+\lambda}} = \sqrt{\frac{\alpha^2-R^2}{\beta^2-\alpha^2}}.$$

Thus,

$$\int_R^\alpha \frac{\bar{R}d\bar{R}}{(\beta^2-\bar{R}^2)^{3/2}(\bar{R}^2-R^2)^{1/2}} = \frac{1}{\beta^2-R^2}\sqrt{\frac{1-\lambda}{1+\lambda}} = \frac{1}{\beta^2-R^2}\sqrt{\frac{\alpha^2-R^2}{\beta^2-\alpha^2}},$$

whence (38).

Acknowledgments. Professor H. M. McConnell, of Stanford University, called attention to the Smoluchowski theory and pointed out its shortcomings for reactions in surfaces. We are indebted to Professor Frits Wiegel, emeritus of Twente University, for sharing his insights.

REFERENCES

[1] E. Beltrami, *On the theory of potential functions symmetrical about an axis of revolution*, R. Accad. Sci. Bologna, Ser. IV, Vol. II, 1881, pp. 461–498.
[2] H. C. Berg and E. M. Purcell, *Physics of chemoreception*, Biophys. J., 20 (1977), pp. 193–219.
[3] R. I. Cukier, *Diffusion-influenced reactions*, J. Statist. Phys., 42 (1986), pp. 69–82.
[4] C. Delisi, *Biophysics of ligand receptor interactions*, Quart. Rev. Biophys., 13 (1980), pp. 201–223.
[5] C. Durning and B. O'Shaughnessy, *Diffusion controlled reactions at an interface*, J. Chem. Phys., 88 (1988), pp. 7717–7128.
[6] J. R. Faeder, W. S. Hlavacek, I. Reischl, M. L. Blinov, H. Metzger, A. Redondo, C. Wofsy, and B. Goldstein, *Investigation of early events in FcepsilonRI-mediated signaling using a detailed mathematical model*, J. Immunol., 170 (2003), pp. 3769–3781.
[7] A. N. Gorban and I. V. Karlin, *Invariant Manifolds for Physical and Chemical Kinetics*, Springer-Verlag, Berlin, 2005.
[8] J. Keizer, *Nonequilibrium statistical thermodynamics and the effect of diffusion on chemical reaction rates*, J. Phys. Chem., 86 (1982), pp. 5052–5067.
[9] T. D. Lamb, *Gain and kinetics of activation in the G-protein cascade of phototransduction*, in Vision: From Photon to Perception (May 20–22, 1995), National Academy of Sciences Colloquium, National Academy Press, Washington, DC, 1999, pp. 10–14.
[10] A. K. Menon, D. Holowka, W. Webb, and B. Baird, *Clustering, mobility, and triggering activity of small oligomers of immunoglobulin E on rat basophilic leukemia cells*, J. Cell Biol., 102 (1986), pp. 534–540.
[11] P. M. Morse and H. Feshbach, *Methods of Mathematical Physics, Part* I, McGraw-Hill, New York, 1953.
[12] B. Noble, *Solution of Bessel function dual integral equations by a multiplying-factor method*, Proc. Cambridge Philos. Soc., 59 (1963), pp. 351–362.
[13] R. M. Noyes, *Effects of diffusion rates on chemical kinetics*, Prog. React. Kinet., 1 (1961), pp. 129–160.
[14] I. Remy, I. A. Wilson, and S. W. Michnick, *Erythropoietin receptor activation by a ligand-induced conformation change*, Science, 283 (1999), pp. 990–993.
[15] A. R. Reynolds, C. Tischer, P. J. Verveer, O. Rocks, and P. I. H. Bastiaens, *EGFR activation coupled to inhibition of tyrosine phosphatases causes lateral signal propagation*, Nature Cell Biol., 5 (2003), pp. 447–453.
[16] M. G. Rudavets, *The role of pair correlations in diffusion limited recombination* $A + B \to 0$, Phys. Lett. A, 176 (1993), pp. 62–66.

[17] J. Schlessinger, W. W. Webb, E. L. Elson, and H. Metzger, *Lateral motion and valence of Fc receptors on rat peritoneal mast cells*, Nature, 264 (1976), pp. 550–552.
[18] C. Torigoe and H. Metzger, *Spontaneous phosphorylation of the receptor with high affinity for IgE in transfected fibroblasts*, Biochem., 13 (2001), pp. 4016–4025.
[19] D. C. Torney and T. T. Warnock, *Rates of diffusion-limited reaction in periodic systems*, Int. J. Supercomput. Appl., 1 (1987), pp. 33–43.
[20] D. Toussaint and F. Wilczek, *Particle-antiparticle annihilation in diffusive motion*, J. Chem. Phys., 78 (1983), pp. 2642–2647.
[21] T. R. Waite, *Theoretical treatment of the kinetics of diffusion-limited reactions*, Phys. Rev., 107 (1957), pp. 463–470.
[22] C. Wofsy, *private communication*.

SIAM J. APPL. MATH.
Vol. 67, No. 4, pp. 1166–1182

COMPUTATION OF EXTENSIONAL FALL OF SLENDER VISCOUS DROPS BY A ONE-DIMENSIONAL EULERIAN METHOD*

B. H. BRADSHAW-HAJEK†, Y. M. STOKES†, AND E. O. TUCK†

Abstract. We develop a one-dimensional Eulerian model suitable for analyzing the behavior of viscous fluid drops falling from rest from an upper boundary. The method allows examination of development and behavior from early time, when a drop and filament begin to form, out to large times when the bulk of the fluid forms a drop at the bottom of a long thin filament which connects it with the upper boundary. This model overcomes problems seen in Lagrangian models, caused by excessive stretching of grid elements, and enables a better examination of the thin fluid filament.

Key words. extensional flow, dripping, moving boundary, viscous flow, free surface

AMS subject classifications. 35K55, 35Q30, 35R35, 76M20, 76D08

DOI. 10.1137/050646743

1. Introduction. Formation of drops via extensional flow and break-off has been much studied (see the review article by Eggers [6]), motivated by a wide range of applications such as ink-jet printing, spinning and drawing of polymer or glass fibers, glass blowing and blow-molding in the manufacture of containers, light bulbs and glass tubing, rheological measurement by fiber extension, and fiber spinning for polymers and glasses [3, 4, 9, 13]. Considerable progress has been made towards an understanding of the breakup of a thin filament into drops, although the exact details of the final stages of breakup are yet to be resolved. However, the evolution of the drop and filament from some initial configuration, and the influence of initial conditions on the final breakup, are still relatively unexplored and have been the focus of our attention for some time [15, 14]. Some work by others on this topic includes [20, 18].

The problem of interest is a drop of very viscous fluid hanging beneath a solid wall/boundary and extending under gravity, similar to honey dripping from an upturned spoon. Analyses with and without inertia have been done and compared by the present authors [15, 14]. Surface tension was neglected in those studies, on the basis that a mean diameter $\ell = \sqrt{R_0 L_0}$ of the drop is large compared to the meniscus scale $\sqrt{\gamma/(\rho g)}$, or equivalently that the Bond number $Bo = \rho g \ell^2/\gamma$ is large. Here g is the gravitational acceleration, ρ, γ are respectively the density and surface tension coefficient of the fluid, R_0 is a length scale for the drop's cross section (e.g., the radius of the drop at the wall), and L_0 is the initial length of the drop. As the fluid filament extends and gets thinner, this neglect of surface tension may become less justifiable, and an examination of the effect of surface tension is desirable.

Because of the slender geometries involved, one-dimensional models are common in analysis of filament breakup [8, 1, 19, 5, 7, 17, 11]. However, the development of a drop and filament may also involve nonslender geometries at early times, requiring numerical solution of the full Navier–Stokes equations. Our previous work [15, 14] has involved both one-dimensional models and numerical solution of the Navier–Stokes

*Received by the editors December 5, 2005; accepted for publication (in revised form) February 5, 2007; published electronically May 29, 2007. This work was supported by Australian Research Council Discovery grant DP0450047.

http://www.siam.org/journals/siap/67-4/64674.html

†Department of Applied Mathematics, University of Adelaide, SA 5005, Australia (bronwyn.hajek@adelaide.edu.au, Yvonne.Stokes@adelaide.edu.au, ernie.tuck@adelaide.edu.au).

equations for axisymmetric drops and two-dimensional sheets.

For all of our work, a Lagrangian reference frame has been used, with grids that move with moving fluid elements. However, as the filament thins and surface tension potentially becomes important, Lagrangian numerics begin to fail due to the stretching of the grid. For example, in finite-element simulations of the full Navier–Stokes equations [14], mesh elements in and near the filament region become excessively elongated or distorted, leading to loss of accuracy. Similarly, in one-dimensional models the grid points become sparse in the filament region while congregating in the main drop, so that we lose the ability to examine the development of the filament. Hence, if we are to better investigate the filament evolution, including possible effects of surface tension, we must modify our methods.

A number of techniques are available to address the resolution problem in the filament region. First, we can begin with an irregular mesh that has a concentration of grid points in the section of the drop that will develop into the filament region. This, however, will only be successful until that section of the mesh becomes very stretched. Another option is to remesh when grid points become too sparse in the filament. This method becomes difficult (although not impossible) with the inclusion of inertia (see, for example, [18]), as all unknowns and their time derivatives must be interpolated from the old mesh onto the new mesh. In this paper, we wish to present a further alternative in which the congregation of mesh points does not occur.

We have therefore developed a one-dimensional model in an Eulerian reference frame, where the Lagrangian coordinate (a fluid particle label equal to the initial distance ξ from the wall) is sought as a function of time t and that particle's physical distance x from the wall. This model may be derived directly from the Navier–Stokes and continuity equations, as described below. It may also be obtained (in the absence of surface tension) by a transformation of our previous one-dimensional Lagrangian model [14] for the cross-sectional area A as a function of time t and Lagrangian coordinate ξ, which will also be outlined here.

The resulting PDE for $\xi = Z(x,t)$ is formally of higher order in space than the original PDE for $A(\xi,t)$. Also, while the original problem could be solved in a fixed spatial domain $0 < \xi < L_0$, where L_0 is the initial drop length, the transformation results in a moving boundary problem in the domain $0 < x < L(t)$, where the actual drop length $L(t)$ must be determined as part of the problem. Both of these aspects mean that the problem in physical coordinates is considerably harder to solve than that in Lagrangian coordinates, but it has the major benefit that grid elements do not become stretched over time and is therefore worth pursuing in order to better understand the filament behavior.

The increased complexity of the problem is partly a result of the transformation employed, with a further element of difficulty added by the inclusion of surface tension. In the absence of surface tension, the equations may be directly integrated, simplifying the numerical problem. In this paper we explore the new model and its solution in the absence of surface tension, which will be considered in a future paper. We will, however, derive here the equations with surface tension included.

2. A one-dimensional Eulerian model. For an axisymmetric column of incompressible fluid, a one-dimensional lubrication approximation to the Navier–Stokes equations yields (see, for example, [5, 11, 10])

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = g - \frac{\gamma}{\rho}\frac{\partial K}{\partial x} + \frac{\nu^*}{h^2}\frac{\partial}{\partial x}\left(h^2\frac{\partial u}{\partial x}\right), \tag{2.1}$$

while the continuity equation becomes

$$(h^2)_t + (uh^2)_x = 0, \tag{2.2}$$

where subscripts denote derivatives, $\nu_* = 3\mu/\rho$ is the elongational (Trouton) kinematic viscosity [16] of a fluid with shear viscosity μ and density ρ, g is gravitational acceleration in the downward (positive) direction, γ is the coefficient of surface tension, $u(x,t)$ is the downward velocity of the fluid at position x and time t, $h(x,t)$ is the radius of the drop, and $K(x,t)$ is the curvature of the drop, given by

$$K = \frac{1}{\sqrt{1+(h_x)^2}}\left[\frac{1}{h} - \frac{h_{xx}}{1+(h_x)^2}\right]. \tag{2.3}$$

The cross-sectional area of the drop is given by $A = \pi h^2$, so (2.2) can be rewritten as

$$A_t + uA_x = -Au_x \tag{2.4}$$

and substituted into (2.1) to obtain

$$u_t + uu_x = g - \frac{\gamma}{\rho}K_x - \frac{\nu^*}{A}\frac{\partial}{\partial x}\left(\frac{\partial A}{\partial t} + u\frac{\partial A}{\partial x}\right)$$

or

$$\frac{Du}{Dt} = g - \frac{\gamma}{\rho}K_x - \frac{\nu^*}{A}\frac{\partial}{\partial x}\frac{DA}{Dt}, \tag{2.5}$$

where $D/Dt = \partial/\partial t + u\,\partial/\partial x$ denotes the material time derivative.

In a Lagrangian reference frame [15, 14, 19] we let $x = X(\xi,t)$, where ξ is a fluid-particle label such that $x = \xi$ at $t = 0$. The initial drop geometry is assumed to have a cross-sectional area distribution given by some function $A_0(\xi)$. That is, $A(\xi,0) = A_0(\xi)$, $0 \le \xi \le L_0$, where $A(\xi,t)$ is the cross-sectional area at label ξ and time t, and L_0 is the initial drop length. Conservation of mass demands [14]

$$A\frac{\partial X}{\partial \xi} = A_0$$

or, on integration,

$$X(\xi,t) = \int_0^\xi \frac{A_0(\xi_1)}{A(\xi_1,t)}d\xi_1. \tag{2.6}$$

Now, defining $\xi = Z(x,t)$, we have

$$A = A_0 Z_x, \quad u = X_t = -\frac{Z_t}{Z_x}, \quad \text{and} \quad \frac{\partial A_0}{\partial x} = A_0' Z_x,$$

where primes denote differentiation with respect to ξ. Substituting for A and u in (2.5) gives

$$\begin{aligned} -\frac{D}{Dt}\left(\frac{Z_t}{Z_x}\right) &= g - \frac{\gamma}{\rho}K_x - \frac{\nu^*}{A_0 Z_x}\frac{\partial}{\partial x}\left[A_0\frac{D}{Dt}(Z_x)\right] \\ &= g - \frac{\gamma}{\rho}K_x - \nu^*\left[\frac{A_0'}{A_0}\frac{D}{Dt}(Z_x) + \frac{1}{Z_x}\frac{\partial}{\partial x}\left(\frac{D}{Dt}(Z_x)\right)\right] \\ &= g - \frac{\gamma}{\rho}K_x - \nu^*\frac{D}{Dt}\left(\frac{A_0'}{A_0}Z_x + \frac{Z_{xx}}{Z_x}\right), \end{aligned} \tag{2.7}$$

with

$$\frac{D}{Dt} = \frac{\partial}{\partial t} - \frac{Z_t}{Z_x}\frac{\partial}{\partial x}.$$

The transformation to the dependent variable $Z(x,t)$ has yielded a PDE (2.7) that is second order in time and third order in space, whereas the Navier–Stokes equation (2.1) is first order in time and second order in space. The main reason for this increase in order is that the dependant variable is a position rather than a velocity (as in the Navier–Stokes equations). In order to solve (2.7), we must also solve for the length of the drop $L(t) = X(L_0, t)$, which is increasing with time. Thus, we need two initial conditions and four boundary conditions. One initial condition is obtained from the definition of the Lagrangian coordinate such that $\xi = x$ at $t = 0$, so that

$$Z(x,0) = x. \tag{2.8}$$

The other comes from the condition that the flow starts from rest, so $u(x,0) = 0$ or

$$Z_t(x,0) = 0. \tag{2.9}$$

With respect to boundary conditions, two (one at each end) come from the definition of the Lagrangian coordinate such that $x = 0$ at $\xi = 0$ and $x = L(t)$ at $\xi = L_0$, giving

$$Z(0,t) = 0, \tag{2.10}$$

$$Z(L(t),t) = L_0. \tag{2.11}$$

Since the drop is falling from under a solid plane boundary where the normal velocity is zero for all time, i.e., $u = 0$ at $x = 0$, then $Du/Dt = 0$ at $x = 0$, and hence, from (2.7),

$$0 = g - \frac{\gamma}{\rho}K_x - \nu^* \frac{D}{Dt}\left(\frac{A_0'}{A_0}Z_x + \frac{Z_{xx}}{Z_x}\right) \quad \text{at} \quad x = 0. \tag{2.12}$$

We require a further boundary condition which comes from a balance between viscous stresses and surface tension at the bottom of the drop $x = L(t)$. One-dimensional theory yields

$$\frac{\partial}{\partial x}\left(\frac{Z_t}{Z_x}\right) = -\frac{\gamma}{\rho\nu^*}K, \quad \text{or, equivalently,} \quad \frac{D}{Dt}(Z_x) = -\frac{\gamma}{\rho\nu^*}Z_x K. \tag{2.13}$$

Equation (2.7) subject to initial and boundary conditions (2.8)–(2.13) describes the fall of a drop of viscous fluid from underneath a solid boundary, starting from a known initial configuration. Gravitational, viscous, inertial, and surface tension effects are all included. The model derived involves the fluid-particle label $\xi = Z(x,t)$ as the dependent variable, with the physical space coordinate x and time t as the independent variables.

For zero surface tension ($\gamma = 0$), (2.7) simplifies considerably, by integration with respect to the material time derivative. With nonzero surface tension ($\gamma \neq 0$), such a procedure is computationally problematic due to the necessity of time-integrating the surface-tension term while holding the particle label $\xi = Z(x,t)$ constant. We leave consideration of this matter to a future paper and, from here on, neglect surface

tension (i.e., set $\gamma = 0$). Integration with respect to t at fixed $\xi = Z$, subject to $Z = x$ and $Z_t = 0$ at $t = 0$, then yields

$$-\frac{Z_t}{Z_x} = gt - \nu^* \left(\frac{A_0'}{A_0}(Z_x - 1) + \frac{Z_{xx}}{Z_x} \right)$$

or

$$Z_t = \nu^* Z_{xx} - gtZ_x - \nu^* \frac{A_0'(Z)}{A_0(Z)} Z_x(1 - Z_x). \tag{2.14}$$

Equation (2.14) is in general a nonlinear PDE which, like the Navier–Stokes equation (2.1), is first order in time and second order in space. It is worth noting in passing that in the special case of an initially cylindrical drop where $A_0 =$ constant, it becomes linear, and in the further special case where gravity can be neglected (such as in a liquid bridge problem [2]), it reduces to the ordinary linear heat-conduction equation, with diffusivity ν^*.

The appropriate initial and boundary conditions are

$$\begin{aligned} Z(x,0) &= x \qquad \text{at } t = 0, \\ Z(0,t) &= 0 \qquad \text{at } x = 0, \\ Z(L(t),t) &= L_0 \qquad \text{at } x = L(t), \\ Z_x(L(t),t) &= 1 \qquad \text{at } x = L(t). \end{aligned} \tag{2.15}$$

Note that we no longer need boundary condition (2.12), which in integrated form is equivalent to $u = 0$ at $x = 0$, and which is automatically satisfied by demanding $Z(0,t) = 0$. Also, with $\gamma = 0$, (2.13) can be integrated with respect to the material time derivative to give $Z_x = 1$ at $x = L(t)$ as we have in (2.15).

The Lagrangian equivalent to (2.14) in terms of the cross-sectional area $A(\xi, t)$ as a function of Lagrangian coordinate ξ and time t is readily (by manipulation of (2.14)) shown to be

$$u = gt - \frac{\nu^*}{A_0} \frac{\partial}{\partial \xi}(A - A_0).$$

Differentiating with respect to ξ, using (2.6), and rearranging gives

$$\frac{\partial A}{\partial t} = \nu_* \frac{A^2}{A_0} \frac{\partial}{\partial \xi} \left(\frac{1}{A_0} \frac{\partial}{\partial \xi}(A - A_0) \right), \quad 0 \le \xi \le L_0. \tag{2.16}$$

The corresponding initial and boundary conditions are

$$A(\xi,0) = A_0(\xi), \quad \frac{\partial}{\partial \xi}(A - A_0)(0,t) = \frac{gt}{\nu^*} A_0(0), \quad A(L_0,t) = A_0(L_0). \tag{2.17}$$

The Lagrangian model given by (2.16) and (2.17) was derived directly in [14] by balancing viscous and gravitational forces; the inertialess version was considered in [15]. Comparison between solutions to these models and those for the new Eulerian model of present interest, (2.14) and (2.15), will be given below. We note that the Eulerian model involves gravity explicitly in the PDE (2.14), whereas the Lagrangian model involves gravity only in a boundary condition at $\xi = 0$ (2.17).

3. Eulerian-model solution. For the remainder of this paper, we will be primarily interested in initially paraboloidal slender drops, given in Lagrangian coordinates by $A_0(\xi) = A_0(0)(1 - \xi/L_0)$ with small aspect ratio $\alpha_r = \sqrt{A_0(0)}/L_0 \ll 1$, as considered in [14].

Defining dimensionless variables (denoted by bars)

$$\overline{A}_0(\overline{\xi}) = \frac{A_0(\xi)}{A_0(0)}, \quad \overline{\xi} = \overline{Z} = \frac{\xi}{L_0} = \frac{Z}{L_0}, \quad \overline{x} = \frac{x}{L_0}, \quad \overline{t} = \frac{gL_0}{\nu^*}t, \tag{3.1}$$

the dimensionless form of (2.14) for the initially paraboloidal drop $\overline{A}_0(\overline{Z}) = 1 - \overline{Z}$ is (after removing the bars)

$$Re\, Z_t = Z_{xx} - tZ_x + \frac{Z_x}{1 - Z}[1 - Z_x], \tag{3.2}$$

with the Reynolds number Re given by

$$Re = \frac{gL_0^3}{\nu^{*2}}.$$

The initial and boundary conditions (2.15) become

$$Z(x,0) = x, \quad Z(0,t) = 0, \quad Z(L(t),t) = 1, \quad \text{and} \quad Z_x(L(t),t) = 1. \tag{3.3}$$

Equation (3.2) subject to (3.3) is most easily solved using the explicit forward-time-centered-space finite difference method. Setting the time step Δt and spatial step Δx, we approximate (3.2) in the usual manner by

$$\begin{aligned} Re\frac{Z_i^{j+1} - Z_i^j}{\Delta t} &= \frac{Z_{i+1}^j - 2Z_i^j + Z_{i-1}^j}{\Delta x^2} - t\frac{Z_{i+1}^j - Z_{i-1}^j}{2\Delta x} \\ &\quad + \frac{1}{1 - Z_i^j}\frac{Z_{i+1}^j - Z_{i-1}^j}{2\Delta x}\left[1 - \frac{Z_{i+1}^j - Z_{i-1}^j}{2\Delta x}\right], \end{aligned} \tag{3.4}$$

where $Z_i^j = Z(x_i, t_j)$ is the value of $Z(x,t)$ at the jth time step and the ith grid point. For numerical stability, we must ensure that the diffusion number $\Delta t/Re(\Delta x)^2 < 0.5$.

The initial and wall boundary conditions are easily specified by setting $Z_i^0 = i\Delta x$ and $Z_0^j = 0$. However, the boundary conditions at the free end are not quite so straightforward to implement, due to the moving boundary. At each time step, the drop becomes longer and some of the drop (at the bottom) will move beyond the current computational domain. Hence we need to extend the grid to the new position of the bottom of the drop.

Specifically, having computed Z_i^{j+1}, $i = 1, \ldots, N_j - 1$, using (3.4), we seek an extrapolation procedure that approximates the bottom of the drop, satisfying the boundary conditions $Z = Z_x = 1$ at the (as yet unknown) drop bottom $x = L(t_{j+1})$, and that matches our already computed solution above the bottom. The most obvious choice is a linear polynomial extrapolation, but this can only be first order accurate, and we prefer to preserve the second order accuracy of the finite difference scheme. To achieve this, we could seek a quadratic polynomial extrapolation

$$Z(x, t_{j+1}) = a_{j+1}(x - x_{N_j - 1})^2 + b_{j+1}(x - x_{N_j - 1}) + c_{j+1},$$

where the unknown coefficients $a_{j+1}, b_{j+1}, c_{j+1}$ and the unknown length $L(t_{j+1})$ of the drop are determined by satisfying the two boundary conditions at $x = L(t)$ and

matching the already computed solution at x_{N_j-2} and x_{N_j-1}. An equally good second order extrapolation is to use the exponential approximation

$$Z(x,t_{j+1}) = e^{x-L(t_{j+1})}, \quad \text{where} \quad e^{-L(t_{j+1})} = Z_{N_j-1}^{j+1} e^{-x_{N_j-1}}. \tag{3.5}$$

In fact, (3.5) is just the local form of the solution from the corresponding Lagrangian model neglecting inertia as in [15], and hence has a stronger physical motivation than the quadratic extrapolant.

Earlier work [15, 14] has shown that, with neglect of surface tension, the drop shape very near to the bottom is given quite accurately by the inertialess solution. This is because at early times, accelerations are very small and Stokes flow solutions are applicable; at later times, the main drop is essentially in free fall and (with neglect of surface tension) does not change in shape. Furthermore, for some initial configurations, including the initially paraboloidal drop considered here, we can obtain an exact analytic solution to the Lagrangian model neglecting inertia and use this to assign appropriate values of $Z(x,t)$ to the new grid points. Specifically, using (3.1), the dimensionless form of the Lagrangian PDE (2.16) is (after removing the bars)

$$Re\frac{\partial A}{\partial t} = \frac{A^2}{A_0}\frac{\partial}{\partial \xi}\left(\frac{1}{A_0}\frac{\partial}{\partial \xi}(A - A_0)\right), \quad 0 \le \xi \le 1, \tag{3.6}$$

with initial and boundary conditions

$$A(\xi,0) = A_0(\xi), \quad \frac{\partial}{\partial \xi}(A - A_0)(0,t) = t, \quad A(1,t) = A_0(1,t). \tag{3.7}$$

In the inertialess limit ($Re = 0$) this has the explicit solution

$$A(\xi,t) = A_0(\xi) - tV(\xi), \quad V(\xi) = \int_\xi^1 A_0(\xi_1)\, \mathrm{d}\xi_1. \tag{3.8}$$

As discussed by Stokes, Tuck, and Schwartz [15], the cross-sectional area of the drop vanishes at the position $\xi = \xi_*$ such that $t = t_* = A_0(\xi_*)/V(\xi_*)$ is a minimum, so that the drop formally breaks with $A(\xi_*, t_*) = 0$. The time t_* is the "crisis" time; at this time the length of the drop, given by (2.6) with $\xi = 1$, formally becomes infinite in this inertialess approximation. No solution exists for $t > t_*$; i.e., we have a finite-time blow up at the crisis time t_*. However, for larger times $t > t_*$ the main drop is effectively falling as a solid body, and in the absence of surface tension it retains the same shape given by (3.8) with $t = t_*$.

For the initially paraboloidal drop $A_0(\xi) = 1 - \xi$, (3.8) becomes

$$A(\xi,t) = (1-\xi)\left(1 - \frac{1}{2}t(1-\xi)\right), \tag{3.9}$$

from which we see that $\xi_* = 0$ and $t_* = 2$; i.e., the drop breaks at the wall at the crisis time $t = 2$. Hence, for all $t \ge 2$,

$$A(\xi) = \xi(1-\xi), \tag{3.10}$$

which is a solution to (3.6). It is readily verified using $Z_x = A/A_0$ that (3.9) satisfies the condition $Z_x = 1$ at $x = L(t)$ (i.e., $\xi = 1$) for all $t \le 2$.

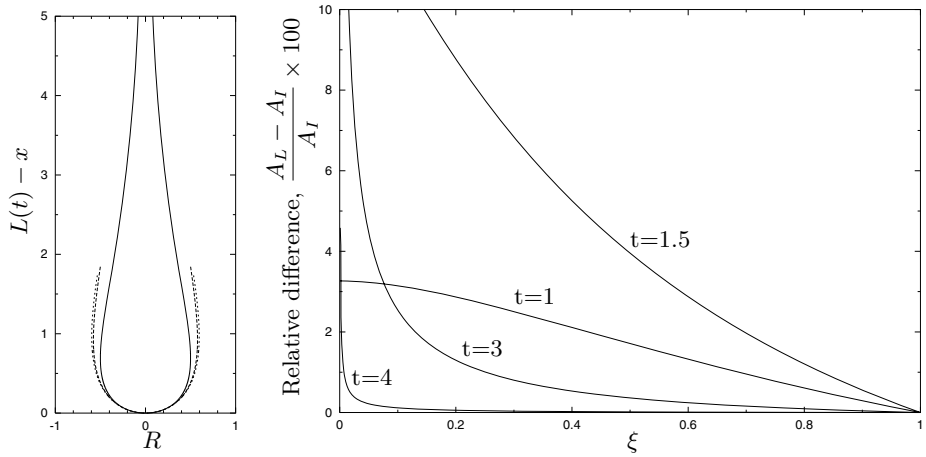


FIG. 3.1. (a) *Drop shape as a function of distance* $L(t) - x$ *from the bottom of the drop. Inertialess solution* (3.9) *at* $t = 1.5$ *(solid); solution to the Lagrangian PDE* (3.6) *at* $t = 1.5$ *(dotted) and* $t = 4.0$ *(dashed); the inertialess large-time solution* (3.10) *is indistinguishable from the dashed curve.* (b) *Percentage relative difference between the solution to the Lagrangian PDE* (3.6) *and the inertialess solution* (3.9) *for* $t < 2$ *or* (3.10) *for* $t \geq 2$. *The percentage relative difference is calculated as* $100 \times (A_L - A_I)/A_I$, *where* A_L *is the calculated solution to* (3.6) *and* A_I *is the inertialess solution* (3.9) *or* (3.10). *Here* (3.6) *was solved using implicit backward differencing with* $Re = 0.1$, $\Delta\xi = 10^{-3}$, $\Delta t = 10^{-3}$.

The extrapolant for $Z(x,t)$ is obtained by substituting $A_0 = 1 - Z$, and $A(Z,t)$ given by (3.9) for $t < 2$ or (3.10) for $t \geq 2$, into $Z_x = A/A_0$. Integrating then yields

$$Z(x,t) = \begin{cases} 1 - \dfrac{2}{t} + c(t)\,\mathrm{e}^{xt/2} & \text{for } \ 0 < t < 2, \\ c(t)\,\mathrm{e}^{x} & \text{for } \ t \geq 2. \end{cases} \tag{3.11}$$

We solve for the value of the unknown function of time $c(t)$ at time t_{j+1} using the already computed value of Z at x_{N_j-1}. The expression so obtained for $t \geq 2$ is (3.5) exactly. It is also readily seen that the expression obtained for $t < 2$ is second-order accurate. Thus, we use (3.11) to calculate values of $Z^{j+1}_{N_j+k}$, $k = 0, 1, \ldots$, stopping when $Z^{j+1}_{N_j+k} > 1$, and thus extending the computational domain to N_{j+1} grid points. The actual position of the bottom of the drop is given by solving $Z(L(t_{j+1}), t_{j+1}) = 1$.

The accuracy of this procedure is demonstrated for $Re = 0.1$ in Figure 3.1. Figure 3.1(a) compares the drop shape at $t = 1.5$ given by the inertialess solution (3.9) and as found by solving (3.6); also shown is the large- (i.e., crisis) time inertialess solution (3.10), which is indistinguishable from the solution to (3.6) at $Re = 0.1$, $t = 4$. Figure 3.1(b) shows the percentage relative difference between solutions to (3.6) at Reynolds number $Re = 0.1$ and the inertialess solution at different times. At the very bottom of the drop, the relative difference is much less than 1%. As the Reynolds number increases, the inertialess solution becomes less accurate as a global approximation for the drop shape, but each (dimensionless) time step represents a decreasing physical time interval so that still only very few grid points are extrapolated. Even for Reynolds numbers as high as $Re = 10$, it gives a good approximation for the local region near the bottom of the drop. Thus the method remains (second-order) accurate. In fact, because only a very few grid points are ever extrapolated, the choice of extrapolation procedure has only a minor effect on the solution.

Having determined the Lagrangian coordinate $Z(x,t)$ over the new, extended, computational domain, the actual shape of the drop can be calculated via $R = \sqrt{A} = \sqrt{A_0 Z_x}$.

4. Results and comparison between Eulerian and Lagrangian models. The numerical solution to (3.2) for the particle label $Z(x,t)$ as a function of physical space and time is shown, for Reynolds number $Re = 0.1$, in Figure 4.1. The growth of the computational domain as a result of the moving boundary at $Z = 1$ can be clearly seen.

The axisymmetric drop shape is shown in Figure 4.2(a), alongside drop shapes from the numerical solution to the Lagrangian equation (3.6), in Figure 4.2(b). The solution to (3.6) was calculated using the implicit backward-time-centered-space finite difference method. The two different models produce the same drop shapes with the same overall length; however, there are some differences to be highlighted.

First, for times $t \gtrsim 2.6$ the computed solutions to the Lagrangian model (Figure 4.2(b)) appear to move away from the wall. This is due to stretching of the grid and a consequent loss of grid points in the filament region and accumulation of grid points in the main drop below the filament, as seen in Figure 4.3(b). The fluid particle that is initially a distance $x = \Delta\xi$ from the wall (i.e., the closest point to the wall for which we calculate $A(\xi,t)$) falls ever downwards, so that there is a continually lengthening region in physical space, which is essentially the fluid filament connecting the drop to the wall, about which we know virtually nothing. Unfortunately, it is in this filament region that our greatest interest lies, since this is where the drop will eventually break. While decreasing the grid spacing near the wall will extend the time over which we have near complete information, there will always come a time (soon after the crisis time t_* of the inertialess theory, when accelerations approach gravitational acceleration) when the grid becomes too stretched in the filament region. This loss of information in the filament region is completely overcome with the Eulerian model (Figure 4.2) because gridpoints are fixed in space and the grid constantly extended as the drop length increases. This leads to a uniform spacing of grid points over the full

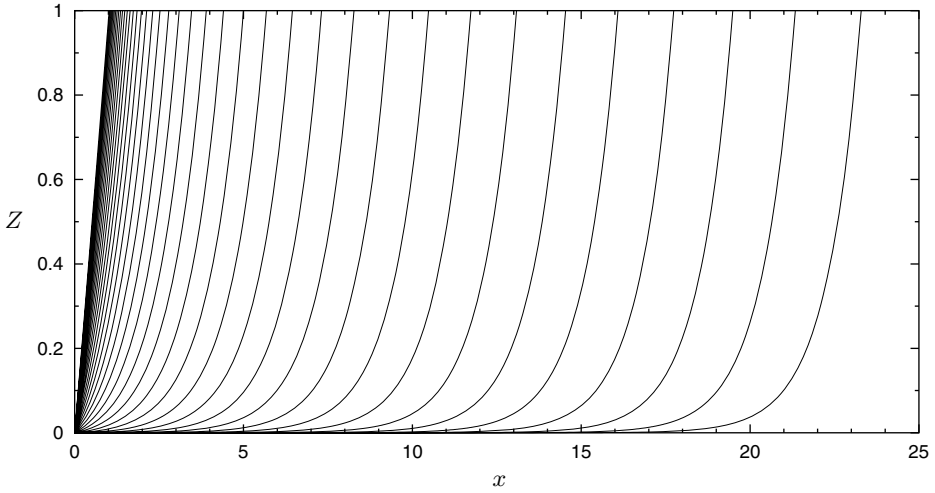


FIG. 4.1. *Solution to PDE* (3.2) *for* $Z(x,t)$*, with* $Re = 0.1$*, at times* $t = 0, 0.1, \ldots, 3.9, 4.0$*. Here* $\Delta x = 10^{-2}$*,* $\Delta t = 4 \times 10^{-6}$*.*

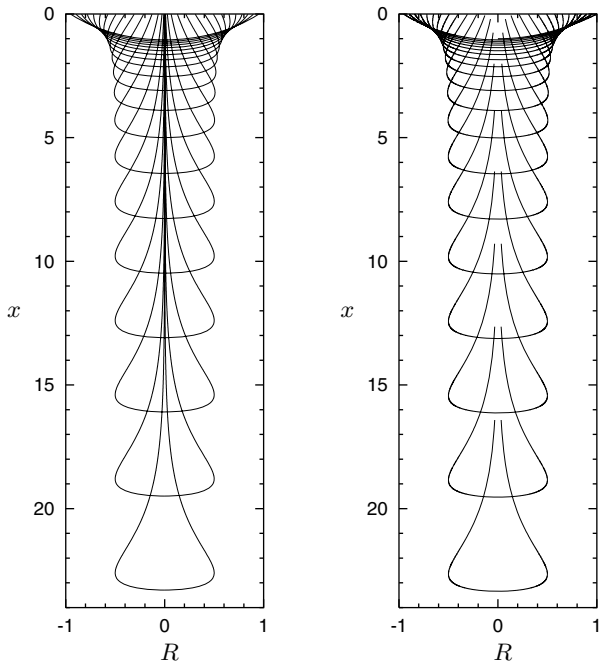


FIG. 4.2. *Drop shapes for* $Re = 0.1$ *at times* $t = 0, 0.2, \ldots, 3.8, 4.0$. (a) *Shape calculated using Eulerian framework* (3.2) *(*$\Delta x = 10^{-2}$, $\Delta t = 4. \times 10^{-6}$*).* (b) *Shape calculated using Lagrangian framework* (3.6) *(*$\Delta\xi = 10^{-3}$, $\Delta t = 10^{-3}$*).*

length of the drop, as seen in Figure 4.3(a). The greater knowledge of the filament region that results from the Eulerian model will better enable a future study of the effect of surface tension on filament breakup and drop pinch-off.

A second point of difference between the Eulerian and Lagrangian models is with respect to the behavior near the wall boundary at $x = \xi = 0$. At this boundary, the Lagrangian boundary condition (3.7) for the initially paraboloidal drop is

$$\frac{\partial A}{\partial \xi}(0,t) + 1 = t. \tag{4.1}$$

For the Lagrangian model it is a simple matter to check that this boundary condition is indeed satisfied, by computing $A_\xi(0,t)$ using the forward-space finite difference formula, i.e.,

$$A_\xi(0,t) = \frac{A(\Delta\xi, t) - A(0,t)}{\Delta\xi}.$$

The value of $A_\xi(0,t) + 1$ so computed is plotted against time t in Figure 4.4 (solid curve). The wall boundary condition is satisfied until $t \approx 2.0$ and then $A_\xi(i,t)$ begins to move away from t. At $t \approx 2.4$ there is a rapid deviation from the correct solution as the value of $A_\xi(0,t)$ decreases and appears to approach a constant unit value. This highlights the fact that the Lagrangian solution cannot be relied upon at large times when the grid becomes excessively stretched in physical space.

The equivalent condition on $A_\xi(0,t)$ for the Eulerian model is obtained by differ-

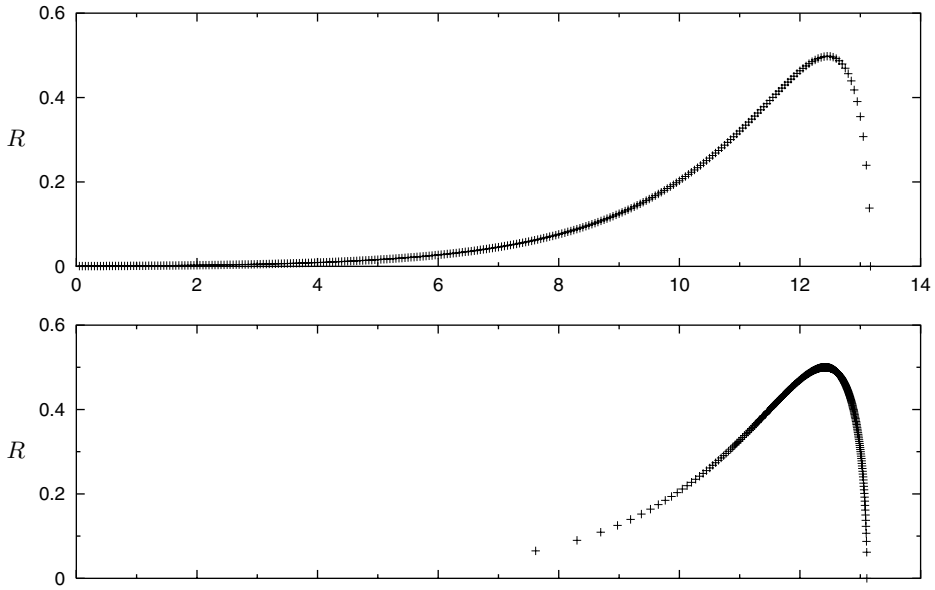


FIG. 4.3. *Comparison of Lagrangian and Eulerian solution methods at* $t = 3.4$. *Each figure has approximately* 260 *grid points.* (a) *Drop shape calculated using the Eulerian model* (3.2), (3.3) *(*$\Delta x = 5 \times 10^{-2}$, $\Delta t = 10^{-4}$*).* (b) *Drop shape calculated using the Lagrangian model* (3.6), (3.7) *(*$\Delta x = 1/260$, $\Delta t = 10^{-3}$*). The extra grid points in the filament region of the Eulerian model can be clearly seen.*

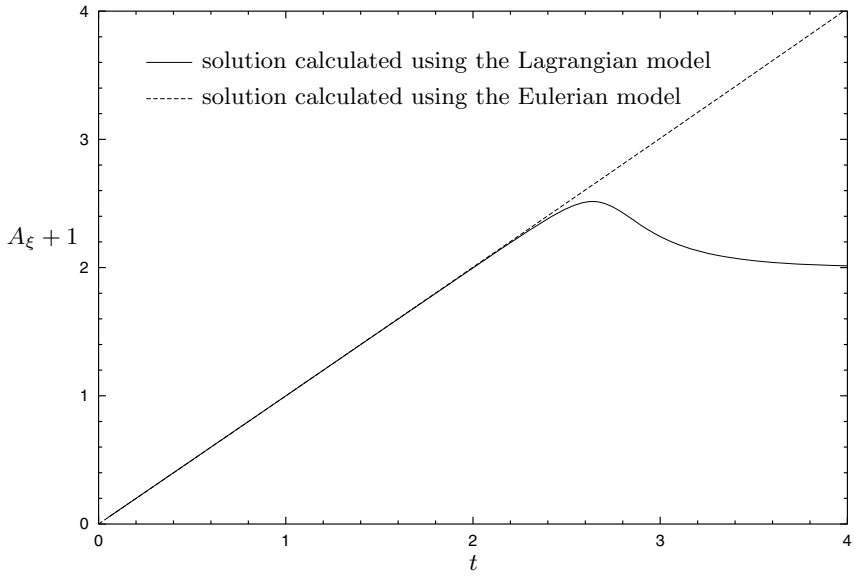


FIG. 4.4. *The accuracy of the Lagrangian and Eulerian models as indicated by the wall boundary condition* $(3.7)_2$. *For an initially paraboloidal drop we require* $A_\xi + 1 \sim t$. *This condition is satisfied by the Eulerian model but not the Lagrangian model.*

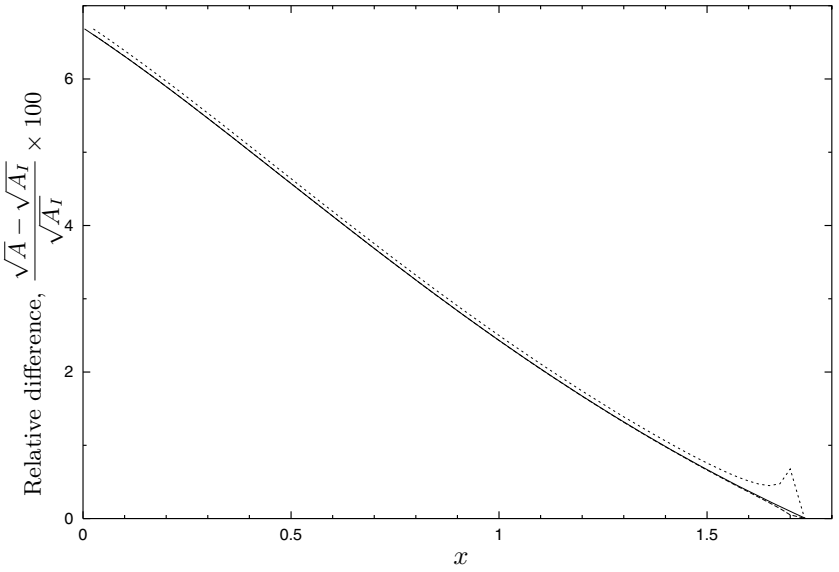


FIG. 4.5. *Percentage relative differences at $t = 1.5$ between the inertialess radius $\sqrt{A_I}$ with A_I given by (3.9), and Lagrangian and Eulerian solutions with $Re = 0.1$. The difference is given by $100 \times (\sqrt{A} - \sqrt{A_I})/\sqrt{A_I}$, where A denotes the Lagrangian solution (solid), the Eulerian solution with extension of the computational domain using the inertialess solution at crisis time (dashed), and the Eulerian solution with extension of the computational domain using the forward-difference representation of $Z_x(L(t), t) = 1$ (dotted).*

entiating $A = A_0 Z_x$ with respect to ξ, i.e. (for the initially paraboloidal drop),

$$\begin{aligned} A_\xi &= A_0' Z_x + \frac{A_0 Z_{xx}}{Z_x} \\ &= -Z_x + (1 - Z)\frac{Z_{xx}}{Z_x}. \end{aligned}$$

The slope, A_ξ at the wall can thus be found from the calculated values of $Z(0, t)$, $Z(\Delta x, t)$, and $Z(2\Delta x, t)$ using first order forward-space finite difference formulae for Z_x and Z_{xx}, i.e.,

$$Z_x(0, t) = \frac{Z(\Delta x, t) - Z(0, t)}{\Delta x} \quad \text{and} \quad Z_{xx}(0, t) = \frac{Z(2\Delta x, t) - 2Z(\Delta x, t) + Z(0, t)}{\Delta x^2}.$$

Figure 4.4 (dashed line) shows that $A_\xi + 1 \sim t$ for times well beyond $t = 2.4$; i.e., the wall boundary condition (4.1) is satisfied. Thus, we see that, for large times, the solution obtained from the Eulerian model is more reliable than that obtained from the Lagrangian model, especially in the filament region.

This is also shown by Figures 4.5 and 4.6. Figure 4.5 shows the relative differences between the inertialess prediction (3.9) of the drop/filament radius ($R = \sqrt{A}$) and solutions at $Re = 0.1$ to the Lagrangian and Eulerian models, as a function of physical distance x from the upper wall boundary, at time $t = 1.5$ before the crisis time of inertialess theory. There is excellent agreement between the Lagrangian and Eulerian models at this time, with a difference visible only at the very bottom of the drop. Figure 4.6 gives the same comparisons but at time $t = 4$, well after the crisis time of inertialess theory. Now we see considerably more difference between the Lagrangian

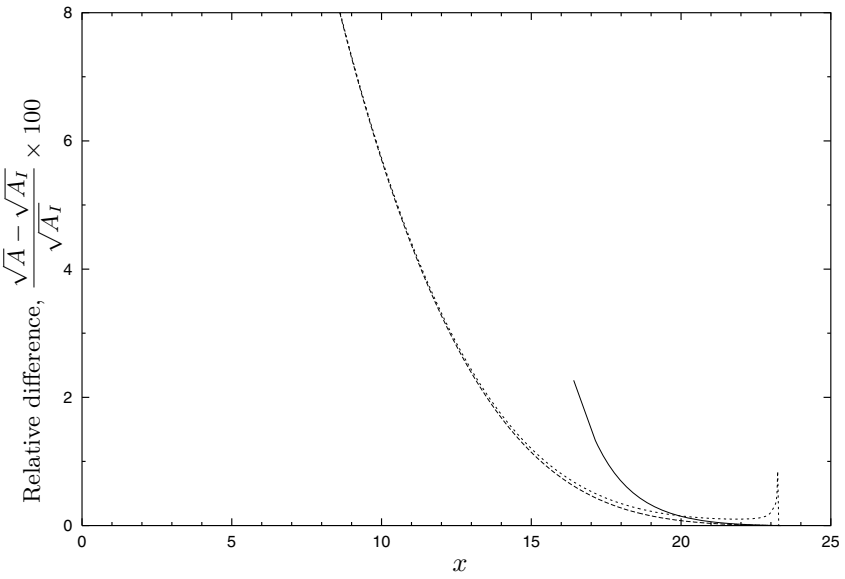


FIG. 4.6. *As for Figure* 4.5 *but at time* $t = 4$*, with* A_I *given by the inertialess large-time solution* (3.10).

and Eulerian solutions, which is due to error in the Lagrangian solution resulting from an excessively stretched grid. Note also that grid stretching limits our comparison to the bottom third of the drop where the Lagrangian solution is available; in the region $0 \leq x < 16$ no information is available from the Lagrangian solution due to a lack of grid points.

For interest, Figures 4.5 and 4.6 also show results for the Eulerian solution obtained using the finite difference approximation to the boundary condition $Z_x = 1$ at $x = L(t)$, discussed earlier as an alternative to pasting of the inertialess solution to the bottom of the drop. This differs from the other curves by only about 0.1% over most of the drop length, with the difference increasing to about 1% at the very bottom; note that the overall drop length is slightly less, as mentioned earlier, although it is not noticable with the grid size used here or at the scales shown.

It is interesting to note that at small Reynolds number the time at which the numerical solution to (3.6) begins to become inaccurate in the filament region (as indicated by Figure 4.4) is approximately equal to the crisis time of inertialess theory, as predicted in [15] ($t_* = 2$), when accelerations increase rapidly up to gravitational acceleration. This correlation between the inertialess crisis time and the time at which the small-Reynolds-number Lagrangian solution becomes inaccurate is also observed with other initial drop shapes.

5. Discussion and conclusions. The major benefit of reformulating the extensional flow problem using an Eulerian framework is that, in contrast to other one-dimensional Lagrangian models, we now include many grid points in the filament region. The Eulerian scheme is computationally more costly, as there is a rapid increase in the number of grid points as the computational domain extends with the falling drop. However, this method provides us with information about the filament region which we cannot obtain using a Lagrangian method. Accuracies such as those achieved in Figure 4.2 can still be obtained in a matter of minutes. This greater res-

olution in the filament region enables us to better study the dynamics and behavior of the developing filament. In particular we are now much better equipped to investigate the effects of surface tension on the filament, the drop shape, and pinch-off of the main drop by solving (2.7) with $\gamma \neq 0$. This will be considered in a future paper.

Meanwhile, reformulating the problem also enables us to address a question previously posed in Stokes and Tuck [14]. In that paper, we saw that at small Reynolds numbers and large times, the main part of the drop is indistinguishable from a solid object that fell from rest at an apparent time t_0. Identification of this apparent time with the crisis time of inertialess theory leads to the conclusion that the large-time drop shape is the drop shape obtained at the crisis time when neglecting inertia. Conversely, it can be shown that equating the large-time drop shape at small Reynolds numbers with the drop shape at the crisis time of inertialess theory, which is strongly supported by the numerical solutions (both here and in [14]), implies that the apparent time t_0 and the crisis time t_* are identical. This relationship between the inertialess theory and the large-time limit of the flow with inertia implies the expected large-time shape for an initially paraboloidal drop [14]

$$A(x,t) = \mathrm{e}^{-(L-x)}\left[1 - \mathrm{e}^{-(L-x)}\right], \tag{5.1}$$

where $L = L(t)$ is the length of the drop at time t. However, the asymptotic theory described in [14] did not provide an estimate of the actual length $L(t)$ of the drop. We can now supply that estimate.

In the physical coordinate system, the cross-sectional area of the drop is given by $A(x,t) = A_0 Z_x$. The expected large-time drop shape obtained from the inertialess theory, for the initially paraboloidal drop, is given by $(3.11)_2$ as $Z(x,t) = c(t)\mathrm{e}^x$, so that

$$\begin{aligned} A(x,t) &= (1 - Z)Z_x \\ &= (1 - c(t)\,\mathrm{e}^x)c(t)\,\mathrm{e}^x. \end{aligned} \tag{5.2}$$

Comparing this with (5.1), we see that

$$c(t) = \mathrm{e}^{-L(t)} \quad \text{or} \quad L(t) = -\ln c(t). \tag{5.3}$$

Furthermore, since $(3.11)_2$ must be a solution to the PDE in physical coordinates, we may substitute it into (3.2) to obtain a first order differential equation for $c(t)$. Upon solving this, we find

$$c(t) = \exp\left[-\frac{1}{2Re}(t-2)^2 - \widetilde{L}_0\right], \tag{5.4}$$

where $\widetilde{L}_0$ is a constant. The length of the drop at large times is then given by

$$L(t) = \frac{1}{2Re}(t-2)^2 + \widetilde{L}_0, \tag{5.5}$$

and the velocity of the bottom of the drop can be found by differentiating to get

$$L'(t) = \frac{1}{Re}(t-2). \tag{5.6}$$

The constant $\widetilde{L}_0$ is seen to be the apparent initial length of the drop at the crisis time $t_* = 2$ of inertialess theory when the main drop essentially enters free fall from rest. That is, at later times, the bottom of the drop falls as if it were dropped from rest at time t_* with apparent initial length $\widetilde{L}_0$.

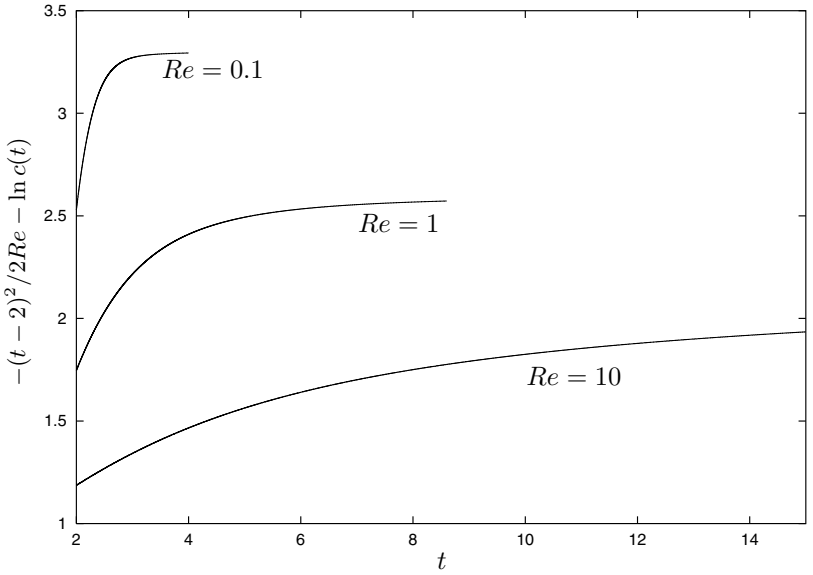


FIG. 5.1. *Plot of the function* $-(t-2)^2/(2Re) - \ln c(t)$ *for* $Re = 0.1, 1.0, 10$, *which at large time* t *gives the apparent initial length* $\widetilde{L}_0$ *of an initially paraboloidal drop. We obtain* $\widetilde{L}_0 \approx 3.3$ *for* $Re = 0.1$, $\widetilde{L}_0 \approx 2.6$ *for* $Re = 1.0$, *and* $\widetilde{L}_0 \approx 2.0$ *for* $Re = 10$.

Our solution of the Eulerian model for initially paraboloidal slender drops involves computation of the function $c(t)$ for extension of the computational domain. Then, at large time, an approximate value for the apparent initial length is given by

$$\widetilde{L}_0 = -\frac{1}{2Re}(t-2)^2 - \ln c(t).$$

With Reynolds numbers $Re = 0.1, 1.0, 10$, we find $\widetilde{L}_0 \approx 3.3, 2.6, 2.0$ (see Figure 5.1). As the Reynolds number increases we must compute to (dimensionless) times well beyond the crisis time $t_* = 2$ of inertialess theory to determine $\widetilde{L}_0$. Figure 5.2 shows $Re\left(L(t) - \widetilde{L}_0\right)$, $Re = 0.1, 1, 10$, versus time t, where $L(t)$ is the length of the drop found by solving (3.2) as described above. At large time this approaches $(t-2)^2/2$, as predicted by (5.5), although, again, as the Reynolds number increases we must compute to times increasingly larger than the crisis time of the inertialess limit to see the agreement.

Another point of interest is that the large-time drop (5.1) (Lagrangian coordinates) or (5.2) (Eulerian coordinates), which derives from the inertialess large-time drop shape (3.10) is observed to be a good representation for the main body of the drop and the lower portion of the filament, as seen by a comparison of Figures 4.6 and 4.2(a); the inertialess large- (crisis) time solution is accurate to within 1% over $x > 15$ (see Figure 4.6), which we see from Figure 4.2(a) is over the bottom third of the drop and filament at this time. The inertialess solution is less accurate in the upper filament region, which is to be expected since inertia and viscous fluid flow are significant in this region of transition from rigid body motion (at increasing velocity) back to zero velocity at the wall; the inertialess solution can be justified only at early time when accelerations are much smaller than gravity, or at larger times in the main drop region which is falling as a rigid body.

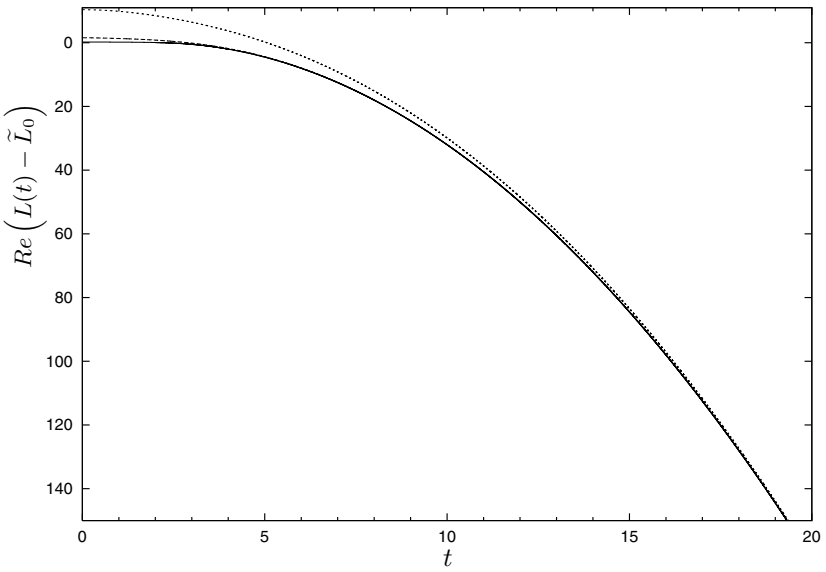


FIG. 5.2. *Plot of* $Re(L(t) - \widetilde{L}_0)$ *with* $L(t)$ *found by solving* (3.2) *for* $Re = 0.1$ *(solid),* 1.0 *(dashed), and* 10 *(dotted), and as predicted at large times* $t > 2$ *by* (5.5), $(t-2)^2/2$ *(solid).*

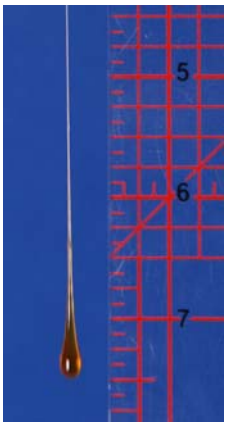


FIG. 5.3. *Dripping golden syrup from a knife. The drop shape is paraboloidal, in contrast to the globular shape of glycerine dripping from a capillary tube as in* [12]. *Note that the scale is in inches.*

Finally we consider the extent to which our computed drop shapes agree with observation. Typically, photographs of viscous fluid drops in the literature (e.g., as in [12]) are of liquids like glycerine dripping from a capillary tube. These drops appear to be considerably more globular in shape than those computed here, but this is essentially a matter of initial conditions. For glycerine-like drops, the initial drop shape is determined by (essentially static) capillarity, as the drop forms slowly at the bottom of the capillary, so that it begins its fall with an already quite globular shape. On the other hand, in the present paper we are assuming an initial shape which is paraboloidal, which is quite like that seen for larger and more viscous drops of liquids like honey falling from a knife or upturned spoon (Figure 5.3). The initial

shape of such rapidly formed drops is influenced very little by surface tension, and their subsequent shape in fall is then quite like those presented here.

Acknowledgments. We gratefully acknowledge valuable discussions with Prof. P. Broadbridge and Dr. Michael Teubner. We are also grateful to an anonymous referee whose helpful comments have resulted in an improved paper.

REFERENCES

[1] L. E. Cram, *A numerical model of droplet formation*, in Computational Techniques & Applications: CTAC-83, J. Noye and C. Fletcher, eds., Elsevier Science/North–Holland, Amsterdam, 1984, pp. 182–188.
[2] P. Concus, R. Finn, and J. McCuan, *Liquid bridges, edge blobs, and Scherk-type capillary surfaces*, Indiana Univ. Math. J., 50 (2001), pp. 411–441.
[3] J. DeWynne, J. R. Ockendon, and P.Wilmott, *On a mathematical model for fiber tapering*, SIAM J. Appl. Math., 49 (1989), pp. 983–990.
[4] P. J. Doyle, *Glass Making Today*, Portcullis Press, Surrey, UK, 1994.
[5] J. Eggers, *Universal pinching of* $3D$ *axisymmetric free surface flow*, Phys. Rev. Lett., 71 (1993), pp. 3458–3460.
[6] J. Eggers, *Nonlinear dynamics and breakup of free surface flows*, Rev. Modern Phys., 69 (1997), pp. 865–929.
[7] J. Eggers and T. F. Dupont, *Drop formation in a one-dimensional approximation of the Navier-Stokes equation*, J. Fluid Mech., 262 (1994), pp. 205–221.
[8] M. A. Matovich and J. R. A. Pearson, *Spinning a molten threadline*, Indust. Engrg. Chem. Fundamentals, 8 (1969), pp. 512–520.
[9] J. E. Matta and R. P. Titus, *Liquid stretching using a falling cylinder*, J. Non-Newton. Fluid, 35 (1990), pp. 215–229.
[10] W. W. Schultz and S. H. Davis, *One-dimensional liquid fibers*, J. Rheol., 26 (1982), pp. 331–345.
[11] S. Senchenko and T. Bohr, *Shape and stability of a viscous thread*, Phys. Rev. E (3), 71 (2005), paper 56301.
[12] X. Shi, M. P. Brenner, and S. R. Nagel, *A cascade of structure in a drop falling from a faucet*, Science, 265 (1994), pp. 219–222.
[13] T. Sridhar, V. Tirtaatmadja, D. A. Nguyen, and R. K. Gupta, *Measurement of extensional viscosity of polymer solutions*, J. Non-Newton. Fluid, 40 (1991), pp. 271–280.
[14] Y. M. Stokes and E. O. Tuck, *The role of inertia in extensional fall of a viscous drop*, J. Fluid Mech., 498 (2004), pp. 205–225.
[15] Y. M. Stokes, E. O. Tuck, and L. W. Schwartz, *Extensional fall of a very viscous fluid drop*, Quart. J. Mech. Appl. Math., 53 (2000), pp. 565–582.
[16] F. T. Trouton, *On the coefficient of viscous traction and its relation to that of viscosity*, Proc. Roy. Soc. A, 77 (1906), pp. 426–440.
[17] D. Vaynblat, J. R. Lister, and T. P. Witelski, *Symmetry and self-similarity in rupture and pinchoff: A geometric bifurcation*, European J. Appl. Math., 12 (2001), pp. 209–232.
[18] E. D. Wilkes, S. D. Phillips, and O. A. Basaran, *Computational and experimental analysis of dynamics of drop formation*, Phys. Fluids, 11 (1999), pp. 3577–3598.
[19] S. D. R. Wilson, *The slow dripping of a viscous fluid*, J. Fluid Mech., 190 (1988), pp. 561–570.
[20] D. F. Zhang and H. A. Stone, *Drop formation in viscous flows at a vertical capillary tube*, Phys. Fluids, 9 (1997), pp. 2234–2242.

SIAM J. APPL. MATH.
Vol. 67, No. 4, pp. 1183–1201

ON THE SHOCKLEY–READ–HALL MODEL: GENERATION-RECOMBINATION IN SEMICONDUCTORS*

THIERRY GOUDON†, VERA MILJANOVIĆ‡, AND CHRISTIAN SCHMEISER§

Abstract. The Shockley–Read–Hall model for generation-recombination of electron-hole pairs in semiconductors based on a quasi-stationary approximation for electrons in a trapped state is generalized to distributed trapped states in the forbidden band and to kinetic transport models for electrons and holes. The quasi-stationary limit is rigorously justified both for the drift-diffusion and for the kinetic model.

Key words. semiconductor, generation, recombination, drift-diffusion, kinetic model

AMS subject classification. 78A35

DOI. 10.1137/060650751

1. Introduction.

The Shockley–Read–Hall (SRH) model was introduced in 1952 [15], [9] to describe the statistics of recombination and generation of holes and electrons in semiconductors occurring through the mechanism of trapping.

The transfer of electrons from the valence band to the conduction band is referred to as the generation of electron-hole pairs (or pair-generation process), since not only is a free electron created in the conduction band, but also a hole in the valence band which can contribute to the charge current. The inverse process is termed recombination of electron-hole pairs. The bandgap between the upper edge of the valence band and the lower edge of the conduction band is very large in semiconductors, which means that a big amount of energy is needed for a direct band-to-band generation event. The presence of trap levels within the forbidden band caused by crystal impurities facilitates this process, since the jump can be split into two parts, each of them "cheaper" in terms of energy. The basic mechanisms are illustrated in Figure 1: (a) hole emission (an electron jumps from the valence band to the trapped level), (b) hole capture (an electron moves from an occupied trap to the valence band, and a hole disappears), (c) electron emission (an electron jumps from the trapped level to the conduction band), (d) electron capture (an electron moves from the conduction band to an unoccupied trap).

Models for this process involve equations for the densities of electrons in the conduction band, holes in the valence band, and trapped electrons. Basic for the SRH model are the drift-diffusion assumption for the transport of electrons and holes, the assumption of one trap level in the forbidden band, and the assumption that

*Received by the editors January 24, 2006; accepted for publication (in revised form) March 7, 2007; published electronically May 29, 2007. This work was supported by the European IHP network "HYKE-Hyperbolic and Kinetic Equations: Asymptotics, Numerics, Analysis" under contract HPRN-CT-2002-00282, and by the Austrian Science Fund under project W008 (Wissenschaftskolleg "Differential Equations").

http://www.siam.org/journals/siap/67-4/65075.html

†Team SIMPAF–INRIA Futurs et Laboratoire Paul Painlevé UMR 8524, CNRS–Université des Sciences et Technologies Lille 1, Cité Scientifique, F-59655 Villeneuve d'Ascq cedex, France (thierry.goudon@math.univ-lille1.fr).

‡Wolfgang Pauli Institut, Universität Wien, Nordbergstraße 15C, 1090 Wien, Austria (vera.miljanovic@univie.ac.at).

§Fakultät für Mathematik, Universität Wien, Nordbergstraße 15C, 1090 Wien, Austria, and RICAM Linz, Österreichische Akademie der Wissenschaften, Altenbergstr. 56, 4040 Linz, Austria (christian.schmeiser@univie.ac.at).

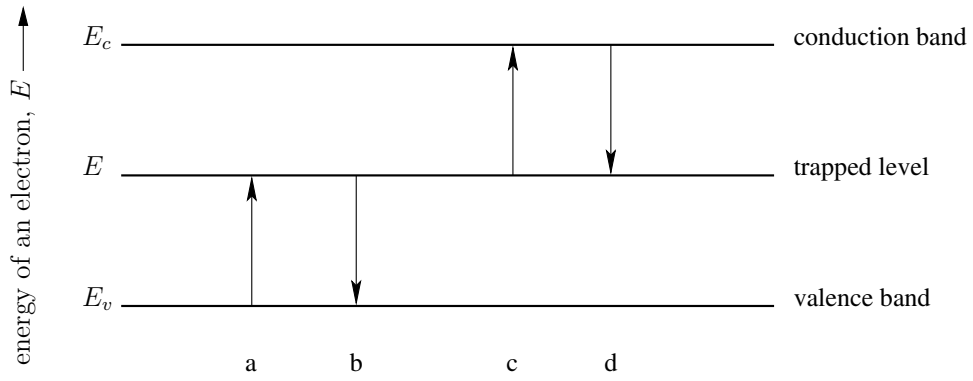


FIG. 1. *The four basic processes of electron-hole recombination.*

the dynamics of the trapped electrons is quasi-stationary, which can be motivated by the smallness of the density of trapped states compared to typical carrier densities. This last assumption leads to the elimination of the density of trapped electrons from the system and to a nonlinear effective recombination-generation rate, reminiscent of Michaelis–Menten kinetics in chemistry. This model is an important ingredient of simulation models for semiconductor devices (see, e.g., [10], [14]).

In this work, two generalizations of the classical SRH model are considered: Instead of a single trapped state, a distribution of trapped states across the forbidden band is allowed, and, in a second step, a semiclassical kinetic model including the fermion nature of the charge carriers is introduced. Although direct band-to-band recombination-generation (see, e.g., [13]) and impact ionization (e.g., [2], [3]) have been modelled on the kinetic level before, this is (to the best of the authors' knowledge) the first attempt to derive a "kinetic SRH model." (We also mention the modelling discussions and numerical simulations in [7], [8].)

For both the drift-diffusion and the kinetic models with self-consistent electric fields existence results and rigorous results concerning the quasi-stationary limit are proven. For the drift-diffusion problem, the essential estimate is derived similarly to [6], where the quasi-neutral limit has been carried out. For the kinetic model Degond's approach [4] for the existence of solutions of the Vlasov–Poisson problem is extended. Actually, the existence theory already provides the uniform estimates necessary for passing to the quasi-stationary limit.

In the following section, the drift-diffusion based model is formulated and nondimensionalized, and the SRH model is formally derived. Section 3 contains the rigorous justification of the passage to the quasi-stationary limit. Section 4 corresponds to section 2, dealing with the kinetic model, and in section 5 existence of global solutions for the kinetic model is proven, and the quasi-stationary limit is justified.

2. The drift-diffusion Shockley–Read–Hall model. We consider a semiconductor crystal with a forbidden band represented by the energy interval (E_v, E_c) with the valence band edge E_v and the conduction band edge E_c. The constant (in space) number density N_{tr} of trapped states is obtained by summing up contributions across the forbidden band:

$$N_{tr} = \int_{E_v}^{E_c} M_{tr}(E)\, dE.$$

Here $M_{tr}(E)$ is the energy dependent density of available trapped states. The position density of occupied traps is given by

$$n_{tr}(f_{tr})(x,t) = \int_{E_v}^{E_c} M_{tr}(E) f_{tr}(x,E,t)\, dE,$$

where $f_{tr}(x,E,t)$ is the fraction of occupied trapped states at position $x \in \Omega$, energy $E \in (E_v, E_c)$, and time $t \geq 0$. Note that $0 \leq f_{tr} \leq 1$ should hold from a physical point of view.

The evolution of f_{tr} is coupled to those of the density of electrons in the conduction band, denoted by $n(x,t) \geq 0$, and the density of holes in the valence band, denoted by $p(x,t) \geq 0$. Electrons and holes are oppositely charged. The coupling is expressed through the following quantities:

$$S_n = \frac{1}{\tau_n N_{tr}}\big[n_0 f_{tr} - n(1-f_{tr})\big], \qquad S_p = \frac{1}{\tau_p N_{tr}}\big[p_0(1-f_{tr}) - p f_{tr}\big], \tag{1}$$

$$R_n = \int_{E_v}^{E_c} S_n M_{tr}\, dE, \qquad R_p = \int_{E_v}^{E_c} S_p M_{tr}\, dE. \tag{2}$$

Indeed, the governing equations are given by

$$\partial_t f_{tr} = S_p - S_n = \frac{p_0}{\tau_p N_{tr}} + \frac{n}{\tau_n N_{tr}} - f_{tr}\left(\frac{p_0+p}{\tau_p N_{tr}} + \frac{n_0+n}{\tau_n N_{tr}}\right), \tag{3}$$

$$\partial_t n = \nabla \cdot J_n + R_n, \quad J_n = \mu_n(U_T \nabla n - n \nabla V), \tag{4}$$

$$\partial_t p = -\nabla \cdot J_p + R_p, \quad J_p = -\mu_p(U_T \nabla p + p \nabla V), \tag{5}$$

$$\varepsilon_s \Delta V = q(n + n_{tr}(f_{tr}) - p - C). \tag{6}$$

For the current densities J_n, J_p we use the simplest possible model, the drift diffusion ansatz, with constant mobilities μ_n, μ_p, and with thermal voltage U_T. Moreover, since the trapped states have fixed positions, no flux appears in (3).

By R_n and R_p we denote the recombination-generation rates for n and p, respectively. The rate constants are $\tau_n(E)$, $\tau_p(E)$, $n_0(E)$, $p_0(E)$, where $n_0(E)p_0(E) = {n_i}^2$ with the energy independent intrinsic density n_i.

Integration of (3) yields

$$\partial_t n_{tr} = R_p - R_n. \tag{7}$$

By adding (4), (5), (7), we obtain the continuity equation

$$\partial_t(p - n - n_{tr}) + \nabla \cdot (J_n + J_p) = 0, \tag{8}$$

with the total charge density $p - n - n_{tr}$ and the total current density $J_n + J_p$.

In the Poisson equation (6), $V(x,t)$ is the electrostatic potential, ε_s the permittivity of the semiconductor material, q the elementary charge, and $C = C(x)$ the given doping profile.

Note that if τ_n, τ_p, n_0, p_0 are independent of E, or if there exists only one trap level E_{tr} with $M_{tr}(E) = N_{tr}\delta(E - E_{tr})$, then $R_n = \frac{1}{\tau_n}[n_0 \frac{n_{tr}}{N_{tr}} - n(1 - \frac{n_{tr}}{N_{tr}})]$, $R_p = \frac{1}{\tau_p}[p_0(1 - \frac{n_{tr}}{N_{tr}}) - p\frac{n_{tr}}{N_{tr}}]$, and (4), (5) together with (7) are a closed system governing the evolution of n, p, and n_{tr}.

We now introduce a scaling of n, p, and f_{tr} in order to render (4)–(6) dimensionless:

Scaling of parameters:

i. $M_{tr} \to \frac{N_{tr}}{E_c - E_v} M_{tr}$.

ii. $\tau_{n,p} \to \bar{\tau}\tau_{n,p}$, where $\bar{\tau}$ is a typical value for τ_n and τ_p.

iii. $\mu_{n,p} \to \bar{\mu}\mu_{n,p}$, where $\bar{\mu}$ is a typical value for $\mu_{n,p}$.

iv. $(n_0, p_0, n_i, C) \to \bar{C}(n_0, p_0, n_i, C)$, where $\bar{C}$ is a typical value of C.

Scaling of unknowns:

v. $(n, p) \to \bar{C}(n, p)$.

vi. $n_{tr} \to N_{tr} n_{tr}$.

vii. $V \to U_T V$.

viii. $f_{tr} \to f_{tr}$.

Scaling of independent variables:

ix. $E \to E_v + (E_c - E_v)E$.

x. $x \to \sqrt{\bar{\mu} U_T \bar{\tau}}\, x$, where the reference length is a typical diffusion length before recombination.

xi. $t \to \bar{\tau} t$, where the reference time is a typical carrier life time.

Dimensionless parameters:

xii. $\lambda = \sqrt{\frac{\varepsilon_s}{q\bar{C}\bar{\mu}\bar{\tau}}} = \frac{1}{\bar{x}}\sqrt{\frac{\varepsilon_s U_T}{q\bar{C}}}$ is the scaled Debye length.

xiii. $\varepsilon = \frac{N_{tr}}{\bar{C}}$ is the ratio of the density of traps to the typical doping density, and will be assumed to be small: $\varepsilon \ll 1$.

The scaled system reads as follows:

(9)
$$\varepsilon \partial_t f_{tr} = S_p(p, f_{tr}) - S_n(n, f_{tr}), \quad S_p = \frac{1}{\tau_p}\big[p_0(1 - f_{tr}) - p f_{tr}\big],$$
$$S_n = \frac{1}{\tau_n}\big[n_0 f_{tr} - n(1 - f_{tr})\big],$$

(10)
$$\partial_t n = \nabla \cdot J_n + R_n(n, f_{tr}), \quad J_n = \mu_n(\nabla n - n\nabla V), \quad R_n = \int_0^1 S_n M_{tr}\, dE,$$

(11)
$$\partial_t p = -\nabla \cdot J_p + R_p(p, f_{tr}), \quad J_p = -\mu_p(\nabla p + p\nabla V), \quad R_p = \int_0^1 S_p M_{tr}\, dE,$$

(12)
$$\lambda^2 \Delta V = n + \varepsilon n_{tr} - p - C, \quad n_{tr}(f_{tr}) = \int_0^1 f_{tr} M_{tr}\, dE,$$

with $n_0(E)p_0(E) = n_i^2$ and $\int_0^1 M_{tr}\, dE = 1$.

By letting $\varepsilon \to 0$ in (9) formally, we obtain $f_{tr} = \frac{\tau_n p_0 + \tau_p n}{\tau_n(p + p_0) + \tau_p(n + n_0)}$, and the reduced system has the following form:

(13) $$\partial_t n = \nabla \cdot J_n + R(n, p),$$

(14) $$\partial_t p = -\nabla \cdot J_p + R(n, p),$$

(15) $$R(n, p) = ({n_i}^2 - np) \int_0^1 \frac{M_{tr}(E)}{\tau_n(E)(p + p_0(E)) + \tau_p(E)(n + n_0(E))\, dE},$$

(16) $$\lambda^2 \Delta V = n - p - C.$$

Note that if τ_n, τ_p, n_0, p_0 are independent of E or if there exists only one trap level, then we would have the standard SRH model, with $R = \frac{{n_i}^2 - np}{\tau_n(p+p_0)+\tau_p(n+n_0)}$. Existence and uniqueness of solutions of the limiting system (13)–(16) under assumptions (21)–(25) stated below is a standard result in semiconductor modelling. A proof can be found in, e.g., [10].

3. Rigorous derivation of the drift-diffusion Shockley–Read–Hall model. We consider the system (9)–(12) with the position x varying in a bounded domain $\Omega \in \mathbb{R}^3$ (all our results are easily extended to the one- and two-dimensional situations), the energy $E \in (0,1)$, and time $t > 0$, subject to initial conditions

$$n(x,0) = n_I(x), \quad p(x,0) = p_I(x), \quad f_{tr}(x,E,0) = f_{tr,I}(x,E) \tag{17}$$

and mixed Dirichlet–Neumann boundary conditions

$$n(x,t) = n_D(x,t), \quad p(x,t) = p_D(x,t), \quad V(x,t) = V_D(x,t), \quad x \in \partial\Omega_D \subset \partial\Omega, \tag{18}$$

and

$$\frac{\partial n}{\partial \nu}(x,t) = \frac{\partial p}{\partial \nu}(x,t) = \frac{\partial V}{\partial \nu}(x,t) = 0, \quad x \in \partial\Omega_N := \partial\Omega \setminus \partial\Omega_D, \tag{19}$$

where ν is the unit outward normal vector along $\partial\Omega_N$. We permit the special case that either $\partial\Omega_D$ or $\partial\Omega_N$ is empty. More precisely, we assume that either $\partial\Omega_D$ has positive $(d-1)$-dimensional measure, or it is empty. In the second situation ($\partial\Omega_D$ empty) we have to assume total charge neutrality; i.e.,

$$\int_\Omega (n + \varepsilon n_{tr} - p - C)\, dx = 0 \qquad \text{if } \partial\Omega = \partial\Omega_N. \tag{20}$$

The potential is then determined only up to a (physically irrelevant) additive constant.

The following assumptions on the data will be used: For the boundary data, given any $0 < T < \infty$,

$$n_D, p_D \in W^{1,\infty}(0,T; W^{1,\infty}_{loc}(\Omega)), \quad V_D \in L^\infty(0,T; W^{1,6}(\Omega)); \tag{21}$$

for the initial data

$$n_I, p_I \in H^1(\Omega) \cap L^\infty(\Omega), \quad 0 \le f_{tr,I} \le 1, \tag{22}$$

$$\int_\Omega (n_I + \varepsilon n_{tr}(f_{tr,I}) - p_I - C)\, dx = 0 \qquad \text{if } \partial\Omega = \partial\Omega_N; \tag{23}$$

for the doping profile

$$C \in L^\infty(\Omega); \tag{24}$$

and for the recombination-generation rate constants

$$n_0, p_0, \tau_n, \tau_p \in L^\infty((0,1)), \quad \tau_n, \tau_p \ge \tau_{min} > 0. \tag{25}$$

With these assumptions, a local existence and uniqueness result for the problem (9)–(12), (17)–(19) for fixed positive ε can be proven by a straightforward extension of the approach in [5] (see also [10]). In the following, local existence will be assumed, and we shall concentrate on obtaining bounds which guarantee global existence and

which are uniform in ε as $\varepsilon \to 0$. For the sake of simplicity, we consider that the data in (21), (22), and (24) do not depend on ε; of course, our strategy works when dealing with sequences of data bounded in the mentioned spaces.

The following result is a generalization of [6, Lemma 3.1], where the case of homogeneous Neumann boundary conditions and vanishing recombination was treated. Our proof uses a similar approach.

LEMMA 3.1. *Let the assumptions* (21)–(25) *be satisfied. Then, the solution of* (9)–(12), (17)–(19) *exists for all times and satisfies* $n, p \in L^\infty(0,T;L^\infty(\Omega)) \cap L^2(0,T;H^1(\Omega))$ *uniformly in* ε *as* $\varepsilon \to 0$ *as well as* $0 \le f_{tr} \le 1$.

Proof. Global existence will be a consequence of the following estimates. Introducing the new variables $\widetilde{n} = n - n_D$, $\widetilde{p} = p - p_D$, $\widetilde{C} = C - \varepsilon n_{tr} - n_D + p_D$, (10)–(12) take the following form:

$$\partial_t \widetilde{n} = \nabla \cdot J_n + R_n - \partial_t n_D, \quad J_n = \mu_n\big[\nabla\widetilde{n} + \nabla n_D - (\widetilde{n} + n_D)\nabla V\big], \tag{26}$$

$$\partial_t \widetilde{p} = -\nabla J_p + R_p - \partial_t p_D, \quad J_p = -\mu_p\big[\nabla\widetilde{p} + \nabla p_D + (\widetilde{p} + p_D)\nabla V\big], \tag{27}$$

$$\lambda^2 \Delta V = \widetilde{n} - \widetilde{p} - \widetilde{C}. \tag{28}$$

As a consequence of $0 \le f_{tr} \le 1$, $\widetilde{C} \in L^\infty((0,\infty)\times\Omega)$ holds. For $q \ge 2$ and even, we multiply (26) by $\widetilde{n}^{q-1}/\mu_n$ and (27) by $\widetilde{p}^{q-1}/\mu_p$, and add:

$$\begin{aligned}
\frac{d}{dt}\int_\Omega \left[\frac{\widetilde{n}^q}{q\mu_n} + \frac{\widetilde{p}^q}{q\mu_p}\right] dx \;=\; & -(q-1)\int_\Omega \widetilde{n}^{q-2}\nabla\widetilde{n}\nabla n\,dx - (q-1)\int_\Omega \widetilde{p}^{q-2}\nabla\widetilde{p}\nabla p\,dx \\
& + (q-1)\int_\Omega \left[\widetilde{n}^{q-2} n\nabla\widetilde{n} - \widetilde{p}^{q-2} p\nabla\widetilde{p}\right]\nabla V\,dx \\
& + \int_\Omega \frac{\widetilde{n}^{q-1}}{\mu_n}(R_n - \partial_t n_D) + \int_\Omega \frac{\widetilde{p}^{q-1}}{\mu_p}(R_p - \partial_t p_D) \\
=:\; & I_1 + I_2 + I_3 + I_4 + I_5.
\end{aligned} \tag{29}$$

Using the assumptions on n_D, p_D and $|R_n| \le C(n+1)$, $|R_p| \le C(p+1)$, we estimate

$$I_4 \le C\int_\Omega |\widetilde{n}|^{q-1}(n+1)\,dx \le C\left(\int_\Omega \widetilde{n}^q\,dx + 1\right), \quad I_5 \le C\left(\int_\Omega \widetilde{p}^q\,dx + 1\right).$$

The term I_3 can be rewritten as follows:

$$\begin{aligned}
I_3 = & \int_\Omega \left[\widetilde{n}^{q-1}\nabla\widetilde{n} - \widetilde{p}^{q-1}\nabla\widetilde{p}\right]\nabla V\,dx \\
& + \int_\Omega \left[\widetilde{n}^{q-2}\nabla\widetilde{n}\right](n_D\nabla V)\,dx - \int_\Omega \left[\widetilde{p}^{q-2}\nabla\widetilde{p}\right](p_D\nabla V)\,dx \\
= & -\frac{1}{\lambda^2 q}\int_\Omega \left[\widetilde{n}^q - \widetilde{p}^q\right](\widetilde{n} - \widetilde{p} - \widetilde{C})\,dx \\
& - \frac{1}{\lambda^2(q-1)}\int_\Omega \widetilde{n}^{q-1}\left(\nabla n_D\nabla V + n_D(\widetilde{n} - \widetilde{p} - \widetilde{C})\right)dx \\
& + \frac{1}{\lambda^2(q-1)}\int_\Omega \widetilde{p}^{q-1}\left(\nabla p_D\nabla V + p_D(\widetilde{n} - \widetilde{p} - \widetilde{C})\right)dx.
\end{aligned}$$

The second equality uses integration by parts and (28). The first term on the right-hand side is the only term of degree $q+1$. It reflects the quadratic nonlinearity of

the problem. Fortunately, it can be written as the sum of a term of degree q and a nonnegative term. By estimation of the terms of degree q using the assumptions on n_D and p_D as well as $\|\nabla V\|_{L^q(\Omega)} \le C(\|\widetilde{n}\|_{L^q(\Omega)} + \|\widetilde{p}\|_{L^q(\Omega)} + \|\widetilde{C}\|_{L^q(\Omega)})$, we obtain

$$\begin{aligned} I_3 &\le -\frac{1}{\lambda^2 q}\int_\Omega [\widetilde{n}^q - \widetilde{p}^q]\,(\widetilde{n}-\widetilde{p})\,dx + C\left(\int_\Omega (\widetilde{n}^q + \widetilde{p}^q)\,dx + 1\right) \\ &\le C\left(\int_\Omega (\widetilde{n}^q + \widetilde{p}^q)\,dx + 1\right). \end{aligned}$$

The integral I_1 can be written as

$$I_1 = -\int_\Omega \widetilde{n}^{q-2}|\nabla n|^2\,dx + \int_\Omega \widetilde{n}^{q-2}\nabla n_D \nabla n\,dx. \tag{30}$$

By rewriting the integrand in the second integral as

$$\widetilde{n}^{q-2}\nabla n_D \nabla n = \widetilde{n}^{\frac{q-2}{2}}\nabla n \widetilde{n}^{\frac{q-2}{2}}\nabla n_D$$

and applying the Cauchy–Schwarz inequality, we have the following estimate for (30):

$$\begin{aligned} I_1 &\le -\int_\Omega \widetilde{n}^{q-2}|\nabla n|^2\,dx + \sqrt{\int_\Omega \widetilde{n}^{q-2}|\nabla n|^2\,dx \int_\Omega \widetilde{n}^{q-2}|\nabla n_D|^2\,dx} \\ &\le -\frac{1}{2}\int_\Omega \widetilde{n}^{q-2}|\nabla n|^2\,dx + C\|\widetilde{n}\|_{L^q}^{q-2} \le -\frac{1}{2}\int_\Omega \widetilde{n}^{q-2}|\nabla n|^2\,dx + C\left(\int_\Omega \widetilde{n}^q\,dx + 1\right). \end{aligned}$$

For I_2, the same reasoning (with n and n_D replaced by p and p_D, respectively) yields an analogous estimate. Collecting our results, we obtain

$$\begin{aligned} \frac{d}{dt}\int_\Omega \left[\frac{\widetilde{n}^q}{q\mu_n} + \frac{\widetilde{p}^q}{q\mu_p}\right]dx &\le -\frac{1}{2}\int_\Omega \widetilde{n}^{q-2}|\nabla n|^2\,dx - \frac{1}{2}\int_\Omega \widetilde{p}^{q-2}|\nabla p|^2\,dx \\ &\quad + C\left(\int_\Omega (\widetilde{n}^q + \widetilde{p}^q)\,dx + 1\right). \end{aligned} \tag{31}$$

Since $q \ge 2$ is even, the first two terms on the right-hand side are nonpositive, and the Gronwall lemma gives

$$\int_\Omega (\widetilde{n}^q + \widetilde{p}^q)\,dx \le e^{qCt}\left(\int_\Omega (\widetilde{n}(t=0)^q + \widetilde{p}(t=0)^q)\,dx + 1\right).$$

A uniform-in-q-and-ε estimate for $\|n\|_{L^q}$, $\|p\|_{L^q}$ follows, and the uniform-in-ε bound in $L^\infty(0,T;L^\infty(\Omega))$ is obtained in the limit $q \to \infty$. The estimate in $L^2(0,T;H^1(\Omega))$ is then derived by returning to (31) with $q = 2$. ∎

Now we are ready to prove the main result of this section.

THEOREM 3.2. *Let the assumptions of Theorem* 3.1 *be satisfied. Then, as $\varepsilon \to 0$, for every $T > 0$, the solution (f_{tr}, n, p, V) of* (9)–(12), (17)–(19) *converges with convergence of f_{tr} in $L^\infty((0,T)\times\Omega\times(0,1))$ weak*, n and p in $L^2((0,T)\times\Omega)$, and V in $L^2(0,T;H^1(\Omega))$. The limits of n, p, and V satisfy* (13)–(19).

Proof. The L^∞-bounds for f_{tr}, n, and p, which are uniform with respect to ε, and the Poisson equation (12) imply ∇V is bounded in $L^2((0,T)\times\Omega)$. From the definition of J_n, J_p (see (4), (5)), it then follows that $J_n, J_p \in L^2((0,T)\times\Omega)$. Then (10) and (11) together with $R_n, R_p \in L^\infty((0,T)\times\Omega)$ imply $\partial_t n, \partial_t p \in L^2(0,T;H^{-1}(\Omega))$.

The previous result and the Aubin lemma (see, e.g., Simon [16, Corollary 4, p. 85]) give compactness of n and p in $L^2((0,T)\times\Omega)$.

We already know from the Poisson equation that $\nabla V \in L^\infty(0,T;H^1(\Omega))$. By taking the time derivative of (12), one obtains

$$\partial_t \Delta V = \nabla\cdot(J_n + J_p),$$

with the consequence that $\partial_t \nabla V$ is bounded (uniformly with respect to ε) in $L^2((0,T)\times\Omega)$. Therefore, the Aubin lemma can again be applied as above to prove compactness of ∇V in $L^2((0,T)\times\Omega)$.

These results and the weak compactness of f_{tr} are sufficient for passing to the limit in the nonlinear terms $n\nabla V$, $p\nabla V$, nf_{tr}, and pf_{tr}. Let us also remark that $\partial_t n$ and $\partial_t p$ are bounded in $L^2(0,T;H^{-1}(\Omega))$, so that n,p are compact in $C^0([0,T];L^2(\Omega)$ weak). With this remark the initial data for the limit equation make sense. By the uniqueness result for the limiting problem (mentioned at the end of section 2), the convergence is not restricted to subsequences. ∎

4. A kinetic Shockley–Read–Hall model. In this section we replace the drift-diffusion model for electrons and holes by a semiclassical kinetic transport model. It is governed by the system

$$\partial_t f_n + v_n(k)\cdot\nabla_x f_n + \frac{q}{\hbar}\nabla_x V\cdot\nabla_k f_n = Q_n(f_n) + Q_{n,r}(f_n,f_{tr}), \tag{32}$$

$$\partial_t f_p + v_p(k)\cdot\nabla_x f_p - \frac{q}{\hbar}\nabla_x V\cdot\nabla_k f_p = Q_p(f_p) + Q_{p,r}(f_p,f_{tr}), \tag{33}$$

$$\partial_t f_{tr} = S_p(f_p,f_{tr}) - S_n(f_n,f_{tr}), \tag{34}$$

$$\varepsilon_s\Delta_x V = q(n + n_{tr} - p - C), \tag{35}$$

where $f_i(x,k,t)$ represents the particle distribution function (with $i=n$ for electrons and $i=p$ for holes) at time $t\ge 0$, at the position $x\in\mathbb{R}^3$, and at the wave vector (or generalized momentum) $k\in\mathbb{R}^3$. All functions of k have the periodicity of the reciprocal lattice of the semiconductor crystal. Equivalently, we shall consider only $k\in B$, where B is the Brillouin zone, i.e., the set of all k which are closer to the origin than to any other lattice point, with periodic boundary conditions on ∂B.

The coefficient functions $v_n(k)$ and $v_p(k)$ denote the electron and hole velocities, respectively, which are related to the electron and hole band diagrams by

$$v_n(k) = \nabla_k\varepsilon_n(k)/\hbar, \quad v_p(k) = -\nabla_k\varepsilon_p(k)/\hbar,$$

where $\hbar$ is the reduced Planck constant. The elementary charge is still denoted by q.

The collision operators Q_n and Q_p describe the interactions between the particles and the crystal lattice. They involve several physical phenomena and can be written in the general form

$$Q_n(f_n) = \int_B \widetilde{\Phi}_n(k,k')[M_n f_n'(1-f_n) - M_n' f_n(1-f_n')]\,dk', \tag{36}$$

$$Q_p(f_p) = \int_B \widetilde{\Phi}_p(k,k')[M_p f_p'(1-f_p) - M_p' f_p(1-f_p')]\,dk', \tag{37}$$

with the primes denoting evaluation at k', with the nonnegative, symmetric scattering cross sections $\widetilde{\Phi}_n(k,k')$ and $\widetilde{\Phi}_p(k,k')$, and with the Maxwellians

$$M_n(k) = c_n\exp(-\varepsilon_n(k)/k_BT), \quad M_p(k) = c_p\exp(-\varepsilon_p(k)/k_BT),$$

where k_BT is the thermal energy of the semiconductor crystal lattice and the constants c_n, c_p are chosen such that

$$\int_B M_n \, dk = \int_B M_p \, dk = 1.$$

The remaining collision operators $Q_{n,r}(f_n, f_{tr})$ and $Q_{p,r}(f_p, f_{tr})$ model the generation and recombination processes and are given by

$$Q_{n,r}(f_n, f_{tr}) = \int_{E_v}^{E_c} \hat{S}_n(f_n, f_{tr}) M_{tr} \, dE, \tag{38}$$

with

$$\hat{S}_n(f_n, f_{tr}) = \frac{\Phi_n(k, E)}{N_{tr}} \left[n_0 M_n f_{tr}(1 - f_n) - f_n(1 - f_{tr}) \right],$$

and

$$Q_{p,r}(f_p, f_{tr}) = \int_{E_v}^{E_c} \hat{S}_p(f_p, f_{tr}) M_{tr} \, dE, \tag{39}$$

with

$$\hat{S}_p(f_p, f_{tr}) = \frac{\Phi_p(k, E)}{N_{tr}} \left[p_0 M_p(1 - f_p)(1 - f_{tr}) - f_p f_{tr} \right],$$

and where $\Phi_{n,p}$ are nonnegative and $M_{tr}(x, E)$ is the same density of available trapped states as for the drift-diffusion model, except that we allow for a position dependence now. This will be commented on below. The parameter N_{tr} is now determined as $N_{tr} = \sup_{x \in \mathbb{R}^3} \int_0^1 M_{tr}(x, E) \, dE$.

The right-hand side in the equation for the occupancy $f_{tr}(x, E, t)$ of the trapped states is defined by

$$S_n(f_n, f_{tr}) = \int_B \hat{S}_n \, dk = \lambda_n[n_0 M_n(1 - f_n)] f_{tr} - \lambda_n[f_n](1 - f_{tr}), \tag{40}$$

with $\lambda_n[g] = \int_B \Phi_n g \, dk$, and

$$S_p(f_p, f_{tr}) = \int_B \hat{S}_p \, dk = \lambda_p[p_0 M_p(1 - f_p)](1 - f_{tr}) - \lambda_p[f_p] f_{tr}, \tag{41}$$

with $\lambda_p[g] = \int_B \Phi_p g \, dk$.

The factors $(1 - f_n)$ and $(1 - f_p)$ take into account the Pauli exclusion principle, which therefore manifests itself in the requirement that the values of the distribution function have to respect the bounds $0 \le f_n, f_p \le 1$.

The position densities on the right-hand side of the Poisson equation (35) are given by

$$n(x, t) = \int_B f_n \, dk, \quad p(x, t) = \int_B f_p \, dk, \quad n_{tr}(x, t) = \int_{E_v}^{E_c} f_{tr} M_{tr} \, dE.$$

The following scaling, which is strongly related to the one used for the drift-diffusion model, will render (32)–(35) dimensionless:

Scaling of parameters:

i. $M_{tr} \to \frac{N_{tr}}{E_v - E_c} M_{tr}$.
ii. $(\varepsilon_n, \varepsilon_p) \to k_B T(\varepsilon_n, \varepsilon_p)$, with the thermal energy $k_B T$.
iii. $(\Phi_n, \Phi_p) \to \tau_{rg}^{-1}(\Phi_n, \Phi_p)$, where τ_{rg} is a typical carrier life time.
iv. $(\widetilde{\Phi}_n, \widetilde{\Phi}_p) \to \tau_{coll}^{-1}(\widetilde{\Phi}_n, \widetilde{\Phi}_p)$.
v. $(n_0, p_0, C) \to \overline{C}(n_0, p_0, C)$, where $\overline{C}$ is a typical value of $|C|$.
vi. $(M_n, M_p) \to \overline{C}^{-1}(M_n, M_p)$.

Scaling of independent variables:

vii. $x \to k_B T \sqrt{\tau_{rg}\tau_{coll}}\overline{C}^{-1/3}\hbar^{-1}\, x$.
viii. $t \to \tau_{rg} t$.
ix. $k \to \overline{C}^{1/3} k$.
x. $E \to E_v + (E_c - E_v)E$.

Scaling of unknowns:

xi. $(f_n, f_p, f_{tr}) \to (f_n, f_p, f_{tr})$.
xii. $V \to U_T V$, with the thermal voltage $U_T = k_B T/q$.

Dimensionless parameters:

xiii. $\alpha^2 = \frac{\tau_{coll}}{\tau_{rg}}$.
xiv. $\lambda = \frac{\hbar}{q\sqrt{\tau_{rg}\tau_{coll}}\overline{C}^{1/6}} \sqrt{\frac{\varepsilon_s}{k_B T}}$.
xv. $\varepsilon = \frac{N_{tr}}{\overline{C}}$, where again we shall study the situation $\varepsilon \ll 1$.

Finally, the scaled system reads as follows:

$$\alpha^2 \partial_t f_n + \alpha v_n(k) \cdot \nabla_x f_n + \alpha \nabla_x V \cdot \nabla_k f_n = Q_n(f_n) + \alpha^2 Q_{n,r}(f_n, f_{tr}), \tag{42}$$

$$\alpha^2 \partial_t f_p + \alpha v_p(k) \cdot \nabla_x f_p - \alpha \nabla_x V \cdot \nabla_k f_p = Q_p(f_p) + \alpha^2 Q_{p,r}(f_p, f_{tr}), \tag{43}$$

$$\varepsilon \partial_t f_{tr} = S_p(f_p, f_{tr}) - S_n(f_n, f_{tr}), \tag{44}$$

$$\lambda^2 \Delta_x V = n + \varepsilon n_{tr} - p - C = -\rho, \tag{45}$$

with $v_n = \nabla_k \varepsilon_n$, $v_p = -\nabla_k \varepsilon_p$, with Q_n and Q_p still having the form (36) and (37), respectively, with the scaled Maxwellians

$$M_n(k) = c_n \exp(-\varepsilon_n(k)), \qquad M_p(k) = c_p \exp(-\varepsilon_p(k)), \tag{46}$$

and with the recombination-generation terms

$$Q_{n,r}(f_n, f_{tr}) = \int_0^1 \hat{S}_n M_{tr}\, dE, \qquad Q_{p,r}(f_p, f_{tr}) = \int_0^1 \hat{S}_p M_{tr}\, dE, \tag{47}$$

with

$$\hat{S}_n = \Phi_n[n_0 M_n f_{tr}(1-f_n) - f_n(1-f_{tr})], \quad \hat{S}_p = \Phi_p[p_0 M_p(1-f_{tr})(1-f_p) - f_p f_{tr}]. \tag{48}$$

The right-hand side of (44) still has the form (40), (41). The position densities are given by

$$n = \int_B f_n\, dk, \qquad p = \int_B f_p\, dk, \qquad n_{tr} = \int_0^1 f_{tr} M_{tr}\, dE. \tag{49}$$

The system (42)–(44) conserves the total charge $\rho = p + C - n - \varepsilon n_{tr}$. With the definition

$$J_n = -\frac{1}{\alpha} \int_B v_n f_n\, dk, \quad J_p = \frac{1}{\alpha} \int_B v_p f_p\, dk$$

of the current densities, the following continuity equation holds formally:

$$\partial_t \rho + \nabla_x \cdot (J_n + J_p) = 0.$$

Formally setting $\varepsilon = 0$ in (44), we obtain

$$\overline{f}_{tr}(f_n, f_p) = \frac{p_0 \lambda_p[M_p(1-f_p)] + \lambda_n[f_n]}{p_0 \lambda_p[M_p(1-f_p)] + \lambda_p[f_p] + \lambda_n[f_n] + n_0 \lambda_n[M_n(1-f_n)]}.$$

Substituting $\overline{f}_{tr}$ into (47) leads to the kinetic SRH recombination-generation operators

$$\begin{aligned} \overline{Q}_{n,r}(f_n, f_p) &= \overline{g}_n[f_n, f_p](1-f_n) - \overline{r}_n[f_n, f_p] f_n, \\ \overline{Q}_{p,r}(f_n, f_p) &= \overline{g}_p[f_n, f_p](1-f_p) - \overline{r}_p[f_n, f_p] f_p, \end{aligned} \tag{50}$$

with

$$\begin{aligned} \overline{g}_n &= \int_0^1 \frac{\Phi_n M_n n_0 \Big(p_0 \lambda_p[M_p(1-f_p)] + \lambda_n[f_n] \Big) M_{tr}}{p_0 \lambda_p[M_p(1-f_p)] + \lambda_p[f_p] + \lambda_n[f_n] + n_0 \lambda_n[M_n(1-f_n)]}\, dE, \\ \overline{r}_n &= \int_0^1 \frac{\Phi_n \Big(\lambda_p[f_p] + n_0 \lambda_n[M_n(1-f_n)] \Big) M_{tr}}{p_0 \lambda_p[M_p(1-f_p)] + \lambda_p[f_p] + \lambda_n[f_n] + n_0 \lambda_n[M_n(1-f_n)]}\, dE, \\ \overline{g}_p &= \int_0^1 \frac{\Phi_p M_p p_0 \Big(n_0 \lambda_n[M_n(1-f_n)] + \lambda_p[f_p] \Big) M_{tr}}{p_0 \lambda_p[M_p(1-f_p)] + \lambda_p[f_p] + \lambda_n[f_n] + n_0 \lambda_n[M_n(1-f_n)]}\, dE, \\ \overline{r}_p &= \int_0^1 \frac{\Phi_p \Big(\lambda_n[f_n] + p_0 \lambda_p[M_p(1-f_p)] \Big) M_{tr}}{p_0 \lambda_p[M_p(1-f_p)] + \lambda_p[f_p] + \lambda_n[f_n] + n_0 \lambda_n[M_n(1-f_n)]}\, dE. \end{aligned}$$

Of course, the limiting model still conserves charge, which is expressed by the identity

$$\int_B \overline{Q}_{n,r}\, dk = \int_B \overline{Q}_{p,r}\, dk.$$

Pairs of electrons and holes are generated or recombine, however, generally not with the same wave vector. This absence of momentum conservation is reasonable since the process involves an interaction with the trapped states fixed within the crystal lattice.

5. Rigorous derivation of the kinetic Shockley–Read–Hall model. The limit $\varepsilon \to 0$ will be carried out rigorously in an initial value problem for the kinetic model: From now on we work with $x \in \mathbb{R}^3$ (and we avoid any discussion on boundary conditions and possible boundary layers). Concerning the behavior for $|x| \to \infty$, we shall require the densities to be in L^1 and use the Newtonian potential solution of the Poisson equation; i.e., (45) will be replaced by

$$\mathcal{E}(x,t) = -\nabla_x V = \lambda^{-2} \int_{\mathbb{R}^3} \frac{x-y}{|x-y|^3} \rho(y,t)\, dy. \tag{51}$$

We define Problem (K) as the system (42)–(44), (51) with (36), (37), (47)–(49), (40), and (41), subject to the initial conditions

$$f_n(x,k,0) = f_{n,I}(x,k), \quad f_p(x,k,0) = f_{p,I}(x,k), \quad f_{tr}(x,E,0) = f_{tr,I}(x,E).$$

We start by stating our assumptions on the data. For the velocities we assume

$$v_n, v_p \in W^{1,\infty}_{per}(B), \tag{52}$$

where here and in the following, the subscript *per* denotes Sobolev spaces of functions of k satisfying periodic boundary conditions on ∂B. Further we assume that the cross sections satisfy

$$\widetilde{\Phi}_n, \widetilde{\Phi}_p \geq 0, \quad \widetilde{\Phi}_n, \widetilde{\Phi}_p \in W^{1,\infty}_{per}(B \times B), \tag{53}$$

and

$$\Phi_n, \Phi_p \geq 0, \quad \Phi_n, \Phi_p \in W^{1,\infty}_{per}(B \times (0,1)). \tag{54}$$

A finite total number of trapped states is assumed:

$$M_{tr} \geq 0, \quad M_{tr} \in W^{1,\infty}(\mathbb{R}^3 \times (0,1)) \cap W^{1,1}(\mathbb{R}^3 \times (0,1)).$$

The L^1-assumption with respect to x is needed for controlling the total number of generated particles. For the initial data we assume

$$\begin{aligned} &0 \leq f_{n,I}, f_{p,I} \leq 1, \quad f_{n,I}, f_{p,I} \in W^{1,\infty}_{per}(\mathbb{R}^3 \times B) \cap W^{1,1}_{per}(\mathbb{R}^3 \times B), \\ &0 \leq f_{tr,I} \leq 1, \quad f_{tr,I} \in W^{1,\infty}_{per}(\mathbb{R}^3 \times (0,1)). \end{aligned} \tag{55}$$

We also assume

$$n_0, p_0 \in L^\infty((0,1)), \qquad C \in W^{1,\infty}(\mathbb{R}^3) \cap W^{1,1}(\mathbb{R}^3). \tag{56}$$

Finally, we need an upper bound for the life time of trapped electrons:

$$\int_B (\Phi_n \min\{1, n_0 M_n\} + \Phi_p \min\{1, p_0 M_p\})\, dk \geq \gamma > 0. \tag{57}$$

The reason for the various differentiability assumptions above is that we shall construct smooth solutions by an approach along the lines of [13], which goes back to [4].

An essential tool relies on the following potential theory estimates:

$$\|\mathcal{E}\|_{L^\infty(\mathbb{R}^3)} \leq C \|\rho\|^{1/2}_{L^1(\mathbb{R}^3)} \|\rho\|^{1/2}_{L^\infty(\mathbb{R}^3)}, \tag{58}$$

$$\|\nabla_x \mathcal{E}\|_{L^\infty(\mathbb{R}^3)} \leq C\big(1 + \|\rho\|_{L^1(\mathbb{R}^3)} + \|\rho\|_{L^\infty(\mathbb{R}^3)} \big[1 + \log(1 + \|\nabla_x \rho\|_{L^\infty(\mathbb{R}^3)})\big]\big). \tag{59}$$

This kind of estimate was already crucial in [17]; for the sake of completeness, we recall the proof in the appendix. We start by rewriting the collision and recombination-generation operators as

$$Q_i(f_i) = a_i[f_i](1 - f_i) - b_i[f_i] f_i, \quad i = n, p,$$

and

$$Q_{i,r}(f_i, f_{tr}) = g_i[f_{tr}](1 - f_i) - r_i[f_{tr}] f_i, \quad i = n, p,$$

with

$$\begin{aligned} a_i[f_i] &= \int_B \widetilde{\Phi}_i M_i f_i'\, dk', \quad b_i[f_i] = \int_B \widetilde{\Phi}_i M_i' (1 - f_i')\, dk', \quad i = n, p, \\ g_n[f_{tr}] &= \int_0^1 \Phi_n n_0 M_n f_{tr} M_{tr}\, dE, \quad g_p[f_{tr}] = \int_0^1 \Phi_p p_0 M_p (1 - f_{tr}) M_{tr}\, dE, \\ r_n[f_{tr}] &= \int_0^1 \Phi_n (1 - f_{tr}) M_{tr}\, dE, \quad r_p[f_{tr}] = \int_0^1 \Phi_p f_{tr} M_{tr}\, dE. \end{aligned}$$

In order to construct an approximating sequence $(f_n^j, f_p^j, f_{tr}^j, \mathcal{E}^j)$ we begin with

$$f_i^0(x,k,t) = f_{i,I}(x,k), \quad i = n,p, \quad f_{tr}^0(x,E,t) = f_{tr,I}(x,E). \tag{60}$$

The field always satisfies

$$\mathcal{E}^j(x,t) = \int_{\mathbb{R}^3} \frac{x-y}{|x-y|^3} \rho^j(y,t)\, dy. \tag{61}$$

Let $(f_n^j, f_p^j, f_{tr}^j, \mathcal{E}^j)$ be given. Then the ${f_i}^{j+1}$ are defined as the solutions of the following problem:

(62)
$$\begin{aligned}
&\alpha^2 \partial_t f_n^{j+1} + \alpha v_n(k) \cdot \nabla_x f_n^{j+1} - \alpha \mathcal{E}^j \cdot \nabla_k f_n^{j+1} \\
&\qquad = (a_n[f_n^j] + \alpha^2 g_n[f_{tr}^j])(1 - f_n^{j+1}) - (b_n[f_n^j] + \alpha^2 r_n[f_{tr}^j]) f_n^{j+1}, \\
&\alpha^2 \partial_t f_p^{j+1} + \alpha v_p(k) \cdot \nabla_x f_p^{j+1} + \alpha \mathcal{E}^j \cdot \nabla_k f_p^{j+1} \\
&\qquad = (a_p[f_p^j] + \alpha^2 g_p[f_{tr}^j])(1 - f_p^{j+1}) - (b_p[f_p^j] + \alpha^2 r_p[f_{tr}^j]) f_p^{j+1}, \\
&\varepsilon \partial_t f_{tr}^{j+1} = (p_0 \lambda_p[M_p(1 - f_p^j)] + \lambda_n[f_n^j])(1 - f_{tr}^{j+1}) - (n_0 \lambda_n[M_n(1 - f_n^j)] + \lambda_p[f_p^j]) f_{tr}^{j+1},
\end{aligned}$$

subject to the initial conditions

(63)
$$f_n^{j+1}(x,k,0) = f_{n,I}(x,k), \quad f_p^{j+1}(x,k,0) = f_{p,I}(x,k), \quad f_{tr}^{j+1}(x,E,0) = f_{tr,I}(x,E).$$

For the iterative sequence we state the following lemma, which is very similar to Proposition 3.1 from [13].

LEMMA 5.1. *Let the assumptions* (52)–(56) *be satisfied. Then the sequence* $(f_n^j, f_p^j, f_{tr}^j, \mathcal{E}^j)$ *defined by* (60)–(63) *satisfies the following for any time* $T > 0$:

(a) $0 \le {f_i}^j \le 1$, $i = n,p,tr$.

(b) f_n^j *and* f_p^j *are uniformly bounded with respect to* $j \to \infty$ *and* $\varepsilon \to 0$ *in* $L^\infty(0,T; L^1(\mathbb{R}^3 \times B))$.

(c) $\mathcal{E}^j$ *is uniformly bounded with respect to* j *and* ε *in* $L^\infty((0,T) \times \mathbb{R}^3)$.

Proof. The first two equations in (62) are standard linear transport equations, and the third equation is a linear ODE. Existence and uniqueness for the initial value problems is therefore a standard result.

Note that the a_i, b_i, g_i, r_i, and λ_i in (62) are nonnegative if we assume that (a) holds for j. Then (a) for $j+1$ is an immediate consequence of the maximum principle.

To estimate the L^1-norms of the distributions, we integrate the first equation in (62) and obtain

$$\|f_n^{j+1}\|_{L^1(\mathbb{R}^3 \times B)} \le \|f_{n,I}\|_{L^1(\mathbb{R}^3 \times B)} + \int_0^t \left\| a_n[f_n^j] \frac{1}{\alpha^2} + g_n[f_{tr}^j] \right\|_{L^1(\mathbb{R}^3 \times B)} (s)\, ds. \tag{64}$$

The boundedness of $\widetilde{\Phi}_n$, Φ_n, and f_{tr}^j and the integrability of M_{tr} imply

$$\|a_n[f_n^j]\|_{L^1(\mathbb{R}^3 \times B)} \le C \|f_n^j\|_{L^1(\mathbb{R}^3 \times B)}, \quad \|g_n[f_{tr}^j]\|_{L^1(\mathbb{R}^3 \times B)} \le C. \tag{65}$$

This is now used in (64). Then an estimate is derived for f_n^j by replacing $j+1$ by j and using the Gronwall inequality. Finally, it is easily seen that this estimate is passed from j to $j+1$ by (64). An analogous argument for f_p^j completes the proof of (b).

A uniform-in-ε ($L^1 \cap L^\infty$)-bound for the total charge density $\rho^j = n^j + \varepsilon n_{tr}^j - p^j - C$ follows from (b) and from the integrability of M_{tr}. Statement (c) of the lemma is now a consequence of (58). ∎

For passing to the limit in the nonlinear terms some compactness is needed. Therefore we prove uniform smoothness of the approximating sequence.

LEMMA 5.2. *Let the assumptions* (52)–(57) *be satisfied. Then for any time* $T > 0$ *the following hold:*

(a) f_n^j *and* f_p^j *are uniformly bounded with respect to* j *and* ε *in* $L^\infty(0,T; W^{1,1}_{per}(\mathbb{R}^3 \times B) \cap W^{1,\infty}_{per}(\mathbb{R}^3 \times B))$.

(b) f_{tr}^j *is uniformly bounded with respect to* j *and* ε *in* $L^\infty(0,T; W^{1,\infty}(\mathbb{R}^3 \times (0,1)))$.

(c) $\mathcal{E}^j$ *is uniformly bounded with respect to* j *and* ε *in* $L^\infty(0,T; W^{1,\infty}(\mathbb{R}^3))$.

Proof. We start by introducing $\nu^j = \nabla_{x,k} f_n^j = (\nu_x^j, \nu_k^j)$, $\pi^j = \nabla_{x,k} f_p^j = (\pi_x^j, \pi_k^j)$, $\phi^j = \nabla_x f_{tr}^j$, and by differentiating the last equation in (62) with respect to x:

$$\begin{aligned}\varepsilon \partial_t \phi^{j+1} = {} & (-p_0 \lambda_p[M_p \pi_x^j] + \lambda_n[\nu_x^j])(1 - f_{tr}^{j+1}) - (-n_0 \lambda_n[M_n \nu_x^j] + \lambda_p[\pi_x^j]) f_{tr}^{j+1} \\ & - (p_0 \lambda_p[M_p(1 - f_p^j)] + \lambda_n[f_n^j] + n_0 \lambda_n[M_n(1 - f_n^j)] + \lambda_p[f_p^j])\phi^{j+1}.\end{aligned}$$

The coefficient of ϕ^{j+1} on the right-hand side is bounded from below by the term appearing in assumption (57) and, thus, bounded away from zero. The maximum principle implies

$$\sup_{(0,t)} \|\phi^{j+1}\|_\infty \le C \left(\sup_{(0,t)} \|\nu_x^j\|_\infty + \sup_{(0,t)} \|\pi_x^j\|_\infty + 1 \right),$$

where here and in the following we use the symbol $\|\cdot\|_\infty$ for the L^∞-norm on $\mathbb{R}^3$, on $\mathbb{R}^3 \times B$, and on $\mathbb{R}^3 \times (0,1)$. The gradient of the first equation in (62) with respect to x and k can be written as

$$\alpha^2 \partial_t \nu^{j+1} + \alpha v_n \cdot \nabla_x \nu^{j+1} - \alpha \mathcal{E}^j \cdot \nabla_k \nu^{j+1} + (a_n + b_n + \alpha^2 g_n + \alpha^2 r_n)\nu^{j+1} = S_n^j,$$

where it is easily seen that, using our assumptions,

$$\|S_n^j\|_\infty \le C \left(1 + \|\nu^j\|_\infty + \|\phi^j\|_\infty + \|\nu^{j+1}\|_\infty (1 + \|\nabla_x \mathcal{E}^j\|_\infty)\right)$$

holds. The analogous treatment of the second equation in (62), the potential theory inequality (59), and the definition

$$\gamma^j(t) = \sup_{(0,t)} (\|\nu^j\|_\infty + \|\pi^j\|_\infty + \|\phi^j\|_\infty)$$

lead to

$$\gamma^{j+1} \le C \left(1 + \int_0^t (\gamma^j + \gamma^{j+1}(1 + \log(1 + \gamma^j)))\, ds \right),$$

implying boundedness of γ^j on arbitrary bounded time intervals (as in [4]). This proves (c) and the L^∞-part of (a). The equation for $\partial_E f_{tr}^{j+1}$ can be treated as above, completing the proof of (b).

By $\int_{\mathbb{R}^3} n_{tr}\, dx \le \int_{\mathbb{R}^3} M_{tr}\, dx$, it is trivial that the total number of trapped electrons is bounded. Therefore, the L^1-estimates in (a) follow the line of [13] since no coupling with the equation for the trapped electrons is necessary. ∎

With the previous results, the first two equations in (62) also give uniform bounds for the time derivatives of f_n^j and f_p^j. Thus, subsequences converge strongly locally in x and t. In the same way, the right-hand side of the time derivative of the Poisson equation is bounded in L^1 and in L^∞, and (58) implies boundedness of the time derivative of the field. So the field also converges strongly. This and the (obvious) weak convergence of f_{tr}^j are sufficient for passing to the limit in the quadratic nonlinearities. Note also that we have enough bounds on the time derivative to define the trace at time $t = 0$. Existence of a global solution of Problem (K) follows. By the same argument, the limit $\varepsilon \to 0$ can be justified, since all estimates are also uniform in ε.

THEOREM 5.3. *Let the assumptions* (52)–(57) *be satisfied. Then Problem* (K) *has a global solution* $(f_n, f_p, f_{tr}, \mathcal{E})$ *with* $f_n, f_p \in L^\infty(0,T; W^{1,\infty}_{per}(\mathbb{R}^3 \times B))$, $f_{tr} \in L^\infty(0,T; W^{1,\infty}(\mathbb{R}^3 \times (0,1)))$, $\mathcal{E} \in L^\infty(0,T; W^{1,\infty}(\mathbb{R}^3))$. *For* $\varepsilon \to 0$, *a subsequence of solutions converges to the formal limit problem. The convergence of* f_n *and* f_p *is in* $L^\infty((0,\infty) \times \mathbb{R}^3 \times B)$, *that of* $\mathcal{E}$ *in* $L^\infty((0,\infty) \times \mathbb{R}^3)$, *and that of* f_{tr} *in* $L^\infty((0,\infty) \times \mathbb{R}^3 \times (0,1))$ *weak*.*

6. Relation between macroscopic and kinetic models. In this section the relation between the two models in sections 2 and 4 is clarified on a formal level. The drift-diffusion model of section 2 can be derived from the kinetic model of section 4 by two simplification steps: a macroscopic and a low density limit.

Starting with the macroscopic limit, i.e., the limit when the Knudsen number α tends to zero in (42), (43), the solutions are expanded in terms of powers of α:

$$f_n = f_n^0 + \alpha f_n^1 + \mathcal{O}(\alpha^2), \qquad f_p = f_p^0 + \alpha f_p^1 + \mathcal{O}(\alpha^2), \tag{66}$$

$$f_{tr} = f_{tr}^0 + \mathcal{O}(\alpha), \qquad V = V^0 + \mathcal{O}(\alpha). \tag{67}$$

The limit of (42), (43) as $\alpha \to 0$ leads to $Q_n(f_n^0) = Q_p(f_p^0) = 0$. With the (frequently used) simplifying assumption that the cross sections $\widetilde{\Phi}_n$ and $\widetilde{\Phi}_p$ are strictly positive, the limiting distributions are of Fermi–Dirac type (see [13]):

$$f_n^0(x,k,t) = \frac{1}{1 + e^{-\mu_n(x,t)}/M_n(k)}, \qquad f_p^0(x,k,t) = \frac{1}{1 + e^{\mu_p(x,t)}/M_p(k)},$$

where the scaled Maxwellians M_n, M_p are given by (46) and the chemical potentials μ_n and μ_p are yet to be specified. Note the one-to-one relations between the chemical potentials and the macroscopic electron and hole densities:

$$n(\mu_n) = \int_B \frac{dk}{1 + e^{-\mu_n}/M_n(k)}, \qquad p(\mu_p) = \int_B \frac{dk}{1 + e^{\mu_p}/M_p(k)}.$$

Now (42) is divided by α, and then again the limit $\alpha \to 0$ is carried out (formally):

$$v_n \cdot \nabla_x f_n^0 + \nabla_x V^0 \cdot \nabla_k f_n^0 = LQ_n(f_n^0) f_n^1, \tag{68}$$

where LQ_n is the linearization of Q_n:

$$LQ_n(f_n^0) f_n^1 = \int_B \widetilde{\Phi}_n \big[(M_n(1 - f_n^0) + M_n' f_n^0) f_n^{1'} - (M_n f_n^{0'} + M_n'(1 - f_n^{0'})) f_n^1 \big] \, dk'.$$

For the following we shall need two facts about the linearized collision operator $LQ_n(f_n^0)$ (see, e.g., [1]): It has a one-dimensional kernel spanned by $f_n^0(1 - f_n^0)$, and

its range consists of functions whose integral with respect to k vanishes. Therefore, for solvability of (68), seen as an equation for f_n^1, the integral with respect to k of the left-hand side has to vanish. This is obvious for the second term $\nabla_x V^0 \cdot \nabla_k f_n^0$ by the periodicity with respect to k. Since the first term can be written as

$$v_n \cdot \nabla_x f_n^0 = \nabla_k \varepsilon_n \cdot \nabla_x \frac{M_n}{M_n + e^{-\mu_n}} = -\nabla_k \cdot \nabla_x \log\left(M_n + e^{-\mu_n}\right),$$

it also satisfies the solvability condition. Now (68) is written as

$$\frac{M_n e^{-\mu_n}}{(M_n + e^{-\mu_n})^2} \nabla_k \varepsilon_n \cdot (\nabla_x V^0 - \nabla_x \mu_n) = LQ_n(f_n^0) f_n^1. \tag{69}$$

Note that the factor in parentheses is independent of k. Thus, choosing a solution $h_n(k, \mu_n)$ of

$$LQ_n(f_n^0) h_n = -\frac{M_n e^{-\mu_n}}{(M_n + e^{-\mu_n})^2} \nabla_k \varepsilon_n, \tag{70}$$

the solution of (69) can be written as

$$f_n^1 = h_n(k, \mu_n) \cdot (\nabla_x V^0 - \nabla_x \mu_n) + \mu_n^1 f_n^0 (1 - f_n^0).$$

Analogously,

$$f_p^1 = h_p(k, \mu_p) \cdot (\nabla_x V^0 + \nabla_x \mu_p) + \mu_p^1 f_p^0 (1 - f_p^0) \tag{71}$$

is obtained (with $\mu_n^1(x,t)$ and $\mu_p^1(x,t)$ not specified, and not needed in the following). Finally, (42), (43) are divided by α^2 and integrated with respect to k, and the limit $\alpha \to 0$ is carried out:

$$\partial_t n + \nabla_x \cdot \int_B v_n f_n^1 \, dk = \int_B Q_{n,r}(f_n^0, f_{tr}^0) \, dk = \int_0^1 S_n(f_n^0, f_{tr}^0) \, dE, \tag{72}$$

$$\partial_t p + \nabla_x \cdot \int_B v_p f_p^1 \, dk = \int_B Q_{p,r}(f_p^0, f_{tr}^0) \, dk = \int_0^1 S_p(f_p^0, f_{tr}^0) \, dE. \tag{73}$$

With the formulas for f_n^1 and f_p^1, we obtain the drift-diffusion fluxes

$$\int_B v_n f_n^1 \, dk = D_n(\mu_n)(\nabla_x V^0 - \nabla_x \mu_n), \qquad \int_B v_p f_p^1 \, dk = D_p(\mu_p)(\nabla_x V^0 + \nabla_x \mu_p),$$

with the diffusion matrices

$$D_n = \int_B v_n \otimes h_n \, dk, \qquad D_p = \int_B v_p \otimes h_p \, dk.$$

For the recombination-generation terms, we obtain

$$S_n(f_n^0, f_{tr}^0) = \lambda_n \left[\frac{e^{-\mu_n}}{1 + e^{-\mu_n}/M_n}\right] (n_0 f_{tr}^0 - e^{\mu_n}(1 - f_{tr}^0)),$$

$$S_p(f_p^0, f_{tr}^0) = \lambda_p \left[\frac{e^{\mu_p}}{1 + e^{\mu_p}/M_p}\right] (p_0(1 - f_{tr}^0) - e^{-\mu_p} f_{tr}^0).$$

Finally, we consider the small densities situation, when μ_n is large and negative and μ_p large and positive. This gives $n(\mu_n) \approx e^{\mu_n}$ and $p(\mu_p) \approx e^{-\mu_p}$. The above recombination-generation terms can then be approximated by the terms in (9) with $1/\tau_n = \lambda_n[M_n]$ and $1/\tau_p = \lambda_p[M_p]$.

Equation (70) for h_n can be approximated by

$$\int_B \widetilde{\Phi}_n \big[M_n h_n' - M_n' h_n\big]\, dk' = -nM_n \nabla_k \varepsilon_n,$$

implying $h_n = n\tilde{h}_n(k)$ and, thus, $D_n = n\tilde{D}_n$. With this and the analogous approximation for holes, the macroscopic model becomes the drift-diffusion model from section 2.

It is worth pointing out that the drift-diffusion SRH model has been obtained from the kinetic model by a two-step approximation procedure: At first, the hydrodynamic limit leads to a more nonlinear system, and we perform additionally the small densities asymptotics. This remark appeals to further mathematical questions:

• It could be interesting to investigate the intermediate macroscopic model that comes directly from the Fermi–Dirac statistics.

• It could be tempting to reverse the limits. Roughly speaking, it means that we do not take into account the Pauli exclusion principle in the kinetic equations, and the collision operator is replaced by a linear Boltzmann operator which relaxes to a Maxwellian (instead of a Fermi–Dirac distribution). Mathematically, this leads to additional difficulties since we lose the natural L^∞-estimate given for free with the exclusion terms. Rigorous derivation of the diffusion regime for the corresponding Boltzmann–Poisson system in a bounded domain, with only one species of charged particles, has been obtained only very recently by using a tricky renormalization argument; see [11] (and [12] for an earlier work on renormalized solutions).

Appendix. Proof of (58) and (59). We recall that the fundamental solution of $-\Delta$ in $\mathbb{R}^N$, $N \geq 3$, reads $E(x) = C_N|x|^{2-N}$. For a given function $\rho : \mathbb{R}^N \to \mathbb{R}^+$, we set

$$\Phi = E * \rho, \qquad \nabla_x \Phi(x) = C_N(2-N) \int_{\mathbb{R}^N} \frac{x-y}{|x-y|}\, \frac{\rho(y)}{|x-y|^{N-1}}\, dy.$$

For any $0 < R < \infty$, we have

$$\int_{\mathbb{R}^N} \frac{\rho(y)}{|x-y|^{N-1}}\, dy = \int_{|x-y|\leq R} \dots\, dy + \int_{|x-y|\geq R} \dots\, dy \leq \|\rho\|_\infty \frac{\Omega_N R^2}{2} + \frac{1}{R^{N-1}} \|\rho\|_1,$$

where Ω_N stands for the surface of the N-dimensional sphere. Optimizing with respect to R yields

$$\|\nabla \Phi\|_\infty \leq K_N \|\rho\|_\infty^{(N-1)/(N+1)} \|\rho\|_1^{2/(N+1)},$$

where K_N is the constant depending only on the dimension.

Since $|x|^{N-1}$ is locally integrable, we compute the second derivatives of the potential as follows. For any $\varphi \in C_c^\infty(\mathbb{R}^N)$, we have

$$\big\langle \partial_{ij}^2 \Phi; \varphi \big\rangle = C_N(2-N) \int_{\mathbb{R}^N} \frac{x_j}{|x|^N}\, \partial_j \varphi(x)\, dx = C_N(2-N) \lim_{\eta \to 0} \int_{|x|\geq \eta} \frac{x_j}{|x|^N}\, \partial_j \varphi(x)\, dx$$

$$= C_N(N-2) \lim_{\eta \to 0} \left(\int_{|x|\geq \eta} \left(\frac{\delta_{ij}}{|x|^N} - N \frac{x_i x_j}{|x|^{N+2}} \right) \varphi(x)\, dx + \int_{|x|=\eta} \frac{x_j}{|x|^N} \frac{x_i}{|x|} \varphi(x)\, d\sigma(x) \right).$$

The second integral in the right-hand side can be recast as

$$\int_{\mathbb{S}^{N-1}} \varphi(\eta\omega)\omega_i\omega_j \, d\omega,$$

and therefore it converges to

$$\frac{\Omega_N}{N}\delta_{ij}\,\varphi(0).$$

Let us introduce the matrix

$$\mathbb{K}_{ij}(x) = C_N \frac{N-2}{|x|^N}\left(\delta_{ij} - N\frac{x_i x_j}{|x|^2}\right).$$

Then, in the sense of distribution the Hessian matrix of E satisfies

$$D^2E(x) = C_N(2-N)\frac{\Omega_N}{N}\mathbb{I}\,\delta(x=0) + \lim_{\eta\to 0}\mathbb{K}(x)\chi_{|x|\geq\eta}.$$

In particular we remark that

$$\int_{\mathbb{S}^{N-1}} \mathbb{K}(r\omega)\, d\omega = 0, \qquad \operatorname{Tr}\mathbb{K} = 0. \tag{74}$$

Accordingly, the Hessian matrix of the potential Φ is given by

$$D^2\Phi(x) = C_N(2-N)\frac{\Omega_N}{N}\mathbb{I}\,\rho(x) + \lim_{\eta\to 0}\int_{|x-y|\geq\eta}\mathbb{K}(x-y)\rho(y)\,dy.$$

Let us discuss the last term. Consider $0<\eta<R_1<R_2<\infty$. Using the notation C to stand for any constant depending only on the dimension, we get

$$\begin{aligned}\left|\int_{|x-y|\geq\eta}\mathbb{K}(x-y)\rho(y)\,dy\right| &\leq C\left(\int_{|x-y|\geq R_2}\frac{\rho(y)}{|x-y|^N}\,dy + \int_{R_1\leq|x-y|\leq R_2}\frac{\rho(y)}{|x-y|^N}\,dy \right.\\ &\qquad \left. + \left|\int_{\eta\leq|x-y|\leq R_1}\mathbb{K}(x-y)\rho(y)\,dy\right|\right)\\ &\leq C\frac{1}{R_2^N}\|\rho\|_1 + C\|\rho\|_\infty \ln\left(\frac{R_2}{R_1}\right) + C\left|\int_{\eta\leq|x-y|\leq R_1}\mathbb{K}(x-y)\big(\rho(y)-\rho(x)\big)\,dy\right|\end{aligned}$$

where we used (74). We deduce the following estimate:

$$\left|\int_{|x-y|\geq\eta}\mathbb{K}(x-y)\rho(y)\,dy\right| \leq C\left(\frac{1}{R_2^N}\|\rho\|_1 + \|\rho\|_\infty\ln\left(\frac{R_2}{R_1}\right) + \|\nabla_x\rho\|_\infty R_1\right).$$

Then, we choose $R_2 = 1 > R_1 = (1+\|\nabla_x\rho\|_\infty)^{-1}$ and conclude that

$$|D^2\Phi(x)| \leq C\Big(1 + \|\rho\|_1 + \|\rho\|_\infty\big(1+\ln(1+\|\nabla_x\rho\|_\infty)\big)\Big)$$

holds. A similar analysis can be done in dimension two; see [17].

Acknowledgment. Part of this work was carried out while the second and third authors enjoyed the hospitality of the Université des Sciences et Technologies Lille 1.

REFERENCES

[1] N. Ben Abdallah and P. Degond, *On a hierarchy of macroscopic models for semiconductors*, J. Math. Phys., 37 (1996), pp. 3306–3333.

[2] I. Choquet, P. Degond, and C. Schmeiser, *Energy-transport models for charge carriers involving impact ionization in semiconductors*, Transport Theory Statist. Phys., 32 (2003), pp. 99–132.

[3] I. Choquet, P. Degond, and C. Schmeiser, *Hydrodynamic model for charge carriers involving strong ionization in semiconductors*, Commun. Math. Sci., 1 (2003), pp. 74–86.

[4] P. Degond, *Global existence of smooth solutions for the Vlasov-Fokker-Planck equation in* 1 *and* 2 *space dimensions*, Ann. Sci. École Norm. Sup. (4), 19 (1986), pp. 519–542.

[5] H. Gajewski, *On existence, uniqueness, and asymptotic behaviour of solutions of the basic equations for carrier transport in semiconductors*, Z. Angew. Math. Mech., 65 (1985), pp. 101–108.

[6] I. Gasser, C. D. Levermore, P. A. Markowich, and C. Schmeiser, *The initial time layer problem and the quasineutral limit in the semiconductor drift-diffusion model*, European J. Appl. Math., 12 (2001), pp. 497–512.

[7] P. González, A. Godoy, F. Gámiz, and J. A. Carrillo, *Accurate deterministic numerical simulation of p-n junctions*, J. Comput. Electron., 3 (2004), pp. 235–238.

[8] P. González, J. A. Carrillo, and F. Gámiz, *Deterministic numerical simulation of* $1d$ *kinetic descriptions of bipolar electron devices*, in Scientific Computing in Electrical Engineering, Math. Ind. 9, Springer, Berlin, Heidelberg, 2006, pp. 339–344.

[9] R. N. Hall, *Electron-hole recombination in Germanium*, Phys. Rev., 87 (1952), p. 387.

[10] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer, Vienna, New York, 1990.

[11] N. Masmoudi and M. L. Tayeb, *Diffusion limit of a semiconductor Boltzmann–Poisson system*, SIAM J. Math. Anal., 38 (2007), pp. 1788–1807.

[12] S. Mischler, *On the initial boundary value problem for the Vlasov-Poisson-Boltzmann system*, Comm. Math. Phys., 210 (2000), pp. 447–466.

[13] F. Poupaud, *On a system of nonlinear Boltzmann equations of semiconductor physics*, SIAM J. Appl. Math., 50 (1990), pp. 1593–1606.

[14] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer, Vienna, New York, 1984.

[15] W. Shockley and W. T. Read, *Statistics of the recombinations of holes and electrons*, Phys. Rev., 87 (1952), pp. 835–842.

[16] J. Simon, *Compact sets in the space* $L^p(0,T;B)$, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.

[17] S. Ukai and T. Okabe, *On the classical solution in the large time of the two dimensional Vlasov equation*, Osaka J. Math., 15 (1978), pp. 245–261.

SIAM J. APPL. MATH.
Vol. 67, No. 4, pp. 1202–1212

ROLL-WAVES IN GENERAL HYPERBOLIC SYSTEMS WITH SOURCE TERMS*

PASCAL NOBLE†

Abstract. The purpose of this article is to prove the existence of particular nonlinear waves, so-called roll-waves, in general hyperbolic systems with a source term: these are periodic and discontinuous traveling waves, the discontinuities satisfying entropy conditions. We show that roll-wave solutions can be seen as zeros of a suitably chosen map. In the vanishing amplitude limit, a particular solution exists, and we prove that this solution persists to small amplitudes, which yields the existence of small amplitude roll-waves. The shock conditions (Rankine–Hugoniot conditions and Lax entropy conditions) are treated as nonlinear boundary conditions.

Key words. nonlinear waves, hyperbolic systems, roll-waves, relaxation

AMS subject classifications. 35L67, 35L60, 76H05

DOI. 10.1137/060672248

1. Introduction.

Roll-waves are hydrodynamic instabilities appearing in shallow water flows downstream from an open inclined channel. This type of flow is described by the Saint Venant equation, which reads in its adimensional form as:

$$\begin{aligned} &h_t + (hu)_x = 0, \\ &(hu)_t + \left(\frac{h^2}{2F} + hu^2\right)_x = h - u^2, \end{aligned} \tag{1}$$

where h is the height of the water, u its velocity, and F is the Froude number. This system is *hyperbolic*, the hyperbolic part being similar to the isentropic Euler equations, with a *source term* taking into account the gravity effects and the rugosity of the channel (through the friction term $-u^2$). When the uniform flow is unstable $F > 4$, it is proved that roll-waves appear as mathematical solutions of (1). More precisely, Dressler [4] proved the existence of periodic traveling waves, which are necessarily discontinuous. The discontinuities satisfy the Rankine–Hugoniot jump conditions and a Lax entropy condition.

In this paper, we consider a general hyperbolic system with a source term

$$u_t + Df(u)\,u_x = g(u), \quad u \in \mathbb{R}^n,\ n \geq 2. \tag{2}$$

For this general class of equations, we study the existence of roll-wave solutions, i.e., discontinuous periodic traveling waves, the discontinuities satisfying the Rankine–Hugoniot jump conditions and the Lax shock conditions.

There are many mathematical studies of the system (2) in the *relaxation case*, where equation (2) has the particular form

$$u_t + Df(u)\,u_x = \frac{1}{\epsilon}\,g(u), \tag{3}$$

*Received by the editors October 13, 2006; accepted for publication (in revised form) February 13, 2007; published electronically June 12, 2007.
http://www.siam.org/journals/siap/67-4/67224.html

†CNRS, UMR 5208 Institut Camille Jordan, Université de Lyon, Université Lyon 1, Batiment du Doyen Jean Braconnier, 43 Blvd du 11 Novembre 1918, F-69622 Villeurbanne Cedex, France (noble@math.univ-lyon1.fr).

with $\epsilon \ll 1$, and the set of equilibrium $g(u) = 0$ forms a j-dimensional manifold, $j < n$. This type of system appears in many physical situations, for example, in kinetic theory [2], gases not in local equilibrium [7], multiphase and phase transition [17], or linear and nonlinear waves [19]. Let us mention a few mathematical results out of a huge literature on the topic. For relaxation equations, reduced systems, inviscid and viscous local conservation laws, and weakly nonlinear limits can be computed through asymptotic expansions. These computations can be justified rigorously for 2×2 systems [3], and in the general case an entropy condition on the full system ensures the hyperbolicity of the reduced inviscid system. For 2×2 systems, this condition is nothing but the stability condition of the equilibrium states. In that case, the first order correction is shown to be dissipative. More recently, the existence and the stability of relaxation shocks have been established in general hyperbolic systems with relaxation [18, 15].

Less studied is the system (2) when the uniform flow is *unstable*: let us just mention the paper of Jin and Katsoulakis [6], where 2×2 hyperbolic systems with supercharacteristic relaxations are considered: a weakly nonlinear limit is identified which is a Burgers equation with a source term, and this equation possesses roll-wave solutions. Such a limit is justified in the presence of artificial viscosity, using the energy method [6]. For the Saint Venant system (1), when the uniform flow is unstable $F > 4$, roll-waves appear and are proved to be stable. This is done formally in [1] for roll-waves of large spatial period, and rigorously in [10], where it is proved that roll-waves are spectrally stable. The nonlinear stability of roll-waves shall be treated in a forthcoming paper [11]. The notion of nonlinear stability of solutions of hyperbolic equations that contain shocks means that for any initial data close to the solution and with the same structure of shocks, the solution of the Cauchy problem exists on a sufficiently small interval and keeps the structure of the initial data; see [14, 16] and the references therein for more details.

In this paper we are going to prove that under suitable conditions system (2) possesses small amplitude roll-wave solutions. For that purpose, we follow the approach of Dressler and search for discontinuous periodic traveling waves, the discontinuities satisfying the Rankine–Hugoniot conditions and the Lax shock entropy condition. We shall see that this is equivalent to proving the existence of a special regular solution of the differential system

$$\big(Df(u) - c\big)u' = g(u) \quad \forall x \in (0, L),$$

where c is the wave speed and L the wavelength of the roll-wave, satisfying nonlinear boundary conditions. This approach has already been used successfully to prove the existence of pulsating roll-waves in Saint Venant equations with a periodic bottom [12].

In order to prove the existence of roll-waves, we show that they are zeros of a suitably chosen map. Then, searching for small amplitude roll-waves, we introduce the scaling used by Jin and Katsoulakis to derive a Burgers equation from 2×2 hyperbolic systems with supercharacteristic relaxation [6]. In the vanishing amplitude limit $\epsilon \to 0$ and $L = 2\epsilon\tau \to 0$, we prove that the map considered has a zero. Then we show that the differential of this map is invertible and, using the implicit function theorem, deduce the existence of small amplitude roll-waves. The plan of the paper is as follows. In section 2, we reduce the problem of finding roll-waves to finding zeros of a particular map. These zeros satisfy a differential system and the Rankine–Hugoniot conditions. The Lax shock condition is treated separately. In section 3, we show that the map has a zero in the limit $\epsilon \to 0$ and $L = 2\epsilon\tau \to 0$. In section 4, we prove that

under suitable conditions, this particular solution persists when $0 < \epsilon \ll 1$. We shall separate two cases. We first consider the "artificial" case where $dg(\overline{u}_0)$ is invertible, $\overline{u}_0$ being a stationary solution of (2) (i.e., such that $g(\overline{u}_0) = 0$): we obtain a family of roll-waves parametrized by the rescaled wavelength $2\tau > 0$. Then we consider the "real" case where $g(u) =^t (0, h(u))$ with $h : \mathbb{R}^n \to \mathbb{R}^{n-j}$. Then assuming that $dh(u_0)$ is surjective, we can prove that the particular solution persists when $\epsilon \neq 0$ and, if the rescaled period is fixed, belongs to a j-dimensional manifold. Under suitable conditions, we show that this family can be parametrized by j-conserved quantities of the motion. Then we draw a conclusion and present some perspectives on this paper.

2. Formulation of the problem. Let us consider the first order system of partial differential equations

$$u_t + Df(u)\,u_x = g(u). \tag{4}$$

The left-hand side of system (4) is supposed to be hyperbolic in a neighborhood $\mathcal{V}(\overline{u}_0)$ of a constant solution $u = \overline{u}_0$ of (4), i.e., so that $g(\overline{u}_0) = 0$. That is, $Df(u)$ has n real eigenvalues $\big(\lambda_k(u)\big)_{k=1,\dots,n}$ so that

$$\lambda_1(u) < \cdots < \lambda_k(u) < \cdots < \lambda_n(u) \qquad \forall u \in \mathcal{V}(\overline{u}_0). \tag{5}$$

We shall consider in what follows that f and g both have a power series expansion at the point $u = \overline{u}_0$ with a domain of convergence $\mathbb{D}$ containing $B(\overline{u}_0, r),\ r > 0$.

The purpose of the paper is to prove the existence of periodic traveling waves with admissible discontinuities given by the Rankine–Hugoniot conditions and a Lax entropy shock condition. Following the approach developed by Dressler [4], we search for a periodic traveling wave with wave speed c and wavelength $2L$ in the form $u(x,t) = U(x - ct)$, with U a $2L$ periodic function with discontinuities at points $x_j = (2j + 1)L$, $j \in \mathbb{Z}$, which satisfies the differential system

$$\big(Df(U(x)) - c\big)U' = g(U(x)) \qquad \forall x \in (-L, L). \tag{6}$$

The discontinuities verify the admissible conditions

$$\begin{aligned} &[f(U)]_{(2j+1)L} = c[U]_{(2j+1)L} \quad \forall j \in \mathbb{Z}, \\ &\lambda_k(U_j^+) < c < \lambda_k(U_j^-), \qquad \lambda_{k-1}(U_j^-) < c < \lambda_{k+1}(U_j^+), \end{aligned} \tag{7}$$

for some k, $1 \le k \le n$, and $U_j^\pm$ denotes $U((2j+1)L)^\pm$ $\forall j \in \mathbb{Z}$. Due to the translational invariance of (6), the problem (6)–(7) is equivalent to finding a smooth solution U defined on $(-L, L)$ satisfying the system (6) with the nonlinear boundary conditions

$$\begin{aligned} &f\big(U(L)\big) - f\big(U(-L)\big) = c\big(U(L) - U(-L)\big), \\ &\lambda_k\big(U(-L)\big) < c < \lambda_k\big(U(L)\big), \qquad \lambda_{k-1}\big(U(L)\big) < c < \lambda_{k+1}\big(U(-L)\big), \end{aligned} \tag{8}$$

for some k, $1 \le k \le n$. Now let us take U as the solution of (6) and integrate this system on $(-L, L)$: this yields

$$f\big(U(L)\big) - f\big(U(-L)\big) - c\big(U(L) - U(-L)\big) = \int_{-L}^{L} g(U(x))dx. \tag{9}$$

As a consequence, the Rankine–Hugoniot jump conditions for solutions in the class of roll-waves are satisfied if and only if

$$\int_{-L}^{L} g(U(x))dx = 0.$$

In what follows we are going to prove the existence of a *small amplitude* roll-wave with a *small spatial period*. For that purpose, we use the scaling introduced by Jin and Katsoulakis to derive a Burgers equation from 2×2 hyperbolic systems with supercharacteristic relaxation [6]. Denote by ϵ the amplitude of the roll-wave and by $L = 2\epsilon\tau$, with $\tau > 0$, its spatial wavelength. Let us write u in the form

$$u(x) = \overline{u} + \epsilon v\left(\frac{x - ct}{2\epsilon\tau}\right), \tag{10}$$

with v a smooth function defined on $(-1, 1)$. The resulting roll-wave is a traveling wave with wave speed c and spatial period $2\epsilon\tau$ for some $\tau > 0$. Note that the amplitude of the roll-wave and its spatial period are of the same order. Inserting (10) into (6), (8) yields

$$\big(Df(u + \epsilon v(x)) - c\big)v'(x) = \tau\, g(u + \epsilon v(x)) \qquad \forall x \in (-1, 1). \tag{11}$$

The nonlinear boundary conditions are given by

$$\begin{aligned} &\frac{f(\overline{u} + \epsilon v(1) - f(\overline{u} + \epsilon v(-1))}{\epsilon} = c(v(1) - v(-1), \\ &\lambda_k(\overline{u} + \epsilon v(-1)) < c < \lambda_k(\overline{u} + \epsilon v(1)), \\ &\lambda_{k-1}(\overline{u} + \epsilon v(1)) < c < \lambda_{k+1}(\overline{u} + \epsilon v(-1)), \end{aligned} \tag{12}$$

for some k, $1 \le k \le n$. It is a classical result that the Lax entropy shock conditions for small amplitude shocks reduce to the single condition

$$\lambda_k(\overline{u} + \epsilon v(-1)) < c < \lambda_k(\overline{u} + \epsilon v(1)), \tag{13}$$

the other condition being automatically satisfied when ϵ is sufficiently small. Moreover, letting $\epsilon \to 0$ in the shock conditions yields

$$Df(\overline{u})\big(v(1) - v(-1)\big) = c(v(1) - v(-1)).$$

Thus $c = \lambda_k(\overline{u})$ is necessarily one of the eigenvalues of $Df(\overline{u})$. In what follows, we suppose that $c = \lambda_k(\overline{u})$.

We are going to prove that there exists a solution $(\overline{u}, v)$ of (11,12) for $0 < \epsilon \ll 1$. For that purpose, let us consider the map $\mathcal{F}_{\epsilon,\tau}$ defined by

$$\begin{aligned} \mathcal{F}_{\epsilon,\tau} &: \mathbb{X} \times \mathbb{R}^n \to \mathbb{Y} \times \mathbb{R}^n, \\ \mathcal{F}_{\epsilon,\tau}(v, \overline{u})_1 &= \Pi_k(\overline{u})\left(\frac{Df(u + \epsilon v) - Df(\overline{u})}{\epsilon}v' - \tau\frac{g(\overline{u} + \epsilon v) - \langle g(u + \epsilon v)\rangle}{\epsilon}\right) \\ &\quad + (1 - \Pi_k(\overline{u}))\Big(\big(Df(u + \epsilon v) - \lambda_k(\overline{u})\big)v'\Big) \\ &\quad + \tau(1 - \Pi_k(\overline{u}))\big(\langle g(u + \epsilon v)\rangle - g(u + \epsilon v)\big), \\ \mathcal{F}_{\epsilon,\tau}(v, \overline{u})_2 &= \int_{-1}^{1} g(\overline{u} + \epsilon v(x))\, dx, \end{aligned}$$

where $\langle u\rangle$ denotes the spatial mean $\langle u\rangle = \frac{1}{2}\int_{-1}^{1} u(x)dx$ and $\Pi_k(\overline{u})$ is the projection on the eigenspace $Ker(Df(\overline{u}) - \lambda_k(\overline{u})I)$ with respect to $Im(Df(\overline{u}) - \lambda_k(\overline{u})I)$. The functional spaces $\mathbb{X}, \mathbb{Y}$ are given by

$$\mathbb{X}_0 = \left\{ f \in C^1(-1,1)/\, f(x) = \sum_{n\geq 0} a_n\, x^n,\ \sum_{n\geq 0}(n+1)|a_n| < +\infty \right\},$$

$$\mathbb{X} = \{ f \in \mathbb{X}_0/\, (1 - \Pi_k(\overline{u}_0))\langle f\rangle = 0\}, \tag{14}$$

$$\mathbb{Y} = \left\{ f \in C(-1,1)/\, f(x) = \sum_{n\geq 0} a_n\, x^n,\ \sum_{n\geq 0}|a_n| < +\infty \right\},$$

endowed with the norms $\|f\|_{\mathbb{X}} = \sum_{n\geq 0}(n+1)|a_n|$ and $\|f\|_{\mathbb{X}} = \sum_{n\geq 0}|a_n|$. The mapping $\mathcal{F}_{\epsilon,\tau}$ is well defined, provided that $v, \tau > 0$ are in a bounded subset of $\mathbb{X}\times\mathbb{R}_*^+$, ϵ is sufficiently small, and u is close enough to $\overline{u}_0$. Moreover, the eigenvalue $\lambda_k(\overline{u})$ of $Df(\overline{u})$ is isolated in a sufficiently small neighbourhood $\mathcal{V}(\overline{u}_0)$ of $\overline{u}_0$: it is standard perturbation theory to prove that the application $u \mapsto \Pi_k(\overline{u})$ is C^1 (see [5] for more details). Consequently, since we have supposed that f, g have power series expansions, the map F is C^1 for $\overline{u} \in \mathcal{V}(\overline{u}_0)$, v, and $\tau > 0$ in a bounded subset of $\mathbb{X}\times\mathbb{R}_*^+$ and ϵ sufficiently small. Since $Ker(Df(\overline{u}) - \lambda_k(\overline{u})I)$ is one-dimensional, we can identify, for any $v \in \mathbb{R}^n$, $\Pi_k(\overline{u})v$ with a real number.

The following proposition relates the zeros of $\mathcal{F}_{\epsilon,\tau}$ with the roll-wave solutions.

PROPOSITION 1. *Let $(\overline{u}, v) \in \mathbb{R}^n \times \mathbb{X}$ be a zero of $\mathcal{F}_{\epsilon,\tau}$; then*

$$\begin{aligned} &\big(Df(\overline{u} + \epsilon v(x)) - c\big)v'(x) = \tau\, g(u + \epsilon v(x)) \qquad \forall x \in (-1,1), \\ &\frac{f(\overline{u} + \epsilon v(1) - f(\overline{u} + \epsilon v(-1))}{\epsilon} = c\big(v(1) - v(-1)\big), \end{aligned} \tag{15}$$

with $c = \lambda_k(\overline{u})$.

Then a zero of $\mathcal{F}_{\epsilon,\tau}$ is a solution of (11) and the Rankine–Hugoniot jump conditions. In what follows, given any $\tau_0 > 0$, we are going to prove that $\mathcal{F}_{0,\tau_0}$ has a particular solution $(\overline{u}_0, v^0)$ and that the derivative at this point $D\mathcal{F}_{0,\tau_0}$ is invertible with a bounded inverse. Using the implicit function theorem, this particular solution *persists* as a zero of $\mathcal{F}_{\epsilon,\tau}$, provided that $\epsilon \approx 0$ and $\tau \approx \tau_0$.

Let us deal with the Lax shock condition: this condition is equivalent to

$$\frac{\lambda_k(\overline{u} + \epsilon v(-1)) - \lambda_k(\overline{u})}{\epsilon} < 0 < \frac{\lambda_k(\overline{u} + \epsilon v(1)) - \lambda_k(\overline{u})}{\epsilon}. \tag{16}$$

Let $\epsilon \to 0$; then

$$d\lambda_k(\overline{u}).v(-1) \leq 0 \leq d\lambda_k(\overline{u}).v(1). \tag{17}$$

It is easy to prove the following lemma

LEMMA 1. *Suppose that $(\overline{u}_0, v^0)$ satisfies the "asymptotic" Lax shock condition*

$$d\lambda_k(\overline{u}_0).v^0(-1) < 0 < d\lambda_k(\overline{u}_0).v^0(1); \tag{18}$$

then $(\overline{u}, v)$, a zero of $\mathcal{F}_{\epsilon,\tau}$, which lies in a sufficiently small neighborhood of $(\overline{u}_0, v^0)$, corresponds to a roll-wave solution of (4).

As a conclusion, we clearly see that the task is two-fold: first we have to prove the existence of a particular zero $(\overline{u}_0, v^0)$ of $\mathcal{F}_{0,\tau_0}$, which satisfies the "asymptotic" Lax shock condition (18). Then we show that $D\mathcal{F}_{0,\tau_0}$ is invertible at point $(\overline{u}_0, v^0)$. Shrinking the neighborhood of $(\overline{u}_0, v^0)$ if necessary, the Lax shock conditions will be automatically satisfied, and this will establish the existence of roll-waves.

3. Roll-wave solutions in the limit $\epsilon \to 0$. In this section we are going to compute a zero $(\overline{u}_0, v^0)$ of $\mathcal{F}_{0,\tau_0}$ which satisfies the "asymptotic" Lax shock condition

$$d\lambda_k(\overline{u}).v(-1) < 0 < d\lambda_k(\overline{u}).v(1). \tag{19}$$

It is easy to see that $(\overline{u}, v)$ as a zero of $\mathcal{F}_{0,\tau_0}$ satisfies

$$\begin{array}{l} g(\overline{u}) = 0, \\ \Pi_k(\overline{u}) D^2 f(\overline{u}).v(x).v'(x) - \Pi_k(\overline{u})\, dg(\overline{u})\big(v(x) - \langle v\rangle\big) = 0, \\ (1 - \Pi_k(\overline{u}))(Df(\overline{u}) - \lambda_k(\overline{u}))v'(x) = 0 \qquad \forall x \in (-1,1). \end{array} \tag{20}$$

Let us choose $\overline{u} = \overline{u}_0$: then $g(\overline{u}) = g(\overline{u}_0) = 0$. The last equation of (20) implies that v is given by

$$\begin{array}{l} v(x) = \Pi_k(\overline{u}_0)v(x) + V_k^0 \quad \forall x \in (-1,1), \\ v(x) = \lambda(x) r_k(u0) + V_k^0 \quad \forall x \in (-1,1), \end{array} \tag{21}$$

where $V_k^0 \in (1 - \Pi_k(\overline{u}_0))\mathbb{R}^n$ is a constant vector. Since $v \in \mathbb{X}$, it is clear that necessarily $V_k^0 = 0$. In order to compute λ, insert (21) into the second equation of (20): this yields

$$\lambda'(x) = \frac{\Pi_k(\overline{u}_0) dg(\overline{u}_0).r_k(\overline{u}_0)\big(\lambda(x) - \langle\lambda\rangle\big)}{\Pi_k(\overline{u}_0) D^2 f(\overline{u}_0).r_k(\overline{u}_0).r_k(\overline{u}_0)\lambda(x)} \qquad \forall x \in (-1,1). \tag{22}$$

The equation has a smooth solution if we choose $\langle\lambda\rangle = 0$. In that case, we find that

$$\lambda(x) = \alpha x, \quad \text{with} \quad \alpha = \tau_0 \frac{\Pi_k(\overline{u}_0) dg(\overline{u}_0).r_k(\overline{u}_0)}{\Pi_k(\overline{u}_0) D^2 f(\overline{u}_0).rk(\overline{u}_0).r_k(\overline{u}_0)}. \tag{23}$$

Hence we have proved the following proposition.

PROPOSITION 2. *Assume that*

$$\Pi_k(\overline{u}_0) D^2 f(\overline{u}_0).r_k(\overline{u}_0).r_k(\overline{u}_0) \neq 0, \qquad \Pi_k(\overline{u}_0)\, dg(\overline{u}_0).r_k(\overline{u}_0) \neq 0.$$

Then there exists a zero $(\overline{u}_0, v^0)$ of $\mathcal{F}_{0,\tau_0}$ with v^0 defined by

$$v^0(x) = \alpha\, x\, r_k(\overline{u}_0), \quad \textit{with} \quad \alpha = \tau_0 \frac{\Pi_k(\overline{u}_0) dg(\overline{u}_0).r_k(\overline{u}_0)}{\Pi_k(\overline{u}_0) D^2 f(\overline{u}_0).r_k(\overline{u}_0).r_k(\overline{u}_0)}. \tag{24}$$

Let us check the "asymptotic" Lax shock condition: this is done in the following lemma.

LEMMA 2. *Under the assumption $d\lambda_k(\overline{u}_0).r_k(\overline{u}_0) > 0$, the particular solution $(\overline{u}_0, \alpha\, x\, r_k(\overline{u}_0))$ satisfies the "asymptotic" Lax shock condition.*

Thus we have proved the existence of a zero $(\overline{u}_0, v^0)$ of $\mathcal{F}_{0,\tau_0}$, which satisfies the asymptotic Lax shock condition. Let us prove that the particular solution $(\overline{u}_0, v^0)$ persists for $\epsilon \neq 0$ and $\tau \approx \tau_0$ as an "admissible" zero of $\mathcal{F}_{\epsilon,\tau}$, (in the sense that they satisfy the Lax shock condition and are zeros of $\mathcal{F}_{\epsilon,\tau}$). This is done in the next section.

4. Persistence of the "asymptotic" roll-waves. In this section, we prove that $D\mathcal{F}_{0,\tau_0}$ is invertible at point $(\overline{u}_0, v^0)$. The expression for $D\mathcal{F}_{0,\tau_0}$ is given in the following lemma.

LEMMA 3. *The differential of $\mathcal{F}_{0,\tau_0}$ at the point $(\overline{u}_0, \alpha\, x\, r_k(\overline{u}_0))$ is given by*

$$\begin{aligned} D\mathcal{F}_{0,\tau_0} : \mathbb{X}\times\mathbb{R}^n &\to \mathbb{Y}\times\mathbb{R}^n \\ (v,u) &\mapsto \begin{array}{l} \Pi_k(\overline{u}_0)L_1(x).v + (1-\Pi_k(\overline{u}_0))L_2.v + M(x).\overline{u} \\ dg(\overline{u}_0).\overline{u}, \end{array} \end{aligned}$$

with

$$\begin{aligned} L_1(x).v &= \alpha x D^2 f(\overline{u}_0).r_k(\overline{u}_0).v' + D^2 f(\overline{u}_0).v.r_k(\overline{u}_0) - dg(\overline{u}_0).(v - \langle v\rangle), \\ L_2.v &= \big(Df(\overline{u}_0) - \lambda_k(\overline{u}_0)\big).v', \\ M(x).u &= \alpha^2 x\big(dA(\overline{u}_0).\overline{u}\big).r_k(\overline{u}_0).r_k(\overline{u}_0) \\ &\quad - \alpha x dB(\overline{u}_0).\overline{u}.r_k(\overline{u}_0) + \alpha dC(\overline{u}_0).\overline{u}.r_k(\overline{u}_0) \end{aligned} \tag{25}$$

and

$$\begin{aligned} &A(u) = \Pi_k(u)D^2 f(u), \\ &B(u) = \Pi_k(u)\, dg(u), \qquad\qquad \forall u\in\mathcal{V}(\overline{u}_0). \\ &C(u) = (1-\Pi_k(u))\big(Df(u) - \lambda_k(u)\big), \end{aligned} \tag{26}$$

Now we prove that $D\mathcal{F}_{0,\tau_0}$ is invertible with bounded inverse at the point $(\overline{u}_0, \alpha x r_k(\overline{u}_0))$: choose $(w,f)\in\mathbb{R}^n\times\mathbb{Y}$ and make the following assumption.

HYPOTHESIS 1. *The operator $dg(\overline{u}_0):\mathbb{R}^n\to\mathbb{R}^n$ is invertible.*

As a consequence, there exists a unique $\overline{u}$ so that

$$dg(\overline{u}_0).\overline{u} = w.$$

Now that $\overline{u}$ is determined, we prove that there exists a unique v so that

$$D\mathcal{F}_{0,\tau_0}(\overline{u},v)_1(x) = f(x) \qquad \forall x\in(-1,1). \tag{27}$$

Projecting the system (27) onto $(1-\Pi_k(\overline{u}_0))\mathbb{R}^n$ yields

$$(1-\Pi_k(\overline{u}_0))\Big(\big(Df(\overline{u}_0) - \lambda_k(\overline{u}_0)\big)v' + M(x).\overline{u}\Big) = (1-\Pi_k(\overline{u}_0))f(x). \tag{28}$$

Integrating (28) yields

$$v(x) = \lambda(x) r_k(\overline{u}_0) + V_k^0 + \mathcal{L}(\overline{u},f)(x), \tag{29}$$

with $V_k^0\in(1-\Pi_k(\overline{u}_0))\mathbb{R}^n$ a constant vector, and $\mathcal{L}(u,f)(x)\in(1-\Pi_k(\overline{u}_0))\mathbb{R}^n$ satisfies

$$\big(Df(\overline{u}_0) - \lambda_k(\overline{u}_0)\big)\mathcal{L}(\overline{u},f)(x) = (1-\Pi_k(\overline{u}_0))\int_0^x M(t).u + f(t)dt.$$

Since $M(x).u$ is polynomial in x and f lies in $\mathbb{Y}$, it is easily seen that

$$\mathcal{L}(\overline{u},f)\in\mathbb{X}.$$

Using the fact that $v\in\mathbb{X}$, V_k^0 is uniquely defined and is given by

$$V_k^0 = \langle \mathcal{L}(\overline{u},f)\rangle. \tag{30}$$

Now let us determine λ. Inserting (29) into the system (27) and projecting onto $\Pi_k(\overline{u}_0)\mathbb{R}^n$ yields the equation on λ:

$$x\lambda'(x) + \langle\lambda\rangle = \mathcal{A}(\overline{u}, f)(x) \qquad \forall x \in (-1, 1) \tag{31}$$

with

$$\begin{aligned}\beta\,\mathcal{A}(\overline{u}, f)(x) = {} & \Pi_k(\overline{u}_0)f(x) - \Pi_k(\overline{u}_0)M(x).\overline{u} \\ & + \Pi_k(\overline{u}_0)\Big(dg(\overline{u}_0)\big(\mathcal{L}(\overline{u}, f)(x) - \langle\mathcal{L}(\overline{u}, f)\rangle\big)\Big) \\ & - \alpha\Pi_k(\overline{u}_0)D^2f(\overline{u}_0)\big(\mathcal{L}(\overline{u}, f)(x) - \langle\mathcal{L}(\overline{u}, f)\rangle\big)r_k(\overline{u}_0) \\ & - \alpha x\Pi_k(\overline{u}_0)D^2f(\overline{u}_0).r_k(\overline{u}_0).\mathcal{L}(\overline{u}, f)'(x),\end{aligned} \tag{32}$$

and $\beta = \Pi_k(\overline{u}_0)dg(\overline{u}_0).r_k(\overline{u}_0)$. Clearly $\mathcal{A}(\overline{u}, f)$ lies in $\mathbb{Y}$. We prove that (31) has a *unique* solution in $\mathbb{X}$. Taking $x = 0$ in (31) yields

$$\langle\lambda\rangle = \mathcal{A}(\overline{u}, f)(0).$$

Then we find that

$$\lambda'(x) = \frac{\mathcal{A}(\overline{u}, f)(x) - \mathcal{A}(\overline{u}, f)(0)}{x} \qquad \forall x \in (-, 1). \tag{33}$$

The right-hand side of (33) lies in $\mathbb{Y}$: thus there exists a *unique* $\lambda \in \mathbb{X}$ which satisfies (33) and

$$\langle\lambda\rangle = \mathcal{A}(\overline{u}, f)(0).$$

Thus we have proved that $D\mathcal{F}_{0,\tau_0} : \mathbb{X} \times \mathbb{R}^n \to \mathbb{Y} \times \mathbb{R}^n$ is invertible. It is an easy computation to prove that this inverse is bounded from $\mathbb{R}^n \times \mathbb{Y}$ to $\mathbb{R}^n \times \mathbb{X}$, and we have proved the following proposition.

PROPOSITION 3. *Under the Hypothesis* 1*, the operator $D\mathcal{F}_{0,\tau_0}$ is invertible at the point $(\overline{u}_0, \alpha x r_k(\overline{u}_0))$. This particular solution persists as a zero of $\mathcal{F}_{\epsilon,\tau}$ to $\epsilon \approx 0$ and $\tau \approx \tau_0 > 0$.*

Using Lemma 2, we have proved that there exist roll-wave solutions of (4) in the case of "artificial" source terms for $0 < \epsilon \ll 1$, and they belong to a curve of roll-waves parametrized by the rescaled wavelength $\tau > 0$. Indeed, in most of the applications in physics, "real" source terms have the form

$$g(u) = \begin{pmatrix} 0 \\ h(u) \end{pmatrix}, \tag{34}$$

with $h : \mathbb{R}^n \to \mathbb{R}^{n-j}$ for some $j \geq 1$. We employ here a terminology similar to the one employed in the case of viscous approximation of conservation laws (we then deal with "artificial" and "real" viscosity). In the case of "real" source terms, the hypothesis (1) is not satisfied. Nevertheless, it is still possible to prove the existence of roll-waves. More precisely, we show the next theorem.

THEOREM 1. *Assume that $dh(\overline{u}_0) : \mathbb{R}^n \to \mathbb{R}^{n-j}$ is surjective. Then under the hypothesis*

$$\Pi_k(\overline{u}_0)D^2f(\overline{u}_0).r_k(\overline{u}_0).r_k(\overline{u}_0) \neq 0, \qquad \Pi_k(\overline{u}_0)dg(\overline{u}_0).r_k(\overline{u}_0) \neq 0,$$

and $\alpha d\lambda_k(\overline{u}_0).r_k(\overline{u}_0) > 0$, there exists a family of roll-wave solutions of (4) *which belongs to a j-dimensional manifold.*

Proof. In the case $g(u) =^t (0, h(u))$, we consider the mapping $\mathcal{G}_{\epsilon,\tau}$ defined by

$$\begin{aligned}\mathcal{G}_{\epsilon,\tau} : \mathbb{X} \times \mathbb{R}^n &\to \mathbb{Y} \times \mathbb{R}^{n-j} \\ (v, \overline{u}) &\mapsto \begin{array}{l} \mathcal{F}_{\epsilon,\tau}(u,v)_1, \\ \displaystyle\int_{-1}^{1} h(\overline{u} + \epsilon v(x))dx, \end{array}\end{aligned}$$

Then a zero $(\overline{u}, v)$ of $\mathcal{G}_{\epsilon,\tau}$ clearly satisfies the differential system (11) and the Rankine–Hugoniot conditions (12). Under the assumptions

$$\Pi_k(\overline{u}_0) D^2 f(\overline{u}_0).r_k(\overline{u}_0).r_k(\overline{u}_0) \neq 0, \qquad \Pi_k(\overline{u}_0) dg(\overline{u}_0).r_k(\overline{u}_0) \neq 0,$$

and $\alpha d\lambda_k(\overline{u}_0).r_k(\overline{u}_0) > 0$, we construct a zero $(\overline{u}_0, v^0 = \alpha\, x\, r_k(\overline{u}_0))$ of $\mathcal{G}_{0,\tau_0}$ (for any $\tau_0 > 0$ fixed). Then it is easily proved that $D\mathcal{G}_{0,\tau_0}(\overline{u}_0, v^0)$ is a submersion, and its kernel is j-dimensional with

$$Ker\, D\mathcal{G}_{0,\tau_0}(\overline{u}_0, v^0) = Ker\, dh(\overline{u}_0) \times \{0\}. \tag{35}$$

Consequently, the zeros of $\mathcal{G}_{0,\tau_0}$ form a j-dimensional manifold. Moreover, since $\mathcal{G}_{\epsilon,\tau}$ is a C^1 perturbation of $\mathcal{G}_{0,\tau_0}$, the zero set $\mathcal{G}_{\epsilon,\tau}^{-1}(0)$ is locally a j-dimensional manifold for $0 < \epsilon \ll 1$ and $\tau \approx \tau_0$. This completes the proof of the theorem. ∎

The question is now to find parameters that describes this j-dimensional family of solutions (the rescaled spatial period τ being fixed): indeed, since $g(u) =^t (0, h(u))$, we find that

$$\big(f_i(\overline{u} + \epsilon v(x)) - \lambda_k(\overline{u}) v_i\big)' = 0 \qquad \forall i = 1, \dots, j. \tag{36}$$

Then there exist $q_1, \dots, q_j$ so that

$$f_i(\overline{u} + \epsilon v(x)) - \lambda_k(\overline{u}) v_i(x) = q_i \qquad \forall i = 1, \dots, j. \tag{37}$$

The question arises whether we can use the conserved quantities $(q_i)_{i=1,\dots,j}$ and the adimensioned period τ to parametrize the solution. For that purpose, let us consider the mapping $\mathcal{H}_{\epsilon,\tau,q}$, where $q =^t (q_1, \dots, q_j)$ is defined by

$$\begin{aligned}\mathcal{H}_{\epsilon,\tau,q} : \mathbb{X} \times \mathbb{R}^n &\to \mathbb{Y} \times \mathbb{R}^n \\ (v, \overline{u}) &\mapsto \begin{array}{l} \mathcal{F}_{\epsilon,\tau}(\overline{u}, v)_1 \\ \pi_j\big(f(u + \epsilon v(1)) - \epsilon \lambda_k(\overline{u}) v(1)\big) - q, \\ \displaystyle\int_{-1}^{1} h(\overline{u} + \epsilon v(x))dx, \end{array}\end{aligned}$$

where $\pi_j : \mathbb{R}^n \to \mathbb{R}^j$ denotes the projection on the j first coordinates. Let us make the following assumption.

HYPOTHESIS 2. *The matrix $A_0 \in M_n(\mathbb{R})$ defined by*

$$A_0 = \begin{pmatrix} \pi_j\, df(\overline{u}_0) \\ dh(\overline{u}_0) \end{pmatrix} \tag{38}$$

is invertible.

Then it is easy to prove that at point $(\overline{u}_0, v^0)$ and for $q^0 = \pi_j(f(\overline{u}_0))$, the operator $D\mathcal{H}_{0,\tau_0,q_0}$ is invertible with a bounded inverse. Then applying the implicit function theorem, we see that for any fixed period $\tau \approx \tau_0$ and for $\epsilon \approx 0$, the j-dimensional manifold of roll-wave solutions is parametrized by the j conserved quantities $q \approx q^0$.

5. Conclusion. In this paper we have proved the existence of small amplitude roll-wave solutions for general hyperbolic systems with a "real" or "artificial" source term. These are periodic traveling waves, piecewise regular with discontinuities periodically distributed, which satisfy the Rankine–Hugoniot and a Lax shock condition (the shocks are indeed k-shocks). We have proved that when the zeros of the source term form a $(n-j)$-manifold, the roll-wave solutions, when they exist, form a $(j+1)$-manifold and are parametrized by their spatial period and j conserved quantities of the motion. This result generalizes the existence result obtained by Dressler in [4] for the Saint Venant equations, which reads in its adimensioned form as

$$\begin{aligned} &h_t + (hu)_x = 0, \\ &(hu)_t + \left(\frac{h^2}{2F} + hu^2\right)_x = h - u^2. \end{aligned} \tag{39}$$

In that case, the relative discharge rate $q = h(c-u)$ is a conserved quantity: Dressler proved in that case the existence of roll-waves parametrized by the spatial period $0 < L < \infty$ and the relative discharge rate $q > 0$. The question of the existence of large amplitude roll-waves for a general hyperbolic system with source term is still open.

Similarly to the study of Dressler roll-waves, several questions arise for inviscid roll-waves in hyperbolic systems with source term. On the one hand, considering artificial or real viscous perturbations of the system (4), one could address the question of the existence of continuous roll-waves. For a fixed size of the viscosity, we shall expect that, under suitable assumptions, one can show the existence of small amplitude roll-waves through a Hopf bifurcation argument. A more interesting question would be the "persistence" of inviscid roll-waves under viscous perturbation already addressed in [8] for Dressler roll-waves: given any inviscid roll-wave solution of (4), is there a viscous roll-wave solution if the viscous system in the vanishing viscosity limit converges to the inviscid roll-wave? This problem seems to be out of reach for general systems but should be possible to treat in low-dimensional systems $n = 2, 3$.

On the other hand, one could ask whether this type of solution is stable. In the viscous case, the question of the linear stability of roll-waves in Saint Venant systems has been treated in [9] using the approach initiated by Oh and Zumbrun [13] for periodic traveling waves in viscous conservation laws: it should be possible to prove for "general" viscous roll-waves that spectral stability implies the linear stability of viscous roll-waves. In the inviscid case, the question of the spectral stability is quite a hard problem but is treated in [10] for the Dressler roll-waves. The question of the nonlinear stability of inviscid roll-waves is also an open problem: this task will be completed for Dressler roll-waves in a forthcoming paper [11] using the framework introduced by Majda [14] for multidimensional compact shocks and generalized by Métivier and co workers [16]. In this case we prove the "persistence" result: more precisely, given any initial data which is close to a roll-wave and satisfies suitable compatibility conditions, one can show that the solution of the Cauchy problem exists on a sufficiently small interval and has a structure analogous to the roll-wave. This question shall be also addressed in a multidimensional framework.

REFERENCES

[1] A. Boudlal and V. Yu Liapidevskii, *Stability of roll waves in open channel flows*, C. R. Math. Acad. Sci. Paris, 303 (2002), pp. 291–295.

[2] C. Cercignani, *The Boltzmann Equation and Its Application*, Springer-Verlag, New York, 1988.

[3] G. Q. Chen, C. D. Levermore, and T.-P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.

[4] R. Dressler, *Mathematical solution of the problem of roll waves in inclined open channels*, Comm. Pure Appl. Math., 2 (1949), pp. 149–194.

[5] T. Kato, *Perturbation Theory for Linear Operators*, Springer, New York, 1995.

[6] S. Jin and M. A. Katsoulakis, *Hyperbolic systems with supercharacteristic relaxations and roll waves*, SIAM J. Appl. Math., 61 (2000), pp. 273–292.

[7] J. Keizer, *Statistical Thermodynamics of Nonequilibrium processes*, Springer-Verlag, New York, 1987.

[8] P. Noble, *Invariant Manifolds Methods for the Saint Venant Equations and the Discrete Hamiltonian Systems*, Laboratory Mathematics for Industry and Physics, University of Toulouse, Toulouse, France, 2003.

[9] P. Noble, *Linear stability of viscous roll-waves.* 2006, Comm. Partial Differential Equations, to appear.

[10] P. Noble, *On the spectral stability of roll-waves*, Indiana. Univ. Math. J., 55 (2006), pp. 795–848.

[11] P. Noble, *Persistence of Roll-Waves for the Saint Venant Equations*, in preparation.

[12] P. Noble, *Existence of pulsating roll-waves for the Saint Venant equations*, Arch. Ration. Mech. Anal, (2006), to appear.

[13] M. Oh and K. Zumbrun, *Stability of periodic solutions of conservation laws with viscosity: Pointwise bounds on the Green function*, Arch. Ration. Mech. Anal., 166 (2003), pp. 167–196.

[14] A. Majda, *The Stability of Multidimensional Shock Fronts*, Mem. Amer Math Soc., 275, AMS, Providence, RI, 1983.

[15] C. Mascia and K. Zumbrun, *Stability of large-amplitude shock profiles of general relaxation systems*, SIAM J. Math. Anal., 37 (2005), pp. 889–913.

[16] G. Métivier, *Stability of multidimensional shocks; Advances in the theory of shock waves*, in Progr. Nonlinear Differential Equations Appl. 47, Birkhäuser Boston, Cambridge, MA, 2001, pp. 25–103.

[17] R. L. Rabie, G. R. Fowles, and W. Fickett, *The polymorphic detonation*, Phys. Fluids A, 22 (1979), pp. 422–435.

[18] W.-A Yong and K. Zumbrun, *Existence of relaxation shock profiles for hyperbolic conservation laws*, SIAM J. Appl. Math., 60 (2000), pp. 1565–1575.

[19] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley, New York, 1974.

SIAM J. APPL. MATH.
Vol. 67, No. 5, pp. 1213–1232

MULTIPHASE IMAGE SEGMENTATION VIA MODICA–MORTOLA PHASE TRANSITION*

YOON MO JUNG†, SUNG HA KANG‡, AND JIANHONG SHEN†

Abstract. We propose a novel multiphase segmentation model built upon the celebrated phase transition model of Modica and Mortola in material sciences and a properly synchronized fitting term that complements it. The proposed sine-sinc model outputs a single multiphase distribution from which each individual segment or phase can be easily extracted. Theoretical analysis is developed for the Γ-convergence behavior of the proposed model and the existence of its minimizers. Since the model is not quadratic nor convex, for computation we adopted the convex-concave procedure (CCCP) that has been developed in the literatures of both computational nonlinear PDEs and neural computation. Numerical details and experiments on both synthetic and natural images are presented.

Key words. multiphase, segmentation, variational, partial differential equation, Modica–Mortola model, phase transition, Γ-convergence, convex splitting, convex-concave procedure

AMS subject classifications. Primary, 94A08; Secondary, 68U10, 49N45

DOI. 10.1137/060662708

1. Introduction.

The literature on segmentation has been the most wealthy and inspiring. From Geman and Geman's mixture random-field models [23] to Mumford and Shah's piecewise smooth variational image models [36], segmentation has been extensively studied by several major stochastic and deterministic machineries of modeling, analysis, and computation. New segmentation models incorporating more complexities or flexibilities have been further proposed by a number of authors in recent years, e.g., the data-driven Monte Carlo Markov chain (DDMCMC) model of Tu and Zhu [52], the graph-cutting and spectral method of Shi and Malik [47], and the variational texture segmentation models by Sandberg, Chan, and Vese [42], and Shen [45] (based on the texture models of Meyer [32] and Osher, Solé, and Vese [40]), just to name a few.

In this paper, we focus on the variational-PDE approach that is closely connected to the Mumford–Shah type of model. Computationally, such models have been implemented in various approaches: the finite-difference or finite-element methods, e.g., by Bourdin and Chambolle [5], Chambolle [8, 9], and Morel and Solimini [35], as well as the influential level-set approach by Chan and Vese [15, 13] (based on the level-set technology of Osher and Fedkiw [38], Osher and Sethian [39], and Sethian [43]). In the level-set approach, in particular, several multiphase computational models have been recently designed by Chan and Vese [16], Chung and Vese [17], as well as Lie, Lysaker, and Tai [26] and Tai and Chan [49]. We emphasize that the current work is more or less related to those ideas explored in [17] and [26], but it is carried out in a completely different framework.

*Received by the editors June 11, 2006; accepted for publication (in revised form) February 16, 2007; published electronically June 15, 2007. This research was supported by the NSF Program in Applied Mathematics under grants DMS-0202565 and DMS-0312223, as well as by the long term visiting program of IMA at the University of Minnesota.

http://www.siam.org/journals/siap/67-5/66270.html

†School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (ymjung@math.umn.edu, jhshen@math.umn.edu).

‡Corresponding author. Department of Mathematics, University of Kentucky, Lexington, KY 40506 (skang@ms.uky.edu).

An alternative approach to modeling and computing segmentation is via the theory of Γ-convergence elliptic approximations, as first developed by Ambrosio and Tortorelli [2, 3] for the Mumford–Shah model. This method has been extensively studied and extended for segmentation, inpainting, and several other applications in image analysis and processing (see, e.g., [20, 21, 30, 29, 46]). We propose a new multiphase segmentation model in the framework of Γ-convergence and phase transition, and develop the relevant mathematical analysis and computational strategies. More specifically, we propose to adopt the celebrated phase transition model of Modica and Mortola [33] with a *sinusoidal* potential. The new model is a self-contained segmentation model, and is different from Ambrosio and Tortorelli's formulation [2], which approximates and computes the Mumford–Shah model.

We hereby emphasize that the similarities are inherent between image segmentation and the phase transition problem in material sciences and fluid mechanics. First, different phases in material sciences are characterized by densities and tensions (e.g., ice versus water), while in image and vision analysis, distinct "object" segments are similarly characterized by some *visual features* such as intensities, orientations, or more general Gabor features (e.g., in texture segmentation [42, 14]). Second, the difficulties in dealing with sharp interfaces emerging from both material phase transitions and image segmentation share the very same roots—the characteristic complexities in handling free boundaries and their geometry. Under such observations, it is beneficial for the imaging community to borrow the successful ideas in contemporary material sciences, e.g., the diffuse-interface model of Cahn and Hilliard [7], and its rigorous mathematical analysis in the framework of Γ-convergence approximation by Modica and Mortola [33] (as initially conjectured by De Giorgi).

Finally, as for all the major segmentation efforts in existence, the phase field based segmentation model proposed herein is also nonlinear and nonconvex, and its robust computation (for local minima) is nontrivial. In the current work, we employ the so-called convex-concave (splitting) procedure (CCCP) as in the literatures of both computational nonlinear PDEs [4, 22, 54] and neural computation [55], and develop the corresponding computational schemes for the proposed energy functional.

The paper has been organized as follows. The new model is developed in section 2. The relevant Mumford–Shah segmentation model and its related literature are briefly reviewed in subsection 2.1, and the proposed Modica–Mortola sine-sinc model is established in subsection 2.2. We analyze the major mathematical properties of the model in section 3, including the Γ-convergence behavior in subsection 3.2 and the existence and compactness theorems in subsection 3.3. Computational schemes are presented in section 4, where we develop the convex-splitting or the CCCP algorithm in subsection 4.2, and demonstrate the numerical performance on generic image examples in subsection 4.3. The conclusion is drawn in section 5.

2. Multiphase segmentation via Modica–Mortola phase transition. In this section, we first motivate and develop the new model based upon the phase transition model of Modica and Mortola in material sciences and fluid dynamics [33], and discuss its connections to the Mumford–Shah segmentation model and some related works. Mathematical analysis will be further developed in the next section.

Let Ω be a bounded Lipschitz domain, and $u_o : \Omega \longrightarrow \mathbb{R}_+ \cup \{0\}$ be a given image. Recall that the classical Mumford–Shah segmentation is to minimize

$$\mathcal{E}_{ms}[u, \Gamma | u_o] = \mathcal{H}^1(\Gamma) + \alpha \int_{\Omega \setminus \Gamma} |\nabla u|^2 dx + \lambda \int_{\Omega} (u - u_o)^2 dx, \tag{2.1}$$

where $\Gamma \in \Omega$ denotes the *edge set* of the ideal image u, and $\mathcal{H}^1$ represents the 1-dimensional Hausdorff measure. This functional is well defined on

$$\mathcal{A}_{ms} = \{(u, \Gamma) \ : \ u \in H^1(\Omega \setminus \Gamma), \mathcal{H}^1(\Gamma) < \infty, \Gamma \text{ is relatively closed in } \Omega\},$$

provided that the given image $u_o \in L^2(\Omega)$. In machine learning [18, 41], $\mathcal{A}_{ms}$ represents the *hypothesis* space (*model* space) of all piecewise smooth functions on Ω.

2.1. Piecewise constant segmentation model. In order to identify individual objects, conceptually one has to carry out a postprocessing step after the Mumford–Shah model outputs the edge set Γ. That is, one has to identify the individual connected components Ω_i's of $\Omega \setminus \Gamma$. If each patch Ω_i is to be called a *phase*, then segmentation automatically bears the nature of *multiple* phases, since a generic image often contains *multiple* objects projected from the 3-dimensional world.

On the other hand, there also exist attempts to *directly* represent and compute different phases. Normally phase separation or identification relies upon the clustering of certain visual features such as frequency, local orientation, and local density, etc. A simple but commonly adopted phase feature is the image intensity value of a pixel. In the Mumford–Shah setting (2.1), this leads to the piecewise constant reduced Mumford–Shah (RMS) model,

$$\mathcal{E}_{rms}[u, \Gamma | u_o] = \mathcal{H}^1(\Gamma) + \lambda \int_\Omega (u - u_o)^2 dx,$$

defined on the following admissible space:

$$\mathcal{A}_{rms} = \{(u, \Gamma) \ : \ Du\big|_{\Omega \setminus \Gamma} = 0, \ \mathcal{H}^1(\Gamma) < \infty, \ \Gamma \text{ is relatively closed in } \Omega\} \subseteq \mathcal{A}_{ms}.$$

Here Du denotes the vectorial Radon measure of the total variation (TV) of u.

The TV constraint requires any admissible u to be constant on any connected component of $\Omega \setminus \Gamma$. For practical convenience, we assume that there are finitely many such patches, say K number of different patches. Then the entire image domain is partitioned into

$$\Omega \setminus \Gamma = \bigcup_{k=0}^{K-1} \Omega_k.$$

On each patch Ω_k, one must have $u|_{\Omega_k} := C_k, \ k = 0, \ldots, K-1$, for a set of distinct intensity values $\mathbf{C} = (C_0, \ldots, C_{K-1})$. Then the RMS energy can be rewritten as

$$\mathcal{E}_{rms}[\mathbf{C}, \Gamma | u_o] = \frac{1}{2} \sum_{k=0}^{K-1} \mathcal{H}^1(\partial \Omega_k) + \lambda \sum_{k=0}^{K-1} \int_{\Omega_k} (C_k - u_o)^2 dx.$$

Notice that the factor $\frac{1}{2}$ is due to the double counting of any two adjacent patches. With successful level-set implementation, the RMS model is also frequently referred to as the *Chan–Vese model* for two phases, honoring its rediscovery from the viewpoint of robust active contours [15]. This particular multiphase model was first used in [26] and has shown that it can keep symmetry for triple junctions.

The main mechanism of the proposed model is to identify *multiple phases* by piecewise constant values, similar to those considered in [17, 26, 49]. However, there are two major differences:

(i) All the aforementioned prior works are in essence still built upon the framework of Mumford and Shah (or its reduced form as discussed above), while our proposed model is not strictly a Mumford–Shah-type model (though some equivalence will be established immediately below).

(ii) All the aforementioned prior works have employed the celebrated level-set technology of Osher and Sethian [39], while our new model adopts the phase field framework in material sciences and fluid mechanics. A level-set function offers remarkable efficiency and robustness for representing and computing free boundaries, yet (strictly speaking) does not participate in the modeling process, while a phase field function is indispensably part of the model itself.

To proceed, we label each phase component with an integer and define a *signature* function z by

$$z(x) = k \ \text{ if } \ x \in \Omega_k, \qquad k = 0, \dots, K-1.$$

In practice, phase extraction is of course the very opposite process from getting the signature function z. That is, one has to first obtain the signature function z before different phase patches Ω_k can be identified and extracted. For convenience, we shall also call z a *phase field*. We then propose the multiphase segmentation model *in this ideal scenario* by minimizing

$$E[\mathbf{C}, z|u_o] = \int_\Omega |Dz(x)| + \lambda \sum_{k=0}^{K-1} \int_\Omega (C_k - u_o)^2 \chi_{\{z=k\}} dx.$$

As in the RMS or the Chan–Vese model [13, 36], the intensity distribution variable $\mathbf{C}$ can be conditionally solved by the following: from any given phase field z,

$$C_k = \langle u_o \rangle_{\Omega_k} = \frac{1}{|\Omega_k|} \int_{\Omega_k} u_o(x)\, dx, \qquad i = 0, \dots, K-1. \tag{2.2}$$

Then the *ideal* model depends only on the phase field z:

$$E[z|u_o] = \int_\Omega |Dz(x)| + \lambda \sum_{k=0}^{K-1} \int_\Omega (C_k - u_o)^2 \chi_{\{z=k\}} dx. \tag{2.3}$$

The admissible class is simply $\mathcal{A} = \{z \in BV(\Omega) : z(x) \in \mathbb{Z} \text{ a.e.}\}$.

The main challenge of the proposed model arises from its mixture of the *continuous* TV Radon measure and the *discrete* constraint. Thus in the next subsection, this *ideal* model will be further polished in the framework of phase transitions and Γ-convergence.

Here we first emphasize that this ideal energy is somewhat equivalent to the piecewise constant Mumford–Shah functional. With the help of the signature function and the formulae for C_k's, $\mathcal{E}_{rms}$ depends only on Γ or z and can be rewritten as

$$\mathcal{E}_{rms}[\Gamma|u_o] = \frac{1}{2} \sum_{k=0}^{K-1} \mathcal{H}^1(\partial\{z = k\}) + \lambda \sum_{k=0}^{K-1} \int_\Omega (C_k - u_o)^2 \chi_{\{z=k\}} dx. \tag{2.4}$$

For any fixed number of phases K, the two functionals (2.3) and (2.4) are then equivalent, in the sense of

$$\mathcal{E}_{rms}[\Gamma|u_o] \le E[z|u_o] \le K\mathcal{E}_{rms}[\Gamma|u_o],$$

since

$$\int_\Omega |Dz| = \int_\Gamma \big|[z]\big| d\mathcal{H}^1 \quad \text{and} \quad 1 \le \big|[z]\big| < K.$$

In most applications (especially in medical imaging), for instance, $K \le 5$.

The difference between the two functionals (2.3) and (2.4) is also obvious, since the former weighs the jumps, while the latter does not. For example, consider a rectangular domain Ω and two disjoint disks Ω_1, Ω_2 in Ω with radius $1/2$. If we assign $z = 0$ on $\Omega_0 = \Omega \setminus (\Omega_1 \cup \Omega_2)$, $z = 1$ on Ω_1, and $z = 2$ on Ω_2, then $\sum_{k=0}^2 \mathcal{H}^1(\partial\Omega_k) = \frac{1}{2}\sum_{k=0}^2 \mathcal{H}^1(\partial\{z = k\}) = \frac{1}{2}(2\pi + \pi + \pi) = 2\pi$, but $\int_\Omega |Dz(x)| = 1\pi + 2\pi = 3\pi$. Thus, $\int_\Omega |Dz|$ is a *weighted length* of Γ. Recall that the TV Radon measure can be decomposed into

$$Dz = \nabla z + [z]\Big|_{S_z} \mathcal{H}^1 \llcorner S_z + \mathcal{C}_z,$$

corresponding to the Lebesgue continuous gradient, the jump set $S_z = \Gamma$ (or reduced boundary) with $[z] = z^+ - z^-$, and the singular Cantor measure. Now if the signature z is ideally piecewise constant, both the Lebesgue and Cantor components must vanish,

$$\int_\Omega |Dz| = \int_{S_z} \big|[z]\big| d\mathcal{H}^1,$$

which clearly shows the weighing nature of $\int_\Omega |Dz|$.

Weighing the object boundaries certainly makes the proposed *ideal* model depending upon the labels. But for a fixed number K of phases, the energies are more or less equivalent as just discussed. More importantly, it allows us to invoke the celebrated phase transition approach in material sciences and fluid dynamics, in order to successfully overcome the major challenge in reconciling the two very opposite characteristics of the segmentation problem: continuum vs. discreteness.

2.2. The sine-sinc model via Modica–Mortola phase transition. The objective functional (2.3)

$$E[z|u_o] = \int_\Omega |Dz(x)| + \lambda \sum_{k=0}^{K-1} \int_\Omega (C_k - u_o)^2 \chi_{\{z=k\}} dx$$

itself does not appear too complicated but becomes immensely baffling under the discrete constraint $z \in \mathbb{Z}$. *This is a very common scenario in integer or discrete programming.* For example, it is difficult to minimize this functional by any ordinary PDE approaches such as Euler–Lagrange equations or gradient-descent time marching.

We thus introduce its relaxed version via the celebrated model of Modica and Mortola [33] on phase transitions in material sciences and fluid mechanics. Recall that in the classical literature on phase transitions, the mixture of two immiscible and incompressible fluids are often modelled so that in equilibrium they separate into two phases with a minimal interface area. Cahn and Hilliard [7] first proposed to use a thin layer of continuous interface (i.e., *diffuse interface*) to model this separation. Later on Modica [34] proved that the Cahn–Hilliard model Γ-converges to the classical model. In another well-known paper [33], Modica and Mortola established that the diffuse-interface energy

$$F_\epsilon[z] = \int_\Omega \left[\epsilon |\nabla z|^2 + \frac{1}{\epsilon}\sin^2 \pi z\right] dx \quad \Gamma\text{-converges to} \quad \frac{4}{\pi}\int_\Omega |Dz(x)|$$

for phase fields that ultimately take only integer values.

Notice that in the Modica–Mortola model, the *discrete* constraint $z \in \mathbb{Z}$ has been *softly* enforced due to the sine potential regulated by the transition bandwidth ϵ in the denominator. In the present work, we adopt this Modica–Mortola diffuse-interface energy F_ϵ to approximate the ideal TV energy in $E[z|u_o]$. This model thus well integrates both the TV regularity and the integer constraint $z \in \mathbb{Z}$.

For the data-fitting term, instead of the ideal indicator $\chi_{\{z=k\}}$ or Kronecker's delta, we also propose to use a properly relaxed version to facilitate model analysis and computing. More specifically, we choose $\mathrm{sinc}^2(z-k)$ to match the sine potential in Modica and Mortola's phase transition energy,

$$G[z|u_o] = \sum_{k=0}^{K-1} \int_\Omega |C_k - u_o|^2 \mathrm{sinc}^2(z-k)\, dx.$$

Recall that the sinc function is defined as $\mathrm{sinc}(z) = \frac{\sin \pi z}{\pi z}$ for $z \in \mathbb{R}$. For a phase field z that takes almost only integer values, sinc leads to desirable approximations to the indicator functions of integer phases and is more appealing computationally. This is because (i) $\mathrm{sinc}(k) = \delta_k$ (Kronecker's delta) for $k \in \mathbb{Z}$, the so-called interpolating property in the celebrated theorem of Shannon interpolation [11, 19, 48]; and (ii) $\mathrm{sinc}(z)$ is an *entire* function for $z \in \mathbb{C}$, and when $z \in \mathbb{R}$,

$$\frac{d}{dz}\mathrm{sinc}(z) = \mathrm{O}\left(\frac{1}{|z|}\right) \quad \text{as } z \to \pm\infty.$$

Thus in particular, $\mathrm{sinc}\,(z) \in W^{1,\infty}(\mathbb{R})$ and is Lipschitz continuous. As a result, the sinc-approximation will facilitate both analysis and computation later on.

In combination, we have arrived at the *relaxed* version of the *ideal* multiphase segmentation model (2.3) with a given number K of phases:

$$\begin{aligned} E_\epsilon[z|u_o] &= F_\epsilon[z] + \lambda G[z|u_o] \\ &= \int_\Omega \left[\epsilon|\nabla z|^2 + \frac{1}{\epsilon}\sin^2 \pi z\right] dx + \lambda \sum_{k=0}^{K-1} \int_\Omega |C_k - u_o|^2 \mathrm{sinc}^2(z-k)\, dx. \end{aligned} \tag{2.5}$$

Here the values C_k's have been conditionally optimized by the following equation with any given phase field z (under the least-square principle):

$$C_k = C_k[z] = \begin{cases} \dfrac{\int_\Omega u_o \,\mathrm{sinc}^2(z-k)\, dx}{\int_\Omega \mathrm{sinc}^2(z-k)\, dx} & \text{if } \displaystyle\int_\Omega \mathrm{sinc}^2(z-k)\, dx > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

When $z(x)$ takes only integer values, this C_k indeed reproduces the value introduced in (2.2). If the denominator in (2.6) vanishes, since

$$\int_\Omega \mathrm{sinc}^2(z-k)\, dx = 0 \iff \mathrm{sinc}(z(x)-k) = 0 \text{ a.e } x \in \Omega \iff z(x) \in \mathbb{Z} \setminus \{k\},$$

we observe that the phase k is empty and redundant. Also in this case the particular value of $C_k[z]$ is unimportant, since the integral $\int_\Omega |u - C_k|^2 \mathrm{sinc}^2(z-k)\, dx$ vanishes.

This relaxed functional (2.5) is our proposed model, and there are essential differences between our model and the closely related work by Lie, Lysaker, and Tai [26] on the piecewise constant level-set method (PCLSM).

- We utilize $\sin(z(x)\pi) = 0$ as our constraints, while, in [26], $\Pi_{i=1}^{k}(z-i) = 0$ is used to get integer value. In our model, we do not need to predetermine the number of phase k, and $\sin(z)$ synchronizes well with $\mathrm{sinc}(z)$ function. The $\mathrm{sinc}(z)$ function is used in place of the usual Heaviside and delta functions.
- The PCLSM gives sharp edges by directly using $\int |\nabla u|$, while one may have to resolve singular layer issues related to renormalization of level sets. This proposed relaxed functional does not have any singular layer issues, and the boundary is identified as a transition band determined by the size of ϵ.

3. Γ-convergence of the model and existence of minimizers. In this section, we develop the necessary analysis of the proposed model. Following a brief review of the Γ-convergence theory, we first show that the relaxed model (2.5) converges to the original piecewise constant model (2.3) and then prove that optimal segmentation does exist for any fixed ϵ. Compactness of the minimizers for all ϵ's is also discussed in the end.

3.1. Brief review of Γ-convergence. Γ-convergence was first introduced by De Giorgi and Franzoni in [24] to facilitate analysis and approximation of PDEs and variational problems. Since then it has been widely applied to phase transition models in material sciences, the modeling of thin films or plates, homogenization of variational problems, and free discontinuity problems (see, e.g., [6, 31]). In image processing, the most influential application is Ambrosio–Tortorelli's Γ-convergence approximation to the Mumford–Shah functional [2]. The definition of Γ-convergence is as follows.

DEFINITION 3.1 (Γ-convergence). *Let X be a metric space and $\mathcal{F}_\epsilon : X \to \bar{\mathbb{R}}$ for $\epsilon > 0$. We say that $\mathcal{F}_\epsilon$ Γ-converges to $\mathcal{F}$ in X as $\epsilon \to 0$ and write $\mathcal{F}_\epsilon \xrightarrow{\Gamma} \mathcal{F}$ if the following two conditions hold for all $u \in X$:*

(i) *(liminf inequality) for every sequence (u_ϵ) converging to u,*

$$\mathcal{F}(u) \le \liminf_{\epsilon \to 0} \mathcal{F}_\epsilon(u_\epsilon);$$

(ii) *(limsup inequality) there exists a sequence (u_ϵ) converging to u such that*

$$\mathcal{F}(u) \ge \limsup_{\epsilon \to 0} \mathcal{F}_\epsilon(u_\epsilon).$$

The most important properties of Γ-convergence are summarized by the following theorem. We refer the reader to [1, 6, 31] for further discussion.

THEOREM 3.2. Γ*-convergence has the following properties:*

(i) *(spatial stability of the limit) the* Γ*-limit $\mathcal{F}$ is lower semicontinuous;*

(ii) *(stability under continuous perturbations) if $\mathcal{F}_\epsilon \xrightarrow{\Gamma} \mathcal{F}$ and $\mathcal{G}$ is continuous, then $\mathcal{F}_\epsilon + \mathcal{G} \xrightarrow{\Gamma} \mathcal{F} + \mathcal{G}$;*

(iii) *(stability of minimizing sequences) if $\mathcal{F}_\epsilon \xrightarrow{\Gamma} \mathcal{F}$ and v_ϵ minimizes $\mathcal{F}_\epsilon$, then every cluster point of (v_ϵ) minimizes $\mathcal{F}$.*

Let us briefly comment on these three properties. Property (i) reveals the necessary condition for designing Γ-convergence approximation to a target functional. Property (ii) paves the way to extending existing Γ-convergence schemes, which is particularly helpful in the current work. Property (iii) reveals the real essence of the entire machinery of Γ-convergence, by which the minimization of a touchy objective $\mathcal{F}$ is tamed or relaxed by a family of better behaved objectives $\mathcal{F}_\epsilon$.

Directly benefiting the current work is the following remarkable theorem in the original paper of Modica and Mortola [33], which historically has played significant roles in the theories of Γ-convergence, phase transitions, and the Cahn–Hilliard model.

THEOREM 3.3 (Modica and Mortola [33]). *Define* $\mathcal{S} = \{z \in BV(\mathbb{R}^n) : z(x) \in \mathbb{Z}$ *a.e.* $x\}$, *and*

$$F_\epsilon(z) := \begin{cases} \displaystyle\int_{\mathbb{R}^n} \left[\epsilon|\nabla z(x)|^2 + \frac{1}{\epsilon}\sin^2(\pi z(x))\right] dx & \text{for } z \in H^1(\mathbb{R}^n) \cap L^1(\mathbb{R}^n), \\ +\infty & \text{for } z \in L^1(\mathbb{R}^n) \setminus H^1(\mathbb{R}^n), \end{cases}$$

$$F(z) := \begin{cases} \displaystyle\frac{4}{\pi}\int_{\mathbb{R}^n} |Dz(x)| & \text{for } z \in S(\mathbb{R}^n), \\ +\infty & \text{for } z \in L^1(\mathbb{R}^n), \text{ but } z \notin S(\mathbb{R}^n). \end{cases}$$

Then the functional F_ϵ Γ*-converges to* F *as* $\epsilon \to 0$ *in* $L^1(\mathbb{R}^n)$.

We note that the domain $\mathbb{R}^n$ can be replaced by any regular open bounded domain Ω, which is the case in the current application. Theorem 3.3 was originally conjectured by De Giorgi and then proven by Modica and Mortola [33] in 1977, shortly after the notion of Γ-convergence was introduced in [24]. The connection with the Cahn–Hilliard model was established in Modica [34].

3.2. Γ-convergence of the sine-sinc model. In image processing, the image range is often bounded by $u_o \in [0, 1]$ in the analog setting and $u_o \in [0, 255]$ in the digital setting with 8 bits. Therefore, we assume $u_o \in L^\infty(\Omega)$ for technical clarity.

In the Modica–Mortola sine-sinc model proposed in (2.5),

$$\begin{aligned} E_\epsilon[z|u_o] &= F_\epsilon[z] + \lambda G[z|u_o] \\ &= \int_\Omega \left[\epsilon|\nabla z|^2 + \frac{1}{\epsilon}\sin^2 \pi z\right] dx + \lambda \sum_{k=0}^{K-1} \int_\Omega |C_k - u_o|^2 \operatorname{sinc}^2(z-k)\, dx, \end{aligned}$$

the first term F_ϵ has already been proven to Γ-converge to F in [33]. In this subsection, we show that the fitting term G is continuous in $L^1(\Omega)$.

In what follows, we shall use the notation $\omega(x|z)$ to denote a spatial function $\omega(\cdot|z)$ that depends on the given phase field $z = z(x)$. Such dependence could be *local* in the form of $g(x, z(x))$, or *global* in the form of $g(x, J[z])$, where $J[z]$ is a functional on z. The following general theorem gives a unified foundation for the proof of Γ-convergence of the proposed model.

THEOREM 3.4. *Suppose that*

(i) $\varphi : \mathbb{R} \to \mathbb{R}$ *is Lipschitz continuous with* $|\varphi(x) - \varphi(y)| \le L|x - y|$,

(ii) $\|w(\cdot|z_n) - w(\cdot|z)\|_{L^\infty(\Omega)} \to 0$ *as* $\|z_n - z\|_{L^1(\Omega)} \to 0$,

(iii) $\|w(\cdot|z)\|_{L^\infty(\Omega)} \le M$ *for some positive* M *and for all* $z \in L^1(\Omega)$;

then $g[z] = \int_\Omega w(x|z)\varphi(z)\, dx$ *is continuous for* $z \in L^1(\Omega)$.

Proof. Let $\{z_n\}$ be a sequence converging to z: $z_n \to z$ in $L^1(\Omega)$. Then

$$\begin{aligned} |g[z_n] - g[z]| &\le \left|\int_\Omega w(x|z_n)\varphi(z_n) - w(x|z)\varphi(z_n) dx\right| + \left|\int_\Omega w(x|z)\varphi(z_n) - w(x|z)\varphi(z) dx\right| \\ &\le \|w(x|z_n) - w(x|z)\|_{L^\infty(\Omega)} \|\varphi(z_n)\|_{L^1(\Omega)} + M\int_\Omega |\varphi(z_n) - \varphi(z)| dx \\ &\le L\|z_n\|_{L^1(\Omega)} \|\omega(x|z_n) - w(x|z)\|_{L^\infty(\Omega)} + ML\|z_n - z\|_{L^1(\Omega)}. \end{aligned}$$

Both terms tend to zero as $\|z_n - z\|_{L^1(\Omega)} \to 0$ by assumption (ii). ∎

This theorem can be applied to establish that $C_k[z]$'s are continuous, and eventually that $G[z|u_o]$ is continuous as well.

COROLLARY 3.5. *Suppose $z_n \to z$ in L^1 and $\int_\Omega \operatorname{sinc}^2(z(x)-k)dx \neq 0$. Then $C_k(z)$ defined in (2.6) is continuous in z, i.e., $C_k[z_n] \to C_k[z]$.*

This is directly proven by Theorem 3.4 with $w(x) = u_o(x)$ (for the denominator) and $w(x) = 1$ (for the numerator), as well as $\varphi(z) = \sin^2(z(x)-k)$.

COROLLARY 3.6 (degenerate case). *If $\varphi(z(x)) = 0$ a.e., then $g[z] = 0$ and condition* (ii) *in Theorem* 3.4 *can be dropped.*

It simply comes from the definition of $g[z]$ and conditions (i) and (iii):

$$|g[z_n]| \le M \int_\Omega |\varphi(z_n)| = M \int_\Omega |\varphi(z_n) - \varphi(z)|dx \le ML \int_\Omega |z_n - z|dx \to 0.$$

PROPOSITION 3.7. *The fitting functional G is continuous in $L^1(\Omega)$:*

$$G[z|u_o] = \sum_{k=0}^{K-1} \int_\Omega |C_k - u_o|^2 \operatorname{sinc}^2(z-k)\,dx$$

with

$$C_k = C_k[z] = \begin{cases} \frac{\int_\Omega u_o \operatorname{sinc}^2(z-k)\,dx}{\int_\Omega \operatorname{sinc}^2(z-k)\,dx} & \text{if } \int_\Omega \operatorname{sinc}^2(z-k)\,dx > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Let $w(x|z) = (u_o(x) - C_k[z])^2$ and $\varphi(z) = \operatorname{sinc}^2(z-k)$ for $z \in L^1(\Omega)$. We now show that $G[z|u_o]$ indeed satisfies all three conditions of Theorem 3.4. Condition (i) is clear, since φ is Lipschitz continuous. For (ii), let $\{z_n\}$ be a converging sequence $z_n \to z$ in $L^1(\Omega)$. Then

$$\begin{aligned} |w(x|z_n) - w(x|z)| &= |(C_k[z_n] - u_o(x))^2 - (C_k[z] - u_o(x))^2| \\ &\le |C_k[z_n] + C_k[z] - 2u_o(x)||C_k[z_n] - C_k[z]| \\ &\le 4\|u_o\|_{L^\infty(\Omega)}|C_k[z_n] - C_k[z]|, \end{aligned}$$

and $|C_k[z_n] - C_k[z]| \to 0$ from the continuity of C_k in Corollary 3.5. Thus $|w(x|z_n) - w(x|z)| \to 0$. For (iii), notice that $|C_k[z]| \le \|u_o\|_{L^\infty(\Omega)}$ for all z. Then

$$w(x|z) \le (\|u\|_{L^\infty(\Omega)} + |C_k[z]|)^2 \le (2\|u_o\|_{L^\infty(\Omega)})^2 \quad \text{for all } x \in \Omega.$$

Therefore, by Theorem 3.4, $G[z|u_o]$ is continuous in $L^1(\Omega)$. ∎

Finally, the Γ-convergence of the Modica–Mortola sine-sinc model becomes evident from the combination of Proposition 3.7, Theorem 3.3, and Theorem 3.2.

THEOREM 3.8 (Γ-convergence). *E_ϵ Γ-converges to E w.r.t. $L^1(\Omega)$ topology.*

3.3. Existence of minimizers of E_ϵ. In this subsection we show that for each ϵ, a minimizer to E_ϵ does exist. We shall also briefly discuss the compactness of a sequence of minimizers for all ϵ's.

THEOREM 3.9 (existence of minimizers). *Suppose the given image $u_o \in L^2(\Omega)$ and denote the admissible space for K-phase segmentation by*

$$\mathcal{A}_K = \{z \in H^1(\Omega) : -1/2 < z < K - 1/2\}.$$

Then for each $\epsilon > 0$, there exists a minimizer of E_ϵ in $\mathcal{A}_K$.

Proof. We first show that the infimum is finite. Consider the uniform phase $z \equiv 0 \in \mathcal{A}_K$. Then the Modica–Mortola energy $F_\epsilon[z] = 0$, and $C_0 = \frac{1}{|\Omega|}\int_\Omega u_o(x)dx < \infty$, while $C_k = 0$ for $k = 1, \dots, K-1$. Therefore,

$$E_\epsilon[0|u_o] = \lambda \int_\Omega |C_0 - u_o(x)|^2 \, dx < \infty, \tag{3.1}$$

and since $E_\epsilon[\,\cdot\,|u_o] \geq 0$, the claim is proved.

Let $\{z_n\}$ be a minimizing sequence for E_ϵ in $\mathcal{A}_K$. Since $\sup_n \int_\Omega |\nabla z_n|^2 \, dx < \infty$ from the Modica–Mortola energy, as well as $-1/2 < z_n < K - 1/2$, the sequence $\{z_n\}$ must be precompact in $L^1(\Omega)$ by Rellich and Kondrachov's compactness theorem. There thus exists a subsequence of $\{z_n\}$, which is still denoted by $\{z_n\}$ after relabelling to simplify notations, such that

$$z_n \to z^* \text{ in } L^2(\Omega) \text{ for some } z^* \in L^2(\Omega).$$

As a result, one has the weak convergence for the gradient fields:

$$\nabla z_n \to \nabla z^* \text{ weakly in } L^2(\Omega).$$

Then, by the lower semicontinuity of the L^2-norm under the weak topology,

$$\int_\Omega |\nabla z^*|^2 \, dx \leq \liminf_{n\to\infty} \int_\Omega |\nabla z_n|^2 \, dx. \tag{3.2}$$

For the other two terms with sine and sinc functions, convergence follows from Lebesgue's dominated convergence theorem (LDCT) as shown below. First, possibly with another round of subsequence refinement and relabelling, one can further assume that $z_n \to z^*$ a.e. Then $\sin^2 \pi z_n \to \sin^2 \pi z^*$ a.e., and by LDCT,

$$\int_\Omega \sin^2 \pi z_n \, dx \to \int_\Omega \sin^2 \pi z^* \, dx. \tag{3.3}$$

Similarly, $\mathrm{sinc}^2\pi(z_n - k) \to \mathrm{sinc}^2\pi(z^* - k)$ a.e. for $k = 0, \dots, K-1$. By LDCT, one has $\int_\Omega \mathrm{sinc}^2\pi(z_n - k)\, dx \to \int_\Omega \mathrm{sinc}^2\pi(z^* - k)\, dx$, and

$$\int_\Omega u_o(x)\mathrm{sinc}^2\pi(z_n - k)\, dx \to \int_\Omega u_o(x)\mathrm{sinc}^2\pi(z^* - k)\, dx,$$

since $u_o \in L^2(\Omega) \subset L^1(\Omega)$. Consequently, if $\int_\Omega \mathrm{sinc}^2\pi(z^* - k)\, dx > 0$ for kth phase, then $(C_k[z_n])_n$ must be a bounded sequence. Since $u_o \in L^2(\Omega)$, by LDCT, as $n \to \infty$,

$$\int_\Omega |u_o - C_k[z_n]|^2 \mathrm{sinc}^2\pi(z_n - k)\, dx \to \int_\Omega |u_o - C_k[z^*]|^2 \mathrm{sinc}^2\pi(z^* - k)\, dx. \tag{3.4}$$

If, on the other hand, $\mathrm{sinc}^2\pi(z^* - k) = 0$ a.e., then

$$\int_\Omega |u_o - C_k[z^*]|^2 \mathrm{sinc}^2(z^* - k)\, dx = 0 \leq \liminf_{n\to\infty} \int_\Omega |u_o - C_k[z_n]|^2 \mathrm{sinc}^2(z^* - k)\, dx. \tag{3.5}$$

Finally, in combination of (3.2), (3.3), (3.4), and (3.5), we have

$$E_\epsilon[z^*|u_o] \leq \liminf_{n\to\infty} E_\epsilon[z_n|u_o] \leq \inf_{z\in\mathcal{A}_K} E_\epsilon[z|u_o],$$

and the limit z^* has to be a minimizer of E_ϵ. This completes the proof. ∎

Notice that in this theorem, we have even allowed the given image $u_o \in L^2(\Omega)$, instead of $u_o \in L^\infty(\Omega)$, which is the default assumption throughout the work.

Finally, we briefly comment on the compactness or stability of the sequence of minimizers from the Γ-convergence approximation.

THEOREM 3.10 (compactness of the sequence of minimizers). *Let z_ϵ minimize E_ϵ for each $\epsilon > 0$. Then there exist a subsequence $(z_{\epsilon'})$ of (z_ϵ) and some $z \in L^1(\Omega)$ such that $z_{\epsilon'} \longrightarrow z$ in $L^1(\Omega)$ as $\epsilon' \to 0$, and z minimizes E.*

Proof. By the Cauchy–Schwarz inequality,

$$\begin{aligned} E_\epsilon[z|u_o] \geq F_\epsilon[z] &= \int_\Omega \left[\epsilon|\nabla z|^2 + \frac{1}{\epsilon}\sin^2 \pi z\right] dx \\ &\geq 2\int_\Omega |\nabla z||\sin \pi z| dx = \int_\Omega |\nabla(H(z))| dx. \end{aligned} \tag{3.6}$$

Here $H : (-\frac{1}{2}, K - \frac{1}{2}) \to \mathbb{R}$ satisfies $H'(r) = 2|\sin \pi r|$ and $H(0) = 0$. By (3.1), there exists $M > 0$ such that $E_\epsilon[z_\epsilon|u_o] \leq M$ for all ϵ. By (3.6), the sequence of functions $(h_\epsilon(x) = H(z_\epsilon(x)))_\epsilon$ must be bounded in $BV(\Omega)$, and thus precompact in $L^1(\Omega)$. By subsequence refinement, one can assume that there exists a subsequence $h_{\epsilon'} \to h$ a.e. for some $h \in L^1(\Omega)$. Since $H(r)$ is continuous and strictly monotone, it admits a continuous inverse. Thus one has $z_{\epsilon'}(x) \longrightarrow z(x) = H^{-1}(h(x))$ a.e. Since $z_\epsilon \in (-\frac{1}{2}, K - \frac{1}{2})$ is uniformly bounded for ϵ, one must have $z_{\epsilon'} \to z$ in $L^1(\Omega)$ by LDCT. Then the rest of the theorem follows from Theorem 3.2. ∎

4. Computation and experiments. In this section, we develop the computational schemes for the proposed model. The major difficulty arises from the fact that the Modica–Mortola sine-sinc functional is nonconvex. In this paper, we apply the method of convex splitting or the concave-convex procedure (CCCP) for robustly computing the local minima of the model. After a brief review on the CCCP method, we detail our computational strategies, and test the schemes on some generic examples involving both synthetic and natural images.

4.1. Review of the CCCP. There are growing interests in how to solve nonconvex functions efficiently. In [22] in the setting of gradient flows, Eyre proposed to split nonconvex functions into two functions, contractive and expansive. It included computational examples of the Cahn–Hilliard equation with different time steps. More analysis on the numerical algorithms for the Cahn–Hilliard or Allen–Cahn equations were studied by Vollmayr-Lee and Rutenberg in [54], where unconditionally stable time step was explored. The idea of convex splitting is also applied to the Cahn–Hilliard inpainting by Bertozzi, Esedoglu, and Gilette [4].

Independent of the computational PDE literature, on the other hand, the similar idea of convex splitting was also explored by Yuille and Rangarajan [55] in a more general setting of neural computation, where the method has been termed the CCCP. The method has found many important applications in computer vision and neural computation.

THEOREM 4.1 (Yuille and Rangarajan [55]). *Let $\mathcal{E}(\vec{x})$ with $\vec{x} \in \mathbb{R}^n$ be an energy function with a bounded Hessian. Then it can be decomposed into the sum of a convex function and a concave function.*

THEOREM 4.2 (Yuille and Rangarajan [55]). *Consider an energy function which is bounded below and is an addition of convex and concave functions:*

$$\mathcal{E}(\vec{x}) = \mathcal{E}_{convex}(\vec{x}) + \mathcal{E}_{concave}(\vec{x}).$$

Then the discrete iterative CCCP algorithm given by

$$\nabla \mathcal{E}_{convex}(\vec{x}^{n+1}) = -\nabla \mathcal{E}_{concave}(\vec{x}^{n}), \qquad n = 0, 1, \ldots, \tag{4.1}$$

is guaranteed to monotonically decrease the energy $\mathcal{E}(\vec{x})$ as a function of time and to converge to a local minimum or a saddle point of $\mathcal{E}(\vec{x})$.

We briefly comment on these results for our application. First, notice that (4.1) is solvable only when

$$\text{Range}(-\nabla \mathcal{E}_{concave}) \subseteq \text{Range}(\nabla \mathcal{E}_{convex}). \tag{4.2}$$

In particular, the condition holds when $\text{Range}(\mathcal{E}_{convex}) = \mathbb{R}^n$, the entire space. The solvability condition (4.2) implies that, in some sense, the convex part must be *stronger* than or *dominant* over the concave part, which is often (practically) true for energy minimization problems when the function f is bounded below and $f \to \infty$ as $|\vec{x}| \to \infty$.

Second, in the present context, the CCCP should be applied to the functional setting instead of the function in $\mathbb{R}^n$. Therefore, the gradients in (4.1) should be naturally replaced by the Fréchet derivatives of the functionals. In our application, the functional is indeed Fréchet differentiable. (In general, the CCCP iteration (4.1) can also be based upon the subgradients of convex function(al)s, since the splitting yields convex components.)

4.2. Details of the computational scheme. To numerically solve (2.5),

$$\begin{aligned} E_\epsilon[z|u_o] &= F_\epsilon[z] + \lambda G[z|u_o] \\ &= \int_\Omega \left[\epsilon |\nabla z|^2 + \frac{1}{\epsilon} \sin^2 \pi z \right] dx + \lambda \sum_{k=0}^{K-1} \int_\Omega |C_k - u_o|^2 \text{sinc}^2(z-k)\, dx, \end{aligned}$$

we compute $E_\epsilon[z, \mathbf{C}|u_o]$ regarding z and $\mathbf{C}$ as independent variables. This allows the application of the alternating minimization (AM) scheme, i.e., to alternatingly optimize the two conditional energies $E_\epsilon[z|\mathbf{C}, u_o]$ and $E_\epsilon[\mathbf{C}|z, u_o]$, under the iterations of $z^n \to \mathbf{C}^n \to z^{n+1}$ given by

$$\mathbf{C}^n = \text{argmin}\, E_\epsilon[\mathbf{C}|z^n, u_o], \tag{4.3}$$

$$z^{n+1} = \text{argmin}\, E_\epsilon[z|\mathbf{C}^n, u_o]. \tag{4.4}$$

It is well known (i.e., Vogel [53] or Shen [44]) that the AM scheme is monotone:

$$E_\epsilon[z^{n+1}, \mathbf{C}^{n+1}|u_o] \le E_\epsilon[z^n, \mathbf{C}^n|u_o].$$

To minimize (4.3), one simply computes at the pixel level,

$$C_k = \frac{\sum_i \sum_j u_{i,j} \text{sinc}^2(z^n_{i,j} - k)}{\sum_i \sum_j \text{sinc}^2(z^n_{i,j} - k)}, \qquad k = 0, \ldots, K-1, \tag{4.5}$$

where $z^n_{i,j}$ denotes computational phase field on the Cartesian image domain. There is a study on treating C_k as an independent variable in image segmentation [50]. However, we update C_k in every alternating step, as in any usual AM schemes.

We apply the CCCP to minimize $E_\epsilon[z|\mathbf{C}^n, u_o]$ in (4.4). For convenience, we shall omit the superscript n of $\mathbf{C}^n$ hereafter. First, we add simple convex functionals to express E_ϵ as the difference of two convex functionals. By noticing that if f is a

convex function from $\mathbb{R}$ to $\mathbb{R}$, then the functional $F(u) = \int_\Omega f(u(x))dx$ is a convex functional, we have the following proposition.

PROPOSITION 4.3. *Let $F(u) = \int_\Omega f(u(x))dx$, where $f \in C^2(\mathbb{R})$ and $f'' \geq -\gamma$ for some $\gamma \geq 0$. Define the splitting*

$$F(u) = \int_\Omega \left(f(u) + \frac{\gamma}{2}u^2\right) dx - \int_\Omega \frac{\gamma}{2}u^2 dx := F^1(u) - F^2(u).$$

Then both F^1 and F^2 are convex.

The proof is trivial, since $f_1''(u) \geq 0$ if $f_1(u) = f(u) + \frac{\gamma}{2}u^2$. We now apply this splitting technique to the proposed model E_ϵ in (2.5).

We shall add two sets of terms, one for the nonconvex Modica–Mortola functional F_ϵ and the other for the nonconvex fitting term G. For the functional F_ϵ, we add $\frac{\pi^2}{\epsilon}\int_\Omega |z|^2\,dx$ by noticing that $\frac{d^2}{dz^2}\sin^2 \pi z \geq -2\pi^2$:

$$F_\epsilon[z] = \left(F_\epsilon[z] + \frac{\pi^2}{\epsilon}\int_\Omega |z|^2\,dx\right) - \frac{\pi^2}{\epsilon}\int_\Omega |z|^2\,dx := F_\epsilon^1[z] - F_\epsilon^2[z].$$

Similarly, the fitting term G in (2.5) can be split into $G[z|\mathbf{C}, u_o] = G^1[z|\mathbf{C}, u_o] - G^2[z|\mathbf{C}, u_o]$, where

$$G^1[z|\mathbf{C}, u_o] = G[z|\mathbf{C}, u_o] + \frac{\pi^2}{3}\sum_{k=0}^{K-1}\int_\Omega |u_o - C_k|^2|z-k|^2\,dx,$$
$$G^2[z|\mathbf{C}, u_o] = \frac{\pi^2}{3}\sum_{k=0}^{K-1}\int_\Omega |u_o - C_k|^2|z-k|^2\,dx,$$

since $\frac{d^2}{dz^2}\mathrm{sinc}^2 z \geq \frac{-2\pi^2}{3}$.

Thus, the functional (2.5) becomes $E_\epsilon = E_\epsilon^1 - E_\epsilon^2 = (F_\epsilon^1 + \lambda G^1) - (F_\epsilon^2 + \lambda G^2)$. We then apply the CCCP algorithm (4.1) via the Fréchet derivative:

$$(F_\epsilon^1 + \lambda G^1)'(z^{n+1}) = (F_\epsilon^2 + \lambda G^2)'(z^n). \tag{4.6}$$

Under integration by parts, (4.6) is equivalent to the PDE

(4.7)
$$\left[-2\epsilon\Delta z^{n+1} + \frac{\pi}{\epsilon}\sin 2\pi z^{n+1}\right] + \frac{2\pi^2}{\epsilon}z^{n+1}$$
$$+ \left[\lambda\sum_{k=0}^{K-1}|u_o - C_k|^2\frac{d}{dz}\mathrm{sinc}^2(z^{n+1}-k)\right] + \lambda\sum_{k=0}^{K-1}|u_o - C_k|^2\frac{2\pi^2}{3}(z^{n+1}-k)$$
$$= \frac{2\pi^2}{\epsilon}z^n + \lambda\sum_{k=0}^{K-1}|u_o - C_k|^2\frac{2\pi^2}{3}(z^n - k).$$

Here the terms in the square brackets come from the Euler–Lagrange equation of E_ϵ.

Numerically, the Laplacian term Δz^{n+1} is computed by the standard 5-pixel stencil, i.e., with h denoting the grid size,

$$h^2\Delta z = z_{i-1,j} + z_{i,j-1} + z_{i+1,j} + z_{i,j+1} - 4z_{i,j}.$$

This could further lead to the Jacobi-type iteration when the central pixel $z_{i,j}$ is assigned to the time step $n+1$, while the other four neighbors still stay at the step n.

We now elaborate on how to develop proper *linearization* schemes for the nonlinear terms that involve sine and sinc. For the second term with $\sin 2\pi z_{i,j}^{n+1}$, we use $\frac{\sin 2\pi z_{i,j}^n}{z^n} z_{i,j}^{n+1} = 2\pi z_{i,j}^{n+1} \mathrm{sinc}(2z_{i,j}^n)$ for linearization. This is inspired by the closely related problem of finding a solution to the nonlinear equation $\sin x = a$ for a given $a \in (0,1)$ and on $[0, \pi/2]$. An effective iteration scheme is given by the same linearization technique:

$$\frac{\sin x^n}{x^n} x^{n+1} = a, \quad \text{or equivalently,} \quad x^{n+1} = a\frac{x^n}{\sin x^n} = \Phi(x^n),$$

where $\Phi(x) = \frac{ax}{\sin x}$. A remarkable property is that for $x \in [0, \pi/2]$ and any given $a \in (0,1)$, Φ is a contractive mapping, i.e., $\max_{x\in[0,\pi/2]} |\Phi'(x)| \le a < 1$. In particular, the linearization iteration indeed leads to a unique fixed point x^*, $\frac{ax^*}{\sin x^*} = x^*$, by Picard's fixed point theorem. In addition, to avoid singularities in the denominators, small value δ, $\delta \ll 1$ (e.g., $\delta = 10^{-16}$, the MATLAB constant), is often added for numerical robustness.

For the fourth term on the left that involves the derivative of the sinc function, we similarly linearize it to

$$\frac{d}{dz}\mathrm{sinc}^2(z_{i,j}^{n+1} - k) = 2\,\mathrm{sinc}(z_{i,j}^n - k) \left[\frac{\pi z \cos \pi z - \sin \pi z}{\pi z^3}\right]_{z=z_{i,j}^n - k} (z_{i,j}^{n+1} - k).$$

Combining all the above steps of finite-difference discretization and function linearization, in the case when the Jacobi iteration is adopted for the Laplacian, we attain the following iteration scheme: at each step n,

$$\begin{aligned} &\left\{ 8\epsilon + \frac{\pi}{\epsilon}\left(\frac{\sin 2\pi z_{i,j}^n}{z_{i,j}^n} + 2\pi\right) \right. \\ &\quad \left. + \lambda \sum_{k=0}^{K-1} |u_{i,j} - C_k|^2 \left(\frac{2\pi^2}{3} + 2\mathrm{sinc}(z_{i,j}^n - k)\frac{\frac{d}{dz}\mathrm{sinc}(z_{i,j}^n - k)}{z_{i,j}^n - k}\right) \right\} z_{i,j}^{n+1} \\ &= 2\epsilon(z_{i-1,j}^n + z_{i,j-1}^n + z_{i+1,j}^n + z_{i,j+1}^n) + \frac{2\pi^2}{\epsilon} z_{i,j}^n \\ &\quad + \lambda \sum_{k=0}^{K-1} |u_{i,j} - C_k|^2 \left(\frac{2\pi^2}{3} z_{i,j}^n + 2k\mathrm{sinc}(z_{i,j}^n - k)\frac{\frac{d}{dz}\mathrm{sinc}(z_{i,j}^n - k)}{z_{i,j}^n - k}\right). \end{aligned} \tag{4.8}$$

The Neumann natural boundary condition is imposed along the boundary of the image domain. The detailed numerical analysis on the CCCP method augmented with all the above linearization techniques is interesting but foreseeably involved. This offers an intriguing open problem to the numerical analysis community.

Finally, once the phase field z is solved from the system, in order to extract each phase or segment, we apply the hard thresholding decision rule: $k - \frac{1}{2} \le z < k + \frac{1}{2}$ for each individual kth phase. A simple morphological transformation (opening) is also employed to remove any spurious dots due to the hard thresholding. Instead of hard thresholding, it is also possible to adopt a slightly more complex local decision rule based on windowing, which will then make the morphological operation obsolete.

4.3. Numerical experiments. In this subsection, we present some generic experimental results based on the theories and computational schemes developed above. In all the experiments, a given image is always normalized to the canonical gray interval $[0,1]$, and the bandwidth parameter (or the diffuse scale) ϵ is in the order of a

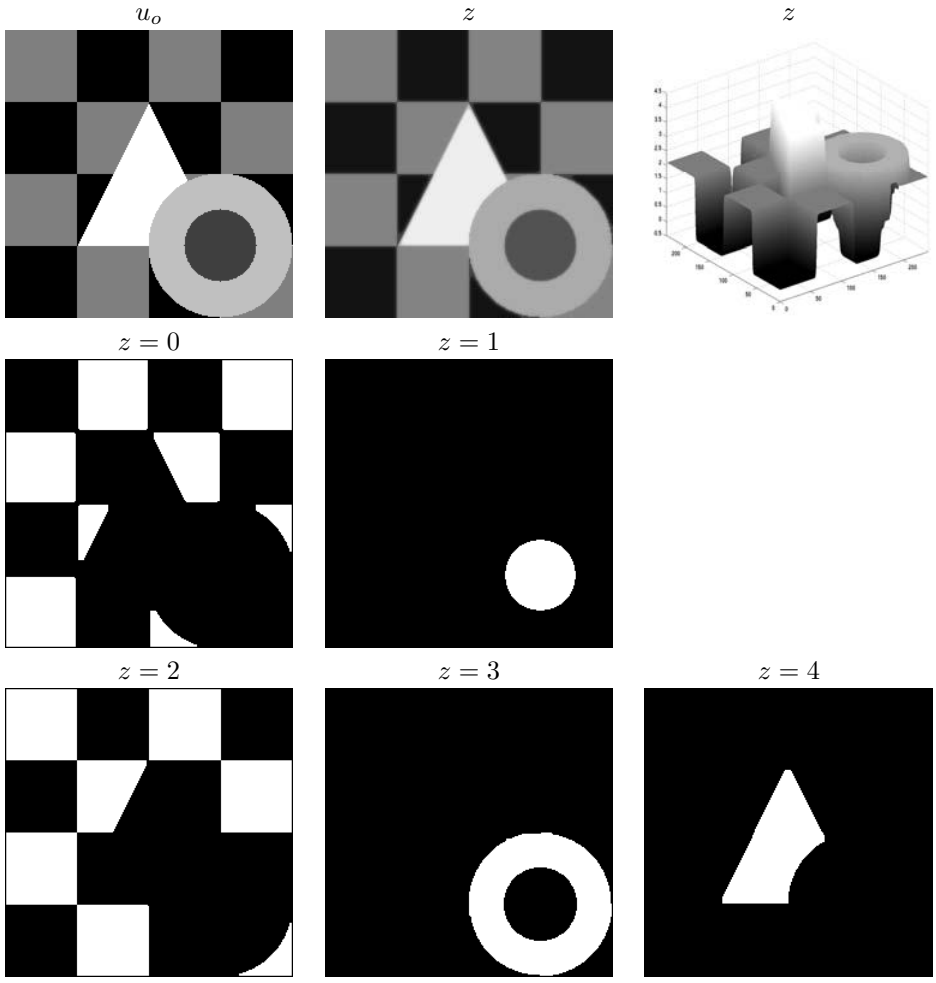


FIG. 4.1. *A complex synthetic image u_o with multiple objects and several generic visual structures. Image size is 240×240, $\epsilon = 2$ (pixels), and $\lambda = 15$ (scaled by the grid size). The calculated C_k values are $0, 0.16, 0.50, 0.74, 0.97$, respectively.*

few pixels. Furthermore, inspired by the simulated annealing technique in stochastic image processing (see, e.g., Geman and Geman [23]) and Gibbs' random fields, we have also experimented with dynamically decreasing ϵ's to speed up convergence, for example, adopting ϵ_1 in the first 50 iterations, while $\epsilon_2 = \epsilon_1/2$ for the rest.

Regarding the initial guess for the phase field z, we have typically adopted random values between -0.5 to $K + 0.5$ as mentioned in the theory (the set $\mathcal{A}_K$). For complex images with large variances in homogeneous regions or with many phases, weak supervision can be used for the initial values; i.e., initial C_k values can be estimated from the assigned supervised "seed" regions. For more discussion on weak supervision and automated stochastic supervision (based on patch statistics), we refer the reader to the recent works of Shen [46], Li et al. [28], and Li and Perona [27].

The first example, Figure 4.1, shows a complex synthetic image that contains several generic visual structures, including an internal hole, occlusion and stacked objects, T-junctions, singular junctions where multiple objects (or phases) meet, as

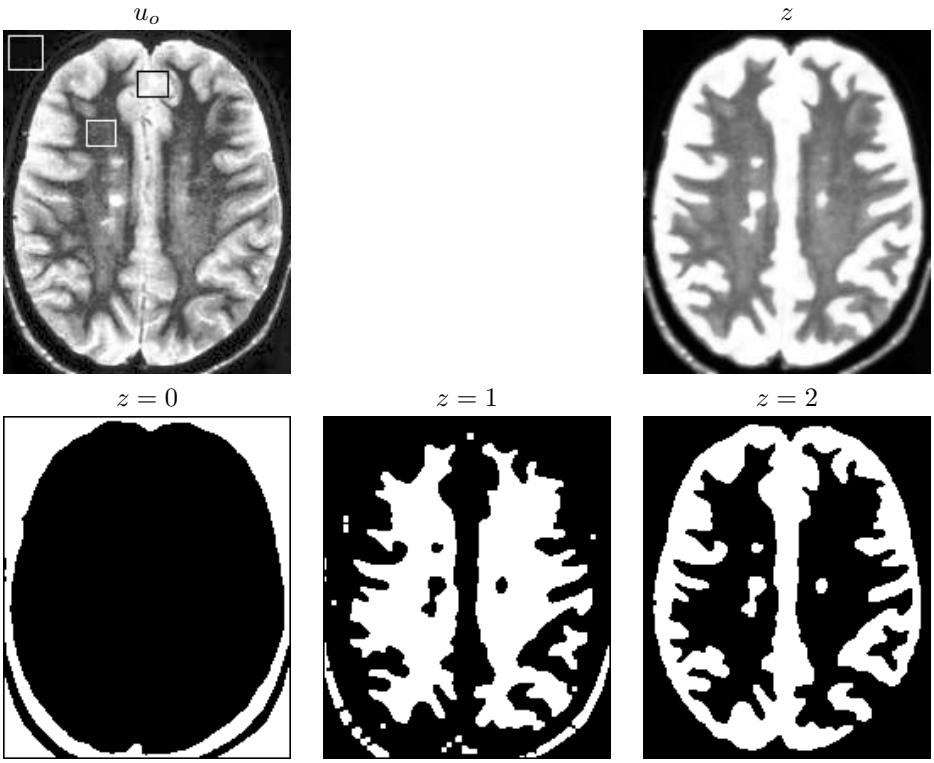


FIG. 4.2. *The performance of the proposed Modica–Mortola sine-sinc model to the segmentation of an MRI brain image. Shown on top of the original image (upper left) are the three "seed" phase patches which are often easily supervised by a radiologist. The resolution details in the segmented phases depend upon the bandwidth parameter* ϵ *(*$\epsilon = 2$ *(pixels) for this particular output).*

well as thin passages that reveal the bottleneck effect (see the recent work of Kohn and Slastikov [25] in material sciences for asymptotic bottleneck analysis in phase transitions). In this case, with $K = 5$ phases, initial condition is weakly supervised.

Figure 4.2 shows the application of the proposed model to an MRI brain image. Even though the intensities fluctuate severely and the boundaries are complex, the proposed method has done a satisfactory job in separating the major different phases. Shown on top of the original image are the three small patches that are in practice easily supervised by a radiologist.

In Figure 4.3, it is shown that the proposed model works well with a noisy image containing a generic T-junction, a universal singular structure crucial in visual perception (see, e.g., Nitzberg, Mumford, and Shiota [37]). The level of noise is exaggerated to show the stability of the proposed method.

The next couple of examples, Figures 4.4 and 4.5, involve color images for which the RGB color space has been employed. We have adopted the Euclidean metric of three color channels as in [51]:

$$|u_o^R - C_k^R|^2 + |u_o^G - C_k^G|^2 + |u_o^B - C_k^B|^2,$$

where u_o^R, u_o^G, u_o^B correspond to the red, green, and blue channels of the given color image u_o. (This may not be optimal for color perception; see, e.g., [10, 12]). In addition, the following two examples deal with either blurry edges or the absence of

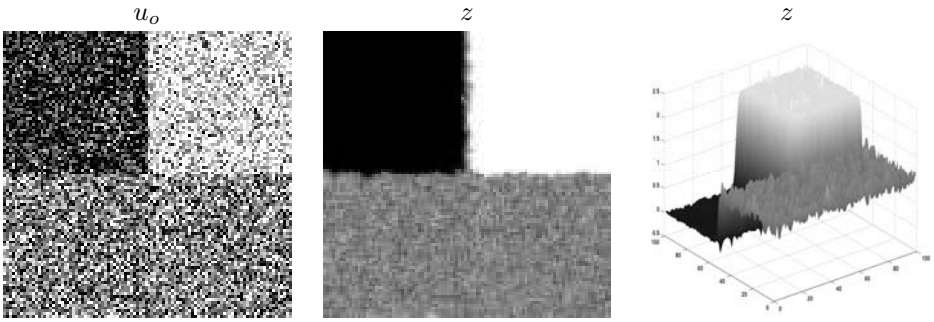


FIG. 4.3. *A noisy synthetic image u_o containing a generic T-junction (left); the phase field z computed by the proposed sine-sinc model (middle and right). The example shows that the model is robust to noise and reconstructs well the geometry.*

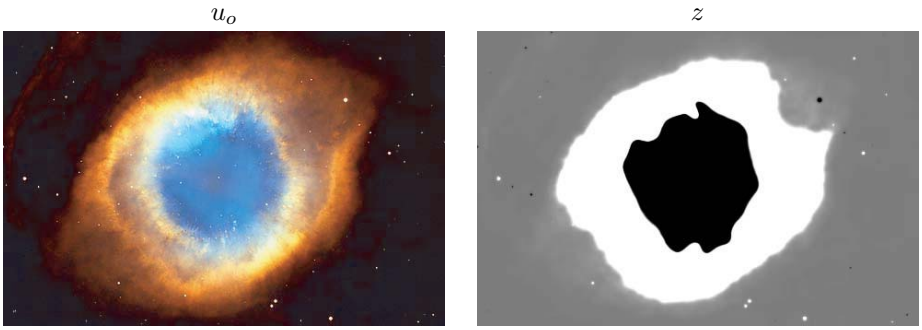


FIG. 4.4. *Example of image with blurry edges. The image of Helix nebula has blurry edges, and the proposed model captures each different level clearly: three phases $z = 0$, 1, and 2.*

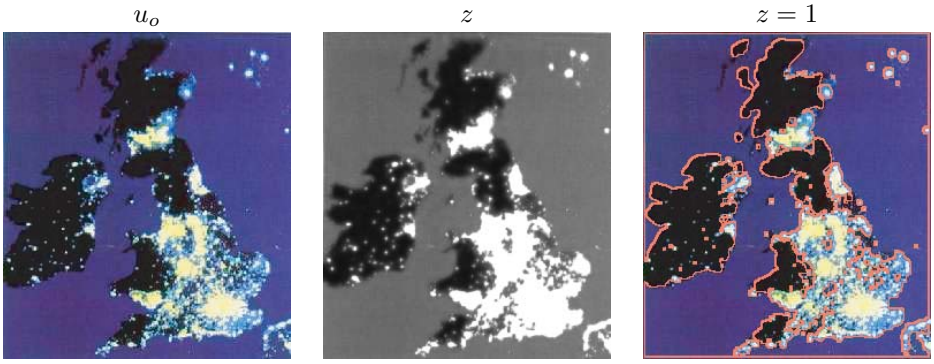


FIG. 4.5. *Example of cluster segmentation. The original (left) and z with three phases (middle) are shown. The $z = 1$ contour is superposed over the original to show the accuracy (right).*

edges. Figure 4.4 is an example of blurring image. The edges are not clearly defined; however, the proposed Modica–Mortola sine-sinc model captures each different level clearly. Figure 4.5 shows an example of cluster segmentation. It is an image of the night light of England and correctly segments three different phases. These two examples further demonstrate the flexibility of the proposed model and its computational algorithm.

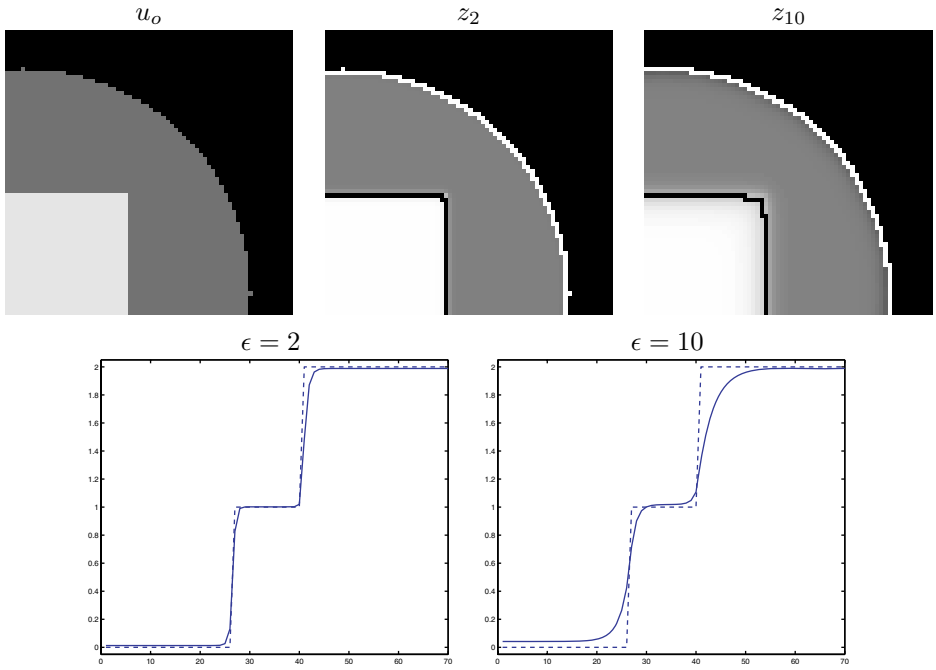


FIG. 4.6. *Effect of different ϵ values. From the same original image u_o, two different ϵ values are used to show the effect of ϵ. In the top row, segmented results z_2 and z_{10} are superposed with black and white contours identifying each segment's $z = 0$, $z = 1$, and $z = 2$. The second row plots show the profile (solid) of the diagonal of z_2 and z_{10} compared to the ideal segmentation (dotted).*

As a final example, we present the effect of using different ϵ in Figure 4.6. From the original image u_o, two different ϵ values are used to show the effect of having different transition bands: as ϵ gets bigger, the transition band gets wider. In Figure 4.6, except for the sharp corner, z_2 and z_{10} show little differences.

5. Conclusion. In this paper, we propose a new multiphase segmentation model based on the celebrated phase transition model of Modica and Mortola [33] in material sciences, fluid mechanics, and the Γ-convergence theory. The sine-sinc model properly synchronizes the fitting term for the given image with the regularity term for the diffuse interfaces. Mathematical analysis is developed for the Γ-convergence behavior of the model and the existence of its minimizers. We also develop in detail the convex-splitting or the CCCP algorithm for minimizing the nonconvex energy functional. Several numerical experiments on both generic synthetic and natural images demonstrate the satisfying performance of the proposed model and its algorithm.

It is our belief that the interplay and integration between physics and information technologies will further blossom in the near future. The current work is a typical example that has substantially benefited from numerous existing contributions in these two fields.

Acknowledgments. Jung is grateful to Prof. Jin Keun Seo for his support and encouragement. Kang would like to thank IMA for supporting a long term visit during the thematic year on imaging which has made this collaboration possible. Shen would like to thank his dear teacher and friend Riccardo March for kindly educating him on the Γ-convergence theory with fun and joy through all these years, and also dear

friends Alan Yuille and Andrea Bertozzi for the enlightenment on the convex splitting or the CCCP algorithm.

REFERENCES

[1] G. Alberti, *Variational models for phase transitions, an approach via Γ-convergence*, in Calculus of Variations and Partial Differential Equations (Pisa, 1996), Springer-Verlag, Berlin, 2000, pp. 95–114.

[2] L. Ambrosio and V. M. Tortorelli, *Approximation of functionals depending on jumps by elliptic functionals via Γ-convergence*, Comm. Pure Appl. Math., 43 (1990), pp. 999–1036.

[3] L. Ambrosio and V. M. Tortorelli, *On the approximation of free discontinuity problems*, Boll. Un. Mat. Ital. B(7), 6 (1992), pp. 105–123.

[4] A. Bertozzi, S. Esedoglu, and A. Gilette, *Inpainting of binary images using the Cahn-Hilliard equation*, IEEE Trans. Image Process., 16 (2007), pp. 285–291.

[5] B. Bourdin and A. Chambolle, *Implementation of an adaptive finite-element approximation of the Mumford-Shah functional*, Numer. Math., 85 (2000), pp. 609–646.

[6] A. Braides, *Γ-Convergence for Beginners*, Oxford Lecture Ser. Math. Appl. 22, Oxford University Press, Oxford, UK, 2002.

[7] J. W. Cahn and J. E. Hilliard, *Free energy of a non-uniform system* I. *Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.

[8] A. Chambolle, *Image segmentation by variational methods: Mumford and Shah functional and the discrete approximations*, SIAM J. Appl. Math., 55 (1995), pp. 827–863.

[9] A. Chambolle, *Finite-differences discretizations of the Mumford-Shah functional*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 261–288.

[10] T. F. Chan, S.-H. Kang, and J. Shen, *Total variation denoising and enhancement of color images based on the CB and HSV color models*, J. Visual Comm. Image Rep., 12 (2001), pp. 422–435.

[11] T. F. Chan, J. Shen, and L. Vese, *Variational PDE models in image processing*, Amer. Math. Soc. Notice, 50 (2003), pp. 14–26.

[12] T. Chan and J. Shen, *Variational restoration of nonflat image features: Models and algorithms*, SIAM J. Appl. Math., 61 (2000), pp. 1338–1361.

[13] T. F. Chan and L. Vese, *A level set algorithm for minimizing the Mumford-Shah functional in image processing*, in Proceedings of the 1st IEEE Workshop on Variational and Level Set Methods in Computer Vision, 2001, pp. 161–168.

[14] T. F. Chan and J. Shen, *Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods*, SIAM, Philadelphia, 2005.

[15] T. F. Chan and L. A. Vese, *Active contours without edges*, IEEE Trans. Image Process., 10 (2001), pp. 266–277.

[16] T. F. Chan and L. A. Vese, *A multiphase level set framework for image segmentation using the Mumford and Shah model*, Int. J. Comp. Vision, 50 (2002), pp. 271–293.

[17] J. T. Chung and L. A. Vese, *Image segmentation using a multilayer level-set approach*, UCLA CAM report 03-53, UCLA, Los Angeles, CA, 2003.

[18] F. Cucker and S. Smale, *On the mathematical foundations of learning*, Bull. Amer. Math. Soc. (N.S.), 39 (2001), pp. 1–49.

[19] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

[20] S. Esedoglu and R. March, *Segmentation with depth but without detecting junctions*, J. Math. Imaging Vision, 18 (2003), pp. 7–15.

[21] S. Esedoglu and J. Shen, *Digital inpainting based on the Mumford-Shah-Euler image model*, European J. Appl. Math., 13 (2002), pp. 353–370.

[22] D. Eyre, *An Unconditionally Stable One-Step Scheme for Gradient Systems*, http://www.math.utah.edu/∼eyre/research/methods/stable.ps (1998).

[23] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Machine Intell., 6 (1984), pp. 721–741.

[24] E. De Giorgi and T. Franzoni, *Su un tipo di convergenza variazionale*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8), 58 (1975), pp. 842–850.

[25] R. V. Kohn and V. V. Slastikov, *Geometrically constrained walls*, Calc. Var. Partial Differential Equations, 28 (2007), pp. 33–57.

[26] J. Lie, M. Lysaker, and X.-C. Tai, *A variant of the level set method and applications to image segmentation*, Math. Comp., 75 (2006), pp. 1155–1174.

[27] F.-F. Li and P. Perona, *A Bayesian hierarchical model for learning natural scene categories*, IEEE CVPR, 2 (2005), pp. 524–531.

[28] F.-F. Li, R. VanRullen, C. Koch, and P. Perona, *Rapid natural scene categorization in the near absence of attention*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 9596–9601.
[29] R. March and M. Dozio, *A variational method for the recovery of smooth boundaries*, Image Vision Comput., 15 (1997), pp. 705–712.
[30] R. March, *Visual reconstruction with discontinuities using variational methods*, Image Vision Comput., 10 (1992), pp. 30–38.
[31] G. Dal Maso, *An Introduction to* Γ*-Convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser Boston, Boston, MA, 1993.
[32] Y. Meyer, *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations*, Univ. Lecture Ser. 22, AMS, Providence, RI, 2001.
[33] L. Modica and S. Mortola, *Un esempio di* Γ^-*-convergenza*, Boll. Un. Mat. Ital. B (5), 14 (1977), pp. 285–299.
[34] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal., 98 (1987), pp. 123–142.
[35] J.-M. Morel and S. Solimini, *Variational Methods in Image Segmentation*, Progr. Nonlinear Differential Equations Appl. 14, Birkhäuser Boston, Boston, MA, 1995.
[36] D. Mumford and J. Shah, *Optimal approximations by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math., 42 (1989), pp. 577–685.
[37] M. Nitzberg, D. Mumford, and T. Shiota, *Filtering, Segmentation, and Depth*, Lecture Notes in Comput. Sci. 662, Springer-Verlag, Berlin, 1993.
[38] S. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, Appl. Math. Sci. 153, Springer-Verlag, New York, 2003.
[39] S. Osher and J. A. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.
[40] S. Osher, A. Solé, and L. Vese, *Image decomposition and restoration using total variation minimization and the* H^{-1} *norm*, Multiscale Model. Simul., 1 (2003), pp. 349–370.
[41] T. Poggio and S. Smale, *The mathematics of learning: Dealing with data*, Amer. Math. Soc. Notice, 50 (2003), pp. 537–544.
[42] B. Sandberg, T. Chan, and L. Vese, *A Level-Set and Gabor-Based Active Contour Algorithm for Segmenting Textured Images*, UCLA CAM report 02-39, UCLA, Los Angeles, CA, 2002.
[43] J. A. Sethian, *Level Set Methods and Fast Marching Methods*, Cambridge Monogr. Appl. Comput. Math. 3, 2nd ed., Cambridge University Press, Cambridge, UK, 1999.
[44] J. Shen, *Bayesian video dejittering by the BV image model*, SIAM J. Appl. Math., 64 (2004), pp. 1691–1708.
[45] J. Shen, *Piecewise* $H^{-1}+H^0+H^1$ *images and the Mumford-Shah-Sobolev model for segmented image decomposition*, AMRX Appl. Math. Res. Express, 4 (2005), pp. 143–167.
[46] J. Shen, *A stochastic-variational model for soft mumford-shah segmentation*, International Journal of Biomedical Imaging, Vol. 2006 (2006), article ID92329.
[47] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Trans. Pattern Anal. Machine Intell., 22 (2000), pp. 888–905.
[48] S. Smale and D.-X. Zhou, *Shannon sampling and function reconstruction from point values*, Bull. Amer. Math. Soc. (N.S.), 41 (2004), pp. 279–305.
[49] X.-C. Tai and T. F. Chan, *A survey on multiple level set methods with applications for identifying piecewise constant functions*, Int. J. Numer. Anal. Model., 1 (2004), pp. 25–48.
[50] X.-C. Tai and C.-H.Yao, *Image segmentation by piecewise constant Mumford-Shah model without estimating the constants*, J. Comput. Math., 24 (2006), pp. 435–443.
[51] L. A. Vese, T. F. Chan, and B. Y. Sandberg, *Active contours without edges for vector-valued images*, J. Visual Comm. Image Rep., 11 (2000), pp. 130–141.
[52] Z. W. Tu and S. C. Zhu, *Image segmentation by data-driven Markov Chain Monte Carlo*, IEEE Trans. Pattern Anal. Machine Intell., 24 (2002), pp. 657–673.
[53] C. R. Vogel, *Computational Methods for Inverse Problems*, SIAM, Philadelphia, 2002.
[54] B. P. Vollmayr-Lee and A. D. Rutenberg, *Fast and accurate coarsening simulation with an unconditionally stable time step*, Phys. Rev. E (3), 68 (2003), 066703.
[55] A. L. Yuille and A. Rangarajan, *The concave-convex procedure (CCCP)*, Neural Comput., 15 (2003), pp. 915–936.

SIAM J. APPL. MATH.
Vol. 67, No. 5, pp. 1233–1258

SCATTERING BY A SEMI-INFINITE PERIODIC ARRAY AND THE EXCITATION OF SURFACE WAVES*

C. M. LINTON†, R. PORTER‡, AND I. THOMPSON†

Abstract. The two-dimensional problem of acoustic scattering of an incident plane wave by a semi-infinite array of either rigid or soft circular scatterers is solved. Solutions to the corresponding infinite array problems are used, together with a novel filtering approach, to enable accurate solutions to be computed efficiently. Particular attention is focused on the determination of the amplitude of the Rayleigh–Bloch waves that can be excited along the array. In general, the far field away from the array consists of the sum of a finite number of plane waves propagating in different directions (the number depending on the observation angle) and a circular wave emanating from the edge of the array. In certain *resonant* cases (characterized by one of the scattered plane waves propagating parallel to the array), a different far field pattern occurs, involving contributions that are neither circular waves nor plane waves. Uniform asymptotic expansions that vary continuously across all of the shadow boundaries that exist are given for both cases.

Key words. acoustic scattering, semi-infinite array, surface wave

AMS subject classifications. 74J20, 74J15, 78A45

DOI. 10.1137/060672662

1. Introduction.

Large array scattering problems are of considerable current interest in many different areas and present significant theoretical and computational challenges. Whereas wave scattering by a small number of scatterers, or by an infinite periodic array, is fairly well understood, scattering by large but finite arrays has received much less attention. Scattering by large finite arrays is of considerable importance in the theory of array antennas and the fabrication of electromagnetic band gap materials [1], [2], [3]; in water waves [4], [5], where offshore structures supported by thousands of cylindrical columns are being designed; and in acoustics, where large periodic arrays continue to be the subject of numerous studies—applications include acoustic filters, noise control, and the design of transducers.

It has long been recognized that one way to approach large finite array scattering is to analyze the effects of each edge of the array in isolation—in other words, to study arrays with just one edge—and this leads to problems formulated on semi-infinite arrays. On the assumption that opposite edges of a finite array are well separated, results from analyses of semi-infinite arrays can then be combined to provide results for finite arrays. Unfortunately, such problems are difficult to analyze and little work has been done on the subject since the pioneering studies of Hills and Karp [6] and Millar [7].

For the case of a semi-infinite periodic array of isotropic point scatterers, it has been shown [8] that progress can be made if the problem is formulated for the difference between unknowns relevant to the infinite and semi-infinite array problems. The situation considered was appropriate to acoustic diffraction by sound-soft scatterers

*Received by the editors October 18, 2006; accepted for publication (in revised form) February 16, 2007; published electronically June 15, 2007.
http://www.siam.org/journals/siap/67-5/67266.html
†Department of Mathematical Sciences, Loughborough University, Leicestershire, LE11 3TU, UK (c.m.linton@lboro.ac.uk, i.thompson@lboro.ac.uk). The research of the third author was supported by the EPSRC under grant EP/C510941/1.
‡Department of Mathematics, University of Bristol, Bristol, BS8 1TW, UK (richard.porter@bristol.ac.uk).

in the limit as the ratio of wavelength to body size tends to infinity and also to the scattering of an E-polarized electromagnetic wave by an array of perfectly conducting wires. Important as these applications are, they do not cover many of the cases in which large array scattering is a serious issue.

As a significant extension we consider here two-dimensional scattering by a semi-infinite row of periodically spaced, identical circular cylinders and show how the diffracted field can be efficiently computed. We extend the techniques developed in [8] so as to investigate the effects of the size of scatterers and the boundary conditions applied on them. Neumann boundary conditions appropriate for rigid bodies give rise to a major complication since diffraction gratings of rigid structures are known to support pure Rayleigh–Bloch surface waves at low frequencies [9], [10] and these may be excited by the edge of the array.

The excitation of surface waves by array edges is a virtually unexplored area, partly because there are very few geometries for which the range of possible Rayleigh–Bloch modes (also called array guided surface waves) is completely understood. They have been observed numerically in arrays of dipoles [2] and are akin to the edge waves that can be excited by the edge of a semi-infinite crack in a thin plate [11]. It is common practice in many applications to assume that the behavior of a large finite array can be approximated well by an infinite array, at least away from the edges. One of the consequences of the presence of Rayleigh–Bloch surface waves is that this is no longer valid. For example, Maniar and Newman [12] showed that the effect of these modes (especially those which are close to standing modes) can be extremely important, giving rise to enormous amplification of the wave field close to the center of a large array. It is thus important to have a good understanding of when and to what extent array guided surface waves are excited.

The structure of the paper is as follows. We begin in section 2 by formulating the scattering problems for both an infinite and a semi-infinite array of circular scatterers using separation of variables. Details of the Rayleigh–Bloch surface waves that can be supported by an infinite array are given. In section 3 the solution to the infinite array problem is used to reformulate the semi-infinite array problem in such a way that the unknown coefficients associated with each cylinder decay to zero as one moves away from the array edge. A number of different approaches are considered, and they are used to compute the amplitude of the Rayleigh–Bloch waves that are excited. The nature of the far field is analyzed in section 4, and a uniform asymptotic approximation is derived. A special treatment is required when the parameters correspond to resonance in the infinite array (when one of the scattered waves propagates along the array), and this is given in section 5. In particular, this allows us to solve the semi-infinite problem in the case of head-on incidence.

2. Formulation. We consider a two-dimensional scattering problem which has application in a number of physical contexts. We will refer primarily to the acoustic setting in which we look for time-harmonic solutions $\mathrm{Re}[\phi(x,y)\exp(-\mathrm{i}\omega t)]$ so that the acoustic potential ϕ satisfies the two-dimensional Helmholtz equation $(\nabla^2+k^2)\phi=0$ in the region exterior to the scatterers, where $k=\omega/c$ and c is the speed of sound. The scatterers can be taken as either rigid (in which case the normal derivative of ϕ must vanish on the boundary; we call this the Neumann problem) or acoustically soft (in which case the appropriate boundary condition is $\phi=0$; we call this the Dirichlet problem). Exactly the same boundary-value problem can be used to study electromagnetic diffraction by an array of perfect conductors or, once the depth variation has been factored out, the scattering of water waves by vertical circular cylinders. In

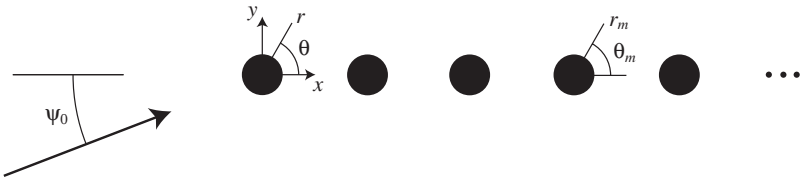


FIG. 2.1. *Definition sketch.*

the latter case, k is the positive solution to the dispersion relation $k \tanh kh = \omega^2/g$, h being the water depth and g the acceleration due to gravity; the appropriate boundary condition is $\partial\phi/\partial n = 0$. The geometry under consideration is sketched in Figure 2.1.

We are concerned with the scattering of a plane wave

$$\phi_{\text{inc}} = \mathrm{e}^{\mathrm{i}(\lambda x + \mu y)}, \tag{2.1}$$

where $\mu = k \sin\psi_0$ and $\lambda = k \cos\psi_0$, by a semi-infinite row of identical circular cylinders of radius a, located at $(x, y) = (j, 0)$, $j = 0, 1, 2, \ldots$. The spacing between the cylinders has been set to unity for convenience, and hence $0 < a \leq 0.5$. We will use polar coordinates (r_j, θ_j), centered on the jth scatterer and defined by

$$x - j = r_j \cos\theta_j, \qquad y = r_j \sin\theta_j, \tag{2.2}$$

and we will usually write (r, θ) for (r_0, θ_0). In terms of (r_j, θ_j) the incident wave is given by

$$\phi_{\text{inc}} = \mathrm{e}^{\mathrm{i}\lambda j} \mathrm{e}^{\mathrm{i}kr_j \cos(\theta_j - \psi_0)}. \tag{2.3}$$

This problem can be formulated using separation of variables. If we write the total field as $\phi = \phi_{\text{inc}} + \phi_{\text{sc}}$ with

$$\phi_{\text{sc}} = \sum_{j=0}^{\infty} \sum_{n=-\infty}^{\infty} A_n^j Z_n \, H_n(kr_j) \mathrm{e}^{\mathrm{i}n\theta_j}, \tag{2.4}$$

where $H_n(\cdot)$ is a Hankel function of the first kind and $Z_n = J_n(ka)/\, H_n(ka)$ if the Dirichlet problem is being studied, or $Z_n = J_n'(ka)/\, H_n'(ka)$ for the Neumann problem, then the unknowns A_n^j are solutions to ([13, eq. (2.11)])

$$A_m^p + \sum_{n=-\infty}^{\infty} Z_n \sum_{\substack{j=0 \\ \neq p}}^{\infty} A_n^j X_{n-m}^{jp} \, H_{n-m}(k|j-p|) = -\mathrm{e}^{\mathrm{i}\lambda p} \mathrm{e}^{\mathrm{i}m(\frac{1}{2}\pi - \psi_0)}, \tag{2.5}$$

$p = 0, 1, 2, \ldots, m \in \mathbb{Z}$, where $X_n^{jp} = 1$ if $p > j$ and $X_n^{jp} = (-1)^n$ if $p < j$.

This system of equations could, in principle, be solved numerically by truncation, but the infinite spatial sum (over j) converges extremely slowly. (The coefficients A_n^j do not decay to zero as $j \to \infty$, and so the terms in this sum decay like $j^{-1/2} \exp(\mathrm{i}j\delta)$ for some δ.) By contrast, the order summation (over n) converges exponentially. The strategy that is followed here is to make use of known properties of the diffraction problem when the array extends to both plus and minus infinity to allow us to sum up the slowly convergent spatial series analytically.

2.1. The infinite array. For the infinite grating problem, with cylinders at $(j,0)$, $j\in\mathbb{Z}$, we can seek a solution of the form

$$\phi_{\rm sc}^{\rm inf}=\sum_{j=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}B_n^jZ_n\,H_n(kr_j)\mathrm{e}^{\mathrm{i}n\theta_j}. \tag{2.6}$$

The periodicity of the geometry and of the incident wave allows us to look for a solution which satisfies

$$B_n^j=\mathrm{e}^{\mathrm{i}\lambda j}B_n^0=\mathrm{e}^{\mathrm{i}\lambda j}B_n, \tag{2.7}$$

say, and then we need only solve for B_n. These coefficients are solutions to the infinite system of equations

$$B_m+\sum_{n=-\infty}^{\infty}B_nZ_n\sigma_{n-m}(\lambda)=-\mathrm{e}^{\mathrm{i}m(\frac12\pi-\psi_0)},\qquad m\in\mathbb{Z}, \tag{2.8}$$

where

$$\sigma_n(\lambda)=\sum_{j=1}^{\infty}\left[(-1)^n\mathrm{e}^{\mathrm{i}\lambda j}+\mathrm{e}^{-\mathrm{i}\lambda j}\right]H_n(kj). \tag{2.9}$$

The quantities σ_n are easily evaluated (though not from the above expression); see [14], [15]. The system (2.8) is straightforward to solve numerically by truncation, the convergence of B_n with $|n|$ being exponential.

The far field for the infinite array problem can be determined as follows. First we define the scattering angles $\psi_m(\lambda)$ by

$$\psi_m=\arccos(\lambda_m/k),\qquad \lambda_m=\lambda+2m\pi. \tag{2.10}$$

If $|\lambda_m|<k$, i.e.,

$$-1<\cos\psi_0+\frac{2m\pi}{k}<1, \tag{2.11}$$

then we say that $m\in\mathcal{M}$ and we have $0<\psi_m<\pi$. If $|\lambda_m|>k$, then ψ_m is no longer real and the appropriate branch of the arccos function is given by

$$\arccos t=\begin{cases}\mathrm{i}\,\mathrm{arccosh}\,t, & t>1,\\ \pi-\mathrm{i}\,\mathrm{arccosh}(-t), & t<-1,\end{cases} \tag{2.12}$$

with $\mathrm{arccosh}\,t=\ln\left(t+\sqrt{t^2-1}\right)$ for $t>1$. Next we use the integral representation

$$H_n(kr)\,\mathrm{e}^{\mathrm{i}n\theta}=\frac{(-\mathrm{i})^{n+1}}{\pi}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{-k\gamma(t)|y|}}{\gamma(t)}\,\mathrm{e}^{\mathrm{i}kxt}\,(t-\gamma(t))^{n\,\mathrm{sgn}(y)}\,\mathrm{d}t, \tag{2.13}$$

in which $\gamma(t)=(t^2-1)^{1/2}$ with $\gamma(0)=-\mathrm{i}$ and the path of integration is indented so as to pass above the branch point at $t=-1$ and below that at $t=1$ (for $n=0$, see [8, Appendix A]; for the extension to all n we use [16, Theorem 2.7]). If this is inserted into (2.6), we can apply the Poisson summation formula to obtain

$$\phi_{\rm sc}^{\rm inf}=\sum_{m=-\infty}^{\infty}\mathcal{F}_m^{\pm}\mathrm{e}^{\mathrm{i}kr\cos(\theta\mp\psi_m)}, \tag{2.14}$$

in which

$$\mathcal{F}_m^{\pm} = \frac{2}{k} \sum_{n=-\infty}^{\infty} (-\mathrm{i})^n B_n Z_n \frac{\mathrm{e}^{\pm \mathrm{i}n\psi_m}}{\sin \psi_m} \tag{2.15}$$

and the superscripts + and − correspond to $y > 0$ and $y < 0$, respectively. The integral representation (2.13) with $n \neq 0$ is valid except on $y = 0$, and thus this expression for the field is valid everywhere outside the scatterers except on $y = 0$, provided that $\sin \psi_m \neq 0$ for any m. If there is a value of m for which $\sin \psi_m = 0$, then the scattering problem is described as resonant and requires a separate treatment. This is discussed elsewhere [17]. As $y \to \pm\infty$, the only contribution comes from those m for which ψ_m is real, i.e., $m \in \mathcal{M}$. Hence the far field consists of a set of plane waves propagating in the directions $\theta = \psi_m$ and $\theta = 2\pi - \psi_m$:

$$\phi_{\mathrm{sc}}^{\mathrm{inf}} \sim \sum_{m \in \mathcal{M}} \mathcal{F}_m^{\pm} \mathrm{e}^{\mathrm{i}kr\cos(\theta \mp \psi_m)} \quad \text{as} \quad y \to \pm\infty. \tag{2.16}$$

2.2. Rayleigh–Bloch surface waves. Crucial to what follows is the fact that in the Neumann problem the solution to the scattering problem described above may not be unique. If we relax the quasi-periodicity condition (2.7) and replace it with another with a different phase, it may be possible to find a value β (maybe more than one), dependent on k, such that the homogeneous infinite system

$$B_m + \sum_{n=-\infty}^{\infty} B_n Z_n \sigma_{n-m}(\beta) = 0, \qquad m \in \mathbb{Z}, \tag{2.17}$$

has a nontrivial solution. The resulting potential

$$\phi^{\mathrm{rb}} = \sum_{j=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\beta j} B_n Z_n \, H_n(kr_j) \mathrm{e}^{\mathrm{i}n\theta_j} \tag{2.18}$$

does not share the same periodicity as the incident wave, but nevertheless satisfies all the boundary conditions of the full problem. Such potentials are referred to as Rayleigh–Bloch surface waves (or array guided surface waves), and for the geometry under consideration here the dispersion relation connecting β and k has been computed in [18] and [19] (the existence of these surface waves was proved in [10]). Since $\exp(\mathrm{i}\beta m) = \exp(\mathrm{i}(\beta + 2\pi)m)$, we can restrict our attention to $0 \leq \beta < 2\pi$. It follows from (2.17) that if there is a solution for a given β, then there is also a solution with β replaced by $2\pi - \beta$ (representing a wave whose energy is travelling in the opposite direction). If we insist that energy is propagating in the positive x-direction, as it will be in the semi-infinite array problem considered below, then we can restrict attention to $0 < \beta < \pi$. The numerical results in [18] and [19] show that Rayleigh–Bloch surface waves exist at discrete values of k for any $\beta < \pi$ and that they satisfy $k < \beta$. The fact that no such modes exist in the Dirichlet problem is proved in [20].

Computations show that a mode which is symmetric about $y = 0$ (for which $B_{-n} = (-1)^n B_n$) exists for all scatterer sizes, and a mode antisymmetric about $y = 0$ (for which $B_{-n} = -(-1)^n B_n$, with $B_0 = 0$) exists for $0.403 \lesssim a \leq 0.5$. For a given value of a, Rayleigh–Bloch waves exist only for a range of values of k: symmetric modes in the range $0 < k < k_{\mathrm{max}}^{\mathrm{s}} < \pi$ and antisymmetric modes in the range $k_{\mathrm{min}}^{\mathrm{a}} < k < k_{\mathrm{max}}^{\mathrm{a}} < \pi$. It turns out that there are three distinct regimes:

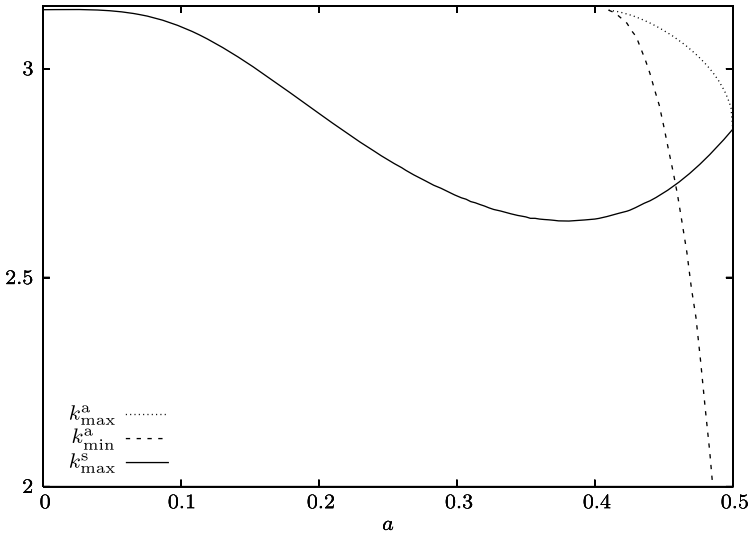


FIG. 2.2. $k_{\rm min}$ *and* $k_{\rm max}$ *for symmetric and antisymmetric Rayleigh–Bloch modes.*

for $a \lesssim 0.403$, only symmetric modes are possible; for $0.403 \lesssim a \lesssim 0.459$, we have $k^{\rm s}_{\rm max} < k^{\rm a}_{\rm min}$, and thus it is possible to have symmetric and antisymmetric modes, but not for the same value of k; and finally, when $0.459 \lesssim a < 0.5$, we have $k^{\rm s}_{\rm max} > k^{\rm a}_{\rm min}$, and hence it is only in this parameter range that it is possible to excite both symmetric and antisymmetric modes at the same time. Figure 2.2 shows values of $k^{\rm s}_{\rm max}$, $k^{\rm a}_{\rm min}$, and $k^{\rm a}_{\rm max}$ for varying scatterer radius a. When $a = 0.5$, the symmetric and antisymmetric Rayleigh–Bloch modes are essentially the same since the cylinders are touching and there is no connection between the two sides of the array.

3. Infinite array subtraction. We will formulate the problem allowing for the excitation of a single Rayleigh–Bloch mode. If both symmetric and antisymmetric modes are present, then this is easily accommodated by first splitting the problem into parts symmetric and antisymmetric about $y = 0$ and treating each of them separately. In fact it is numerically efficient to make this decomposition irrespective of the parameter values, and this was done in all the computations presented below. For the semi-infinite grating we would like to construct an infinite system of equations in which, unlike in (2.5), the unknowns decay to zero as one moves along the array. To this end we first introduce new unknowns which are the differences between the solutions to the infinite and semi-infinite array problems (as in [8], [21]). There are then a number of different ways in which the Rayleigh–Bloch waves can be handled. These fall into two broad categories: filtering methods, in which knowledge about the phase of the unknown coefficients is used to filter out unwanted terms (used for a simpler quasi–one-dimensional scattering problem in [22]), and explicit methods, where the amplitude of the Rayleigh–Bloch waves that are excited is introduced as an extra unknown and an extra equation is therefore required. We will consider the latter type first.

3.1. Explicit methods. We define a new set of unknowns, $\hat{C}^p_m$, as follows:

$$A^p_m = \hat{C}^p_m + {\rm e}^{{\rm i}\lambda p} B_m + \alpha\, {\rm e}^{{\rm i}\widetilde{\beta} p} \widetilde{B}_m. \tag{3.1}$$

Here B_m is the solution to the infinite array problem (2.8), $\widetilde{B}_m$ is a solution to the homogeneous system (2.17) with $\beta \equiv \widetilde{\beta}$, normalized so that

$$\sum_{m=-\infty}^{\infty} |Z_m \widetilde{B}_m|^2 = 1, \tag{3.2}$$

and α is an unknown constant representing the (complex) amplitude of the Rayleigh–Bloch mode. We expect that as $p \to \infty$ (i.e., as we move away from the edge) the coefficients A_m^p will tend to the values appropriate to a fully infinite array, plus possibly the effect of any Rayleigh–Bloch waves, and hence that $\hat{C}_m^p \to 0$ as $p \to \infty$ provided that α is chosen appropriately.

If we substitute from (3.1) into (2.5) and use (2.8) and (2.17), we get a system of equations for the coefficients $\hat{C}_m^p$ which is the same as (2.5) except with a different right-hand side:

$$\begin{aligned} &\hat{C}_m^p + \sum_{n=-\infty}^{\infty} Z_n \sum_{\substack{j=0 \\ \neq p}}^{\infty} \hat{C}_n^j X_{n-m}^{jp} H_{n-m}(k|j-p|) \\ &\qquad = \sum_{n=-\infty}^{\infty} Z_n \left(B_n S_{n-m}^p(\lambda) + \alpha \widetilde{B}_n S_{n-m}^p(\widetilde{\beta}) \right), \qquad p = 0, 1, 2, \ldots, m \in \mathbb{Z}, \end{aligned} \tag{3.3}$$

where

$$S_n^p(\beta) = \sum_{j=p+1}^{\infty} \mathrm{e}^{\mathrm{i}\beta(p-j)} H_n(kj). \tag{3.4}$$

The slowly convergent series (3.4) does not contain unknown coefficients and thus can be treated analytically and computed efficiently; see [23]. It can be shown that (see [8, Appendix D] for the method though the final result in that paper is in error), provided that θ is not an integer multiple of 2π,

$$\sum_{j=p+1}^{\infty} \frac{\mathrm{e}^{\mathrm{i}j\theta}}{j^{1/2}} \sim \frac{-p^{-1/2}\mathrm{e}^{\mathrm{i}\theta p}}{1 - \mathrm{e}^{-\mathrm{i}\theta}} \quad \text{as} \quad p \to \infty. \tag{3.5}$$

Hence

$$S_n^p(\beta) \sim -\sqrt{\frac{2}{\pi k p}} \frac{(-\mathrm{i})^n \, \mathrm{e}^{-\frac{1}{4}\mathrm{i}\pi} \, \mathrm{e}^{\mathrm{i}kp}}{\left(1 - \mathrm{e}^{-\mathrm{i}(k-\beta)}\right)}. \tag{3.6}$$

It follows that the right-hand side of (3.3) decays as $p \to \infty$; therefore the behavior of the coefficients $\hat{C}_n^p$ in this limit must be such that the sum on the left-hand side converges. In fact, it turns out that (see section 4), as $p \to \infty$,

$$\hat{C}_n^p \sim C_n p^{-3/2} \mathrm{e}^{\mathrm{i}kp}. \tag{3.7}$$

The simplest way to determine α is to set $\hat{C}_N^P = 0$, treat α as an unknown, and solve (3.3) for $\hat{C}_n^0, \ldots, \hat{C}_n^P$, $|n| \le N$, by truncation, a procedure which was used for a related problem in [24]. This works, in that as P gets large, the value obtained for α converges, but the convergence is slow, and very large values of P are therefore required in order to obtain accurate results. We will refer to this approach to the

determination of α as the direct method. We can also make use of a subsequent result from the analysis of the far field as a means of obtaining an additional equation. Thus, the asymptotic behavior of the coefficients $\hat{C}_n^p$ given by (3.7) means that we must take $g(0) = 0$ in (4.7) (see section 4.2), and thus

$$\sum_{n=-\infty}^{\infty} (-\mathrm{i})^n Z_n \left(\frac{B_n}{1 - \mathrm{e}^{\mathrm{i}(\lambda - k)}} + \frac{\alpha \widetilde{B}_n}{1 - \mathrm{e}^{\mathrm{i}(\tilde{\beta} - k)}} + \sum_{j=0}^{\infty} \hat{C}_n^j \mathrm{e}^{-\mathrm{i}kj} \right) = 0. \tag{3.8}$$

Note that this is not an explicit formula for α since the coefficients $\hat{C}_n^j$ are the solutions to (3.3), which contains α on the right-hand side. For the antisymmetric part of the field this identity is trivially satisfied; thus it is of use only for the symmetric part. The use of (3.3) combined with (3.8) will be referred to as the far field method for the determination of α. Unfortunately, the sum involving $\hat{C}_n^j$ converges too slowly for this approach to work well. In cases where there is no Rayleigh–Bloch wave (i.e., in the Neumann problem with $k > k_{\max}^{\mathrm{s}}$ or in the Dirichlet case), $\alpha = 0$, and equation (3.8) is an identity that can be used as a check on the results. It can also be used as a numerical check in cases where Rayleigh–Bloch waves do exist if α is calculated by some other means.

In order to make the best use of (3.8) we use a simple acceleration procedure. This procedure has been used wherever possible in what follows and we will refer to it as asymptotic acceleration. Thus, we substitute the asymptotic form for the coefficients $\hat{C}_n^j$ for all values of j greater than the truncation parameter $j = J$, giving

$$\sum_{j=0}^{\infty} \hat{C}_n^j \mathrm{e}^{-\mathrm{i}kj} = \sum_{j=0}^{J} \hat{C}_n^j \mathrm{e}^{-\mathrm{i}kj} + C_n \sum_{j=J+1}^{\infty} j^{-3/2}, \tag{3.9}$$

the coefficient C_n being determined from the computed value of $\hat{C}_n^J \approx C_n J^{-3/2} \exp(\mathrm{i}kJ)$ and the final sum being a generalized zeta function which can easily be computed from standard packages.

3.2. Filtering. We have found that filtering methods yield the best results in terms of accuracy and efficiency, and these are described next. This time we define new unknowns C_m^p via

$$A_m^p = C_m^p + \mathrm{e}^{\mathrm{i}\lambda p} B_m \tag{3.10}$$

(so that $C_m^p = \hat{C}_m^p + \alpha\, \mathrm{e}^{\mathrm{i}\tilde{\beta}p} \widetilde{B}_m$). Instead of (3.3) we now have

$$C_m^p + \sum_{n=-\infty}^{\infty} Z_n \sum_{\substack{j=0 \\ \neq p}}^{\infty} C_n^j X_{n-m}^{jp}\, H_{n-m}(k|j-p|) = \Gamma_m^p, \tag{3.11}$$

$p = 0, 1, 2, \ldots, m \in \mathbb{Z}$, where for future convenience we have defined

$$\Gamma_m^p = \sum_{n=-\infty}^{\infty} Z_n B_n S_{n-m}^p(\lambda). \tag{3.12}$$

If Rayleigh–Bloch modes are excited, then the coefficients C_m^p will not decay to zero as $p \to \infty$. Instead, we expect that $C_m^p \sim \mathrm{e}^{\mathrm{i}\tilde{\beta}} C_m^{p-1}$ in this limit. We thus introduce

$$D_m^p = \begin{cases} C_m^p - \mathrm{e}^{\mathrm{i}\tilde{\beta}} C_m^{p-1}, & p = 1, 2, \ldots, \\ C_m^0, & p = 0, \end{cases} \tag{3.13}$$

so that D_m^p decays to zero as $p \to \infty$. This recurrence relation can be solved for C_m^p to give

$$C_m^p = \sum_{j=0}^{p} \mathrm{e}^{\mathrm{i}\tilde{\beta}(p-j)} D_m^j, \qquad p = 0, 1, 2, \ldots. \tag{3.14}$$

A system of equations for D_m^p can then be derived in more than one way. If (3.14) is substituted into (3.11), we get

$$\sum_{j=0}^{p} \mathrm{e}^{\mathrm{i}\tilde{\beta}(p-j)} D_m^j + \sum_{n=-\infty}^{\infty} Z_n \sum_{j=0}^{\infty} D_n^j \sum_{\substack{l=j \\ \neq p}}^{\infty} \mathrm{e}^{\mathrm{i}\tilde{\beta}(l-j)} X_{n-m}^{lp} H_{n-m}(k|l-p|) = \Gamma_m^p, \tag{3.15}$$

$p = 0, 1, 2, \ldots, m \in \mathbb{Z}$. Alternatively we can combine equations in (3.11) in the obvious way so that

$$D_m^p + \sum_{n=-\infty}^{\infty} Z_n \sum_{\substack{j=0 \\ \neq p}}^{\infty} D_n^j X_{n-m}^{jp} H_{n-m}(k|j-p|) = \Gamma_m^p - \mathrm{e}^{\mathrm{i}\tilde{\beta}} \Gamma_m^{p-1}, \tag{3.16}$$

$p = 1, 2, \ldots, m \in \mathbb{Z}$. The system (3.16) then needs to be supplemented by an equation for D_m^0 which can be obtained from (3.11) with $p = 0$ by substituting for C_n^j from (3.14). This yields

$$D_m^0 + \sum_{n=-\infty}^{\infty} (-1)^{n-m} Z_n \sum_{j=0}^{\infty} D_n^j \sum_{l=\max(j,1)}^{\infty} \mathrm{e}^{\mathrm{i}\tilde{\beta}(l-j)} H_{n-m}(kl) = \Gamma_m^0. \tag{3.17}$$

The sums over l in either of the formulations can be expressed in terms of the sums S_n^p defined in (3.4) and can be computed efficiently and accurately using results from [23]. The two formulations are equivalent, but we have found that the second (i.e., using (3.16) and (3.17)) is easier to implement. In either case, the finite sum in (3.14) has been interchanged with the (infinite) spatial sum in (3.11). This crucial step has the effect of continuing the filtered term (in this case the Rayleigh–Bloch mode) to infinity, so that it is unaffected by spatial truncation. This is the essence and great advantage of infinite array subtraction and filtering methods: only the part of the solution which decays as one moves along the array is subject to errors caused by spatial truncation.

Once the coefficients D_m^p have been computed via truncation, the coefficients C_m^p can be reconstructed from (3.14). If we truncate at $p = P$, a value for α can then be deduced from

$$C_m^P \mathrm{e}^{-\mathrm{i}\tilde{\beta}P} = \sum_{p=0}^{P} \mathrm{e}^{-\mathrm{i}\tilde{\beta}p} D_m^p \to \alpha \widetilde{B}_m \quad \text{as} \quad P \to \infty. \tag{3.18}$$

The coefficients D_m^p have the asymptotic behavior $D_m^p \sim D_m p^{-3/2} \exp(\mathrm{i}kp)$ exactly as for $\hat{C}_m^p$ because they both model the behavior of the scattered field once the Rayleigh–Bloch wave has been removed. Thus asymptotic acceleration can be used in (3.18).

It is possible to use the known asymptotic behavior of D_m^p to create a set of coefficients which decay like $p^{-5/2}$ by filtering again. In other words we define a new

TABLE 3.1

Convergence of different methods for the determination of $|\alpha|$ for the symmetric Rayleigh–Bloch mode that is excited when $a = 0.25$, $\psi_0 = \pi/10$, and $k = 2$. The numbers in parentheses are the results when asymptotic acceleration is not used (note that this is not available when using the direct method). Eleven modes have been used in the order summations.

Spatial truncation	Direct	Far field		Single filtering		Double filtering	
50	(0.0938)	(0.1148)	0.1022	(0.0940)	0.1007	(0.1015)	0.1014
100	(0.1015)	(0.1213)	0.0991	(0.1017)	0.1017	(0.0999)	0.1014
150	(0.1035)	(0.1022)	0.1016	(0.1035)	0.1016	(0.1014)	0.1015
200	(0.1020)	(0.0869)	0.1023	(0.1019)	0.1015	(0.1018)	0.1015
250	(0.1007)	(0.0972)	0.1016	(0.1007)	0.1014	(0.1016)	0.1015
300	(0.1010)	(0.1117)	0.1011	(0.1010)	0.1015	(0.1014)	0.1015

set of coefficients via

$$E_m^p = \begin{cases} D_m^p - \mathrm{e}^{\mathrm{i}k} D_m^{p-1}, & p = 1, 2, \ldots, \\ D_m^0, & p = 0. \end{cases} \tag{3.19}$$

Details of the resulting equations can be found in the appendix. In most situations there is very limited gain from this second filtering. However, there are certain situations (which we will mention below) where it is absolutely essential.

3.3. Numerical results. We have described four methods which can be used to determine α: the direct method, the far field method, single filtering, and double filtering. The final three can all be improved via asymptotic acceleration. Table 3.1 shows the relative performance of these different approaches for a typical, rather than an extreme, case. We have taken $a = 0.25$, $\psi_0 = \pi/10$, and $k = 2$. For these parameters, there is a symmetric Rayleigh–Bloch mode with $\widetilde{\beta} \approx 2.0268$. The table, which lists values of $|\alpha|$, clearly demonstrates the superiority of the filtering methods, and also the increased convergence that results from using asymptotic acceleration. An important caveat to note is that double filtering does not work well when k and $\widetilde{\beta}$ are too close together. This happens for symmetric Rayleigh–Bloch waves when k is small. However, this is mitigated against by the fact that in long waves smaller truncations in the order summations are necessary for a given accuracy, and hence large spatial truncations can easily be used.

When antisymmetric surface waves are excited, computing α accurately is more of a challenge. One factor is that such modes exist only for large values of a when the cylinders are close together, and this entails the use of many more terms in the order summations so as to accurately model the interactions. The second is that the problematic case $k \approx \widetilde{\beta}$ occurs not for very long waves but when k is near $k^{\mathrm{a}}_{\mathrm{min}}$. Table 3.2 shows the relative performance of the different approaches for computing α for the case $a = 0.49$, $\psi_0 = \pi/10$, and $k = 2.5$. For these parameters, there is an antisymmetric Rayleigh–Bloch mode with $\widetilde{\beta} \approx 2.5096$. (There is also a symmetric mode excited with $\widetilde{\beta} \approx 2.5644$, and thus the problem must be decomposed into its symmetric and antisymmetric parts before the Rayleigh–Bloch amplitudes are calculated.) Note that the far field method cannot be used in the antisymmetric case. Again, the filtering methods are seen to converge fastest as the spatial truncation is increased. It is also evident that the convergence is not as good as in the symmetric case presented in Table 3.1.

TABLE 3.2

Convergence of different methods for the determination of $|\alpha|$ for the antisymmetric Rayleigh–Bloch mode that is excited when $a = 0.49$, $\psi_0 = \pi/10$, and $k = 2.5$. The numbers in parentheses are the results when asymptotic acceleration is not used (note that this is not available when using the direct method). Twenty-one modes have been used in the order summations.

Spatial truncation	Direct	Single filtering		Double filtering	
50	(0.2212)	(0.2212)	0.2400	(0.3236)	0.2597
100	(0.2197)	(0.2198)	0.2380	(0.2571)	0.2445
150	(0.2258)	(0.2259)	0.2384	(0.2418)	0.2413
200	(0.2325)	(0.2325)	0.2392	(0.2375)	0.2404
250	(0.2379)	(0.2380)	0.2400	(0.2369)	0.2402
300	(0.2418)	(0.2418)	0.2405	(0.2377)	0.2402

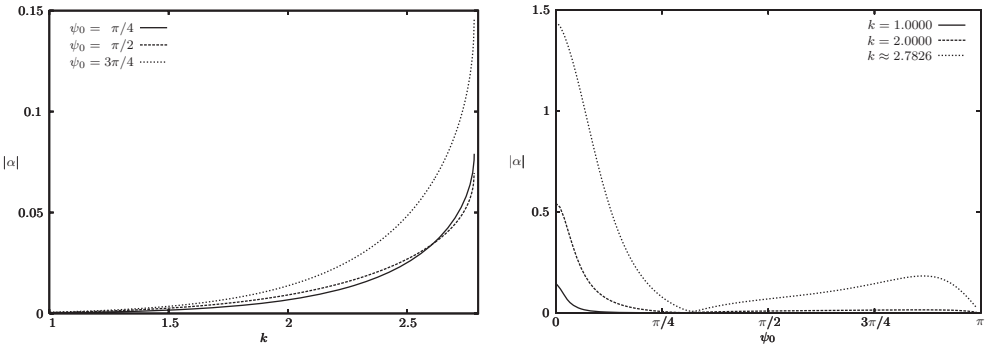


FIG. 3.1. *Left: variation with k of $|\alpha|$ for the excited symmetric surface wave, for three different angles of incidence, when $a = 0.25$. Right: variation with ψ_0 of $|\alpha|$ for the excited symmetric surface wave, for three different values of k, when $a = 0.25$.*

Figure 3.1 shows the variation in the amplitude of the excited symmetric surface wave with k for three different angles of incidence and with ψ_0, for three different wavenumbers, when $a = 0.25$. For this value of a, symmetric Rayleigh–Bloch waves are excited for all k in the range $0 < k < k^{\mathrm{s}}_{\max} \approx 2.783$ (and antisymmetric Rayleigh–Bloch waves are never excited). For $0 < k < 1$ (not shown in the figure), $|\alpha|$ is essentially zero. This corresponds to wavelength-to-spacing ratios greater than 2π. As k increases so the amplitude increases, reaching a maximum at $k^{\mathrm{s}}_{\max}$ (at which point $\widetilde{\beta} = \pi$). The variation in $|\alpha|$ with ψ_0 is not monotonic, and this is illustrated clearly in Figure 3.1. For a given k, the amplitude is greatest at head-on incidence (note that this case requires special treatment as described in section 5 below). It then reduces to approximately zero at an angle somewhere near $\pi/3$ (independently of the value of k) before increasing and then getting smaller again as the incident wave grazes the array.

Figure 3.2 shows the variation in the amplitude of the excited antisymmetric surface wave with k for three different angles of incidence and with ψ_0, for three different wavenumbers, when $a = 0.49$. For this value of a, antisymmetric Rayleigh–Bloch waves are excited for all k in the range $1.796 \lesssim k \lesssim 2.969$ (and symmetric Rayleigh–Bloch waves are also excited). For k just above 1.796 we have problems computing α accurately caused by the closeness of k and $\widetilde{\beta}$. The qualitative behavior of the amplitude as a function of k is very similar to that for the symmetric mode shown in Figure 3.1. At head-on incidence the problem is entirely symmetric about

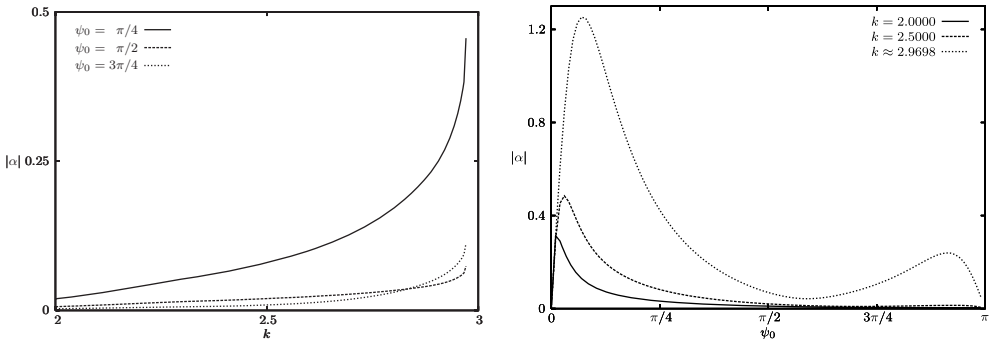


FIG. 3.2. *Left: variation with k of $|\alpha|$ for the excited antisymmetric surface wave, for three different angles of incidence, when $a = 0.49$. Right: variation with ψ_0 of $|\alpha|$ for the excited antisymmetric surface wave, for three different values of k, when $a = 0.49$.*

the line of the array, and thus the amplitude of the antisymmetric mode tends to zero as the incidence angle tends to zero. The amplitude rises sharply as ψ_0 increases from zero, and the maximum amplitude occurs for quite small angles.

Figures 3.1–3.2 clearly show that the amplitude of the surface waves that are excited are greatest when the frequency parameter k is close to its maximum possible value for the modes to exist. When k is just less than $k_{\max}$, $\widetilde{\beta}$ is just less than π, and as $\widetilde{\beta} \to \pi$ from below, the Rayleigh–Bloch mode approaches a standing wave and its group velocity, $c_{\rm g}$, tends to zero. The energy in the Rayleigh–Bloch wave is proportional to $|\alpha|^2$, and thus the rate of energy transport is proportional to $|\alpha|^2 c_{\rm g}$, which tends to zero as $k \to k_{\max}$. Hence the large amplitudes correspond to situations where the energy is transported slowly away from the array edge.

4. The far field. From (2.4), (2.13), and (3.1), the scattered field can be written as

$$\phi_{\rm sc} = \sum_{j=0}^{\infty}\sum_{n=-\infty}^{\infty}\left(\mathrm{e}^{\mathrm{i}\lambda j}B_n + \alpha\,\mathrm{e}^{\mathrm{i}\widetilde{\beta}j}\widetilde{B}_n + \hat{C}_n^j\right)Z_n \times \frac{(-\mathrm{i})^{n+1}}{\pi}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{-k\gamma(t)|y|}}{\gamma(t)}\,\mathrm{e}^{\mathrm{i}k(x-j)t}\,(t-\gamma(t))^{n\,\mathrm{sgn}(y)}\,\mathrm{d}t. \tag{4.1}$$

The spatial sums involving B_n and $\widetilde{B}_n$ can be evaluated using the result

$$\sum_{j=0}^{\infty}\int_{-\infty}^{\infty} f(u)\,\mathrm{e}^{-\mathrm{i}ju}\,\mathrm{d}u = \int_{-\infty}^{\infty}\frac{f(u)}{1-\mathrm{e}^{-\mathrm{i}u}}\,\mathrm{d}u \tag{4.2}$$

(see [8], [25]). Thus

$$\phi_{\rm sc} = \sum_{n=-\infty}^{\infty} Z_n\frac{(-\mathrm{i})^{n+1}}{\pi}\int_{-\infty}^{\infty}\left[b_n(t)+\tilde{b}_n(t)+\hat{c}_n(t)\right](t-\gamma(t))^{n\,\mathrm{sgn}(y)}\,\mathrm{e}^{-k\gamma(t)|y|+\mathrm{i}kxt}\frac{\mathrm{d}t}{\gamma(t)}, \tag{4.3}$$

in which

$$b_n(t) = \frac{B_n}{1 - \mathrm{e}^{\mathrm{i}(\lambda - kt)}}, \qquad \tilde{b}_n(t) = \frac{\alpha \tilde{B}_n}{1 - \mathrm{e}^{\mathrm{i}(\tilde{\beta} - kt)}}, \tag{4.4}$$

$$\hat{c}_n(t) = \sum_{j=0}^{\infty} \hat{C}_n^j \mathrm{e}^{-\mathrm{i}kjt}, \quad \mathrm{Im}(t) \leq 0. \tag{4.5}$$

The path of integration passes below all of the singularities, apart from the branch point at $t = -1$.

In general the far field asymptotics of (4.3) can be obtained via a straightforward application of the method of steepest descents, provided that the point of observation is not close to the array. The function $\hat{c}_n(t)$ is analytic in $\mathrm{Im}(t) < 0$ by definition, and from (3.7) and [26, section 3.4] the only singularities on $\mathbb{R}$ are branch points, on approach to which $\hat{c}_n(t)$ remains bounded; these do not contribute to the leading order far field behavior of ϕ_{sc}. For $\mathrm{Im}(t) > 0$, $\hat{c}_n(t)$ represents the meromorphic continuation of (4.5) into some cut upper half-plane. While $\hat{c}_n(t)$ may possess singularities in this region, these will yield an exponentially small contribution to the far field should they be encountered in the process of making the steepest descents deformation. Also, the poles of $\tilde{b}_n(t)$ lie outside the interval $[-1, 1]$, so that their contribution is also exponentially small. Thus, only the term involving $b_n(t)$ requires special treatment. First, assume that the saddle point $t = \cos\theta$ does not coincide with any of the poles of $b_n(t)$. Ignoring evanescent contributions, we find that as $kr \to \infty$ with $\theta \in (0, 2\pi)$ (provided $\psi_m \neq 0$ for any m),

$$\phi_{\mathrm{sc}} \sim \widetilde{H}(kr) g(\theta) + \sum_{\substack{m \in \mathcal{M} \\ \psi_m > \theta}} \mathcal{F}_m^+ \mathrm{e}^{\mathrm{i}kr\cos(\theta - \psi_m)} + \sum_{\substack{m \in \mathcal{M} \\ 2\pi - \psi_m < \theta}} \mathcal{F}_m^- \mathrm{e}^{\mathrm{i}kr\cos(\theta + \psi_m)}. \tag{4.6}$$

Here, $\mathcal{F}_m^{\pm}$ is defined in (2.15), $\widetilde{H}(kr) = \sqrt{\frac{2}{\pi kr}}\, \exp(\mathrm{i}(kr - \frac{1}{4}\pi))$, and

$$g(\theta) = \sum_{n=-\infty}^{\infty} (-\mathrm{i})^n \mathrm{e}^{\mathrm{i}n\theta} Z_n\big(b_n(\cos\theta) + \tilde{b}_n(\cos\theta) + \hat{c}_n(\cos\theta)\big). \tag{4.7}$$

The diffracted field takes the form of a circular wave of directivity $g(\theta)$ plus a sum of plane waves which propagate in the same directions as in the infinite grating case. However, unlike in the grating problem, the plane waves do not exist everywhere, and the wave making an angle ψ_m (resp., $-\psi_m$) with the x-axis is found only in the sector $0 < \theta < \psi_m$ (resp., $2\pi > \theta > 2\pi - \psi_m$). Crucially, the coefficients $\hat{C}_n^m$ affect only the circular wave. The plane wave field is determined entirely from the solution to the infinite grating problem; in fact, where the plane waves exist, their amplitude is precisely as in the infinite grating problem. Thus it is only the circular wave which causes any computational difficulties.

4.1. Uniform asymptotics. The approximation (4.6) is nonuniform in the sense that $b_n(\cos\theta)$ is singular at the shadow boundaries where $\theta = \psi_p$ or $\theta = 2\pi - \psi_p$ (a case which corresponds to a pole of $b_n(t)$ coinciding with the saddle point). This limitation can be overcome by adding correction terms, each of which includes an error function which rapidly but continuously activates and deactivates the appropriate plane wave as the shadow boundary is crossed. These correction terms have appeared in the literature in numerous forms and with various regions of validity. The appropriate form for use here is that given by Thompson [27], since this accounts for limits

in which a shadow boundary in $y > 0$ approaches its counterpart in $y < 0$. Thus, the uniform approximation to $\phi_{\rm sc}$ is

$$\phi_{\rm sc} \sim \widetilde{H}(kr)g(\theta) + \tfrac{1}{2}\mathrm{e}^{\mathrm{i}kr} \sum_{m\in\mathcal{M}} \left[\mathcal{F}_m^+ \left(w\left(\zeta_m^- \mathrm{e}^{\mathrm{i}\pi/4}\right) - \frac{\mathrm{e}^{\mathrm{i}\pi/4}}{\zeta_m^- \sqrt{\pi}} \right) + \mathcal{F}_m^- \left(w\left(\zeta_m^+ \mathrm{e}^{\mathrm{i}\pi/4}\right) - \frac{\mathrm{e}^{\mathrm{i}\pi/4}}{\zeta_m^+ \sqrt{\pi}} \right) \right], \tag{4.8}$$

where $\zeta_m^\pm = \sqrt{2kr}\sin\frac{1}{2}(\theta \pm \psi_m)$ and $w(z) = \exp(-z^2)\operatorname{erfc}(-\mathrm{i}z)$ is the scaled complex error function. It is not difficult to show that (4.8) is continuous at all of the shadow boundaries, as the singularities in $\widetilde{H}(kr)g(\theta)$ are cancelled by those in the series (note that $w(0) = 1$). Now, if $\zeta_m^\pm > 0$, we can use the result [28, equation (7.1.23)]

$$w(z) \sim \mathrm{i}/(z\sqrt{\pi}) + O(z^{-3/2}), \quad z \to \infty, \quad -\pi/4 < \arg(z) < 5\pi/4, \tag{4.9}$$

to show that the correction term vanishes to leading order as $kr \to \infty$. On the other hand, for $\zeta_p < 0$, we must first apply the identity

$$w(z) + w(-z) = 2\mathrm{e}^{-z^2} \tag{4.10}$$

and then use (4.9) to deduce that

$$\frac{\mathrm{e}^{\mathrm{i}kr}}{2}\left(w\left(\zeta_p^\pm \mathrm{e}^{\mathrm{i}\pi/4}\right) - \frac{\mathrm{e}^{\mathrm{i}\pi/4}}{\zeta_p^\pm \sqrt{\pi}} \right) \sim \mathrm{e}^{\mathrm{i}kr\cos(\theta\pm\psi_p)} + O((kr)^{-3/2}) \tag{4.11}$$

as $kr \to \infty$. Thus each error function term in (4.8) (which is an exact solution to the Helmholtz equation) includes a plane wave in the appropriate region.

Note that the limit $y \to 0$ of (4.3) can be taken directly, provided that $x < 0$, since then convergence can be maintained by deforming the path of integration into the lower half-plane. The value taken for $\operatorname{sgn}(0)$ is immaterial. To see this, subtract (4.3) with $y = 0^+$ from the same equation with $y = 0^-$. The resulting integrand has no branch point at $t = -1$, in view of the identity $(t - \gamma(t))^n = (t + \gamma(t))^{-n}$, and therefore evaluates to zero. This shows that (4.6) and (4.8) are valid at $\theta = \pi$. For $x > 0$, the required upwards deformation cannot be performed since the sums of residues from $b(t)$ and $\tilde{b}(t)$ and of branch point contributions from $\hat{c}_n(t)$ all diverge when $y = 0$. Consequently, (4.8) represents only a part of the far field at $\theta = 0$ and $\theta = 2\pi$, and not necessarily the most significant.

Figure 4.1 shows a contour plot of the real part of the scattered field, with $a = 0.25$, $\psi_0 = 0.25\pi$, and $k = 5.0$, using Dirichlet boundary conditions. Table 4.1 contains the amplitudes of the propagating modes for this case, and the real part of the leading order contribution to the far field given by (4.8) is plotted in Figure 4.2, with $r = 5$. The imaginary part exhibits qualitatively similar behavior and is not shown. The black disks indicate the locations of the shadow boundaries. The data from which the dashed line is plotted includes contributions from plane waves and is therefore continuous. The solid line represents the same uniform approximation, but with the plane waves removed via (4.10). The two lines coincide in the region $0.69\pi < \theta < 1.25\pi$ where no plane waves exist; here the circular wave is clearly visible in Figure 4.1. The sizes of the discontinuities in the solid line are consistent with the mode amplitudes in Table 4.1; all of the associated shadow boundaries are clearly evident in Figure 4.1.

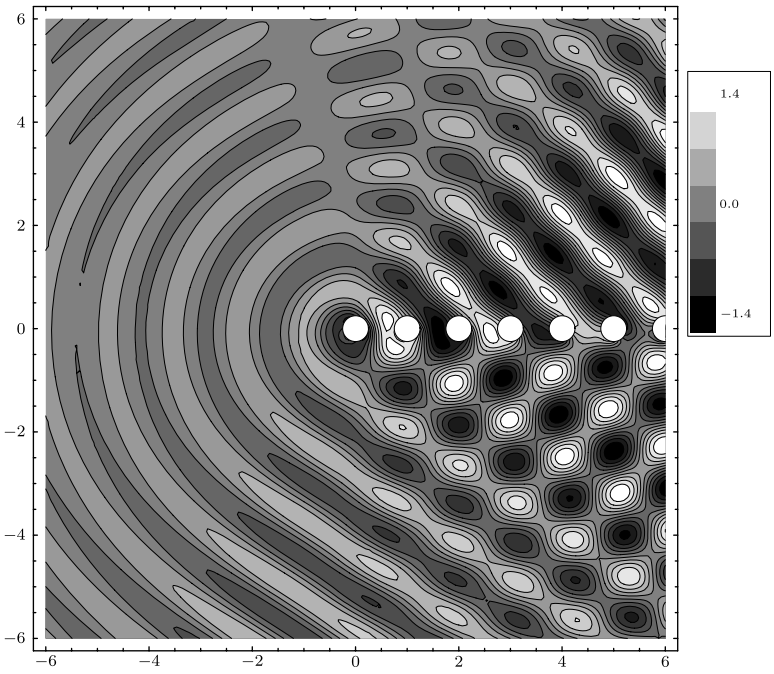


FIG. 4.1. *Contour plot of* $\mathrm{Re}[\phi_{\mathrm{sc}}]$*, with* $a = 0.25$*,* $\psi_0 = 0.25\pi$*, and* $k = 5.0$*, using Dirichlet boundary conditions.*

TABLE 4.1
Propagating mode amplitudes and directions for the contour plot shown in Figure 4.1.

j	ψ_j	$\lvert\mathcal{F}_j^+\rvert$	$\lvert\mathcal{F}_j^-\rvert$
-1	0.69π	0.177	0.578
0	0.25π	0.170	0.734

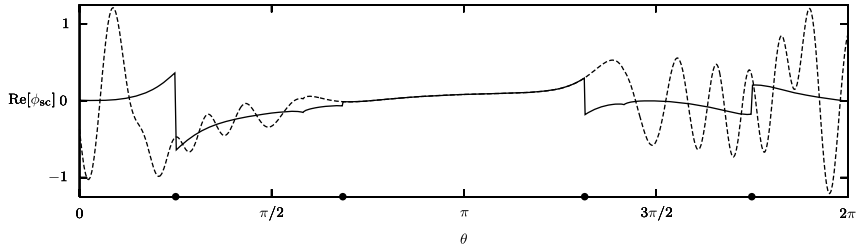


FIG. 4.2. *Far field plots for the parameters used in Figure* 4.1*. Correction terms are included for all shadow boundaries; the dashed line includes plane wave contributions, whereas the solid line does not.*

4.2. Behavior of $\hat{C}_n^p$ as $p \to \infty$. Finally, we give some justification for (3.7), though this will not amount to a rigorous proof. The total field in the region $r_p < 1/2$ (i.e., local to scatterer p) can be written in the form [13]

$$\phi = \sum_{n=-\infty}^{\infty} A_n^p \left[Z_n H_n(kr_p) - J_n(kr_p) \right] \mathrm{e}^{\mathrm{i}n\theta_p}. \tag{4.12}$$

If we write similar expressions for $\phi^{\rm inf}$ and $\phi^{\rm rb}$ (see section 2) and introduce $\hat{\phi} = \phi - (\phi^{\rm inf} + \phi^{\rm rb})$, then from (3.1) we have

$$\hat{\phi} = \sum_{n=-\infty}^{\infty} \hat{C}_n^p \left[Z_n \, H_n(kr_p) - J_n(kr_p) \right] \mathrm{e}^{\mathrm{i}n\theta_p}. \tag{4.13}$$

For sufficiently large p, $\hat{\phi}$ represents the field due to end effects other than the Rayleigh–Bloch wave, and its asymptotic behavior is evidently determined by that of $\hat{C}_n^p$. It is not difficult to show that $\hat{\phi}$ can also be represented by (4.3), but with the path of integration now passing above the poles. Nevertheless, we cannot set $y = 0$, since then the sum of contributions from the branch points of $\hat{c}_n(t)$ would diverge. Instead, we can impose the restriction $y < a$, so that letting $x \to \infty$ causes θ to approach zero. Then, we deform the contour into the upper half-plane and deduce the leading order behavior of $\hat{\phi}$ as $x \to \infty$. The phase dependence and rate of decay of $\hat{C}_n^p$ as $p \to \infty$ must be such that the result of this calculation is consistent with (4.13). Now we introduce the ansatz

$$\hat{C}_n^p \sim C_n \mathrm{e}^{\mathrm{i}p\xi} p^{-u}, \tag{4.14}$$

in which C_n is a constant, $u > 0$, and $\xi > 0$, and substitute this into (4.5). Clearly, the critical points will be those at which the phase disappears; therefore we define

$$t_m = (\xi + 2m\pi)/k. \tag{4.15}$$

The behavior of $\hat{c}_n(t)$ in the vicinity of $t = t_m$ now follows from [26, section 3.4]. Thus, as $(t_m - t) \to 0^+$, we have

$$\hat{c}_n(t) = C_n \Gamma(1-u) \mathrm{e}^{\mathrm{i}\pi(1-u)/2} \xi^{u-1} (t_m - t)^{u-1} + f_n(t), \quad u \neq 1, \tag{4.16}$$

where $f_n(t)$ is regular at $t = t_m$ and the terms with exponent $u - 1$ are positive real. If $u = 1$, then the integrand in (4.3) possesses logarithmic singularities whose contribution cannot be consistent with (4.13). There are now two cases to consider. If $\xi \neq k$, then the leading order behavior must be due to a singularity of the function $\hat{c}_n(t)$, since the contribution from the branch point at $t = 1$ (which in general is $O((kr)^{-1/2})$) has the wrong phase, according to (4.13). Consequently, we must have $u < 1/2$ in this case. On the other hand, if $\xi = k$, then the singularity at $t = 1$ yields a contribution whose rate of decay is slower than that predicted by (4.13), unless a sufficient number of terms vanish as $\theta \to 0$ (or 2π) so as to achieve consistency. Note that we must have $u > 1/2$, or else the sum from $j = 0$ to $j = p - 1$ on the left-hand side of (3.3) would diverge as $p \to \infty$ due to phase cancellation. If $u = 3/2$, only the leading order term must disappear, that is, $g(0) = g(2\pi) = 0$, and now both (4.3) and (4.13) predict that the leading order far field behavior of $\hat{\phi}$ on the array is $O((kr)^{-3/2})$. This is borne out by numerical computations and is also the behavior proved in [8] using the discrete Wiener–Hopf technique for a semi-infinite array of point scatterers (which is equivalent to taking only the monopole terms in the Dirichlet problem here).

5. Resonance. Resonance occurs when one of the modes in (4.6) propagates in a direction parallel to the array. This requires that either $\psi_m = 0$ or $\psi_m = \pi$ for some m; see (2.10). In general, resonances can occur only if $k > \pi$, thereby precluding the possibility of simultaneous occurrence with Rayleigh–Bloch waves. The special case of head-on incidence ($\psi_0 = 0$) is resonant for *all* k; symmetric Rayleigh–Bloch waves

may be excited if $k < \pi$. We will take (3.11) as the starting point for solving resonant problems. Now, in order to use the infinite array subtraction technique we must first compute the coefficients B_n in the resonant case; this is complicated by the fact that the Schlömilch series in (2.9) are now divergent. However, it can be achieved using the method developed in [17]. The form for the mode amplitude $\mathcal{F}_m^{\pm}$ in the limit $\sin\psi_m \to 0$ is also given in [17]. It was noted in [6, 29] that for point scatterers all the nonresonant scattered modes disappear at resonance. To see this, one need only note that the solution for point scatterers is retrieved by truncating all order summations at zero. Then, at a resonance, the divergence of the Schlömilch series in (2.8) requires that $B_0 = 0$, which in turn implies that $\mathcal{F}_p^{\pm} = 0$, unless mode p is resonant in which case we have $\mathcal{F}_p^{\pm} = -1$; see [17]. For finite size scatterers, however, singular behavior in the Schlömilch series requires that the coefficients B_n satisfy a modified system of equations which permits nonzero values (except in the case of head-on incidence), and thus all the modes are present in the scattered field.

5.1. Outward resonance. The case in which $\psi_m = \pi$ is known as outward resonance, since now mode m has the form $\mathcal{F}_m^{\pm}\mathrm{e}^{-\mathrm{i}kx}$. This can occur only if $k > \pi$; therefore no Rayleigh–Bloch waves are excited. Outward resonant modes exist in all space; however, in general they have different amplitudes in regions $y < 0$ and $y > 0$. Once the resonant solution to the infinite array problem has been obtained, no further special treatment is required, and the coefficients C_m^p can be computed from (3.11). Note that, in the case of point scatterers, in which $B_0 = 0$, the right-hand side of (3.11) vanishes, so that $C_0^p = 0$ for all p, and therefore the circular wave term in (4.6) also disappears. For finite sized scatterers, this is not the case, since $B_n \neq 0$ in general.

Figure 5.1 shows a contour plot of the real part of the scattered field, with $a = 0.25$, $\psi_0 = 0.6\pi$, and the wavenumber chosen so that mode -1 is outward resonant ($k \approx 9.1$). Including the resonant mode, there are three propagating plane waves; the amplitudes above and below the array are shown in Table 5.1. Since the amplitude of mode 1 is relatively small, its shadow boundaries are not visible in Figure 5.1; however, those at $\theta = 0.6\pi$, $\theta = \pi$, and $\theta = 1.4\pi$ are clearly evident. The presence of mode 0 when $\theta < 0.6\pi$ and $\theta > 1.4\pi$ accounts for the interference in this region. The real part of the leading order contribution to the far field given by (4.8) is plotted in Figure 5.2, with $r = 5$. As before, the black disks indicate the locations of the shadow boundaries, and the dashed line includes contributions from plane waves, whereas the solid line does not. Notice in particular the smooth transition that occurs in the amplitude of the resonant mode across $\theta = \pi$. This effect is due to the coincidence of two shadow boundaries directly opposite the array; thus, as the observer moves from the region where $\theta < \pi$ to that where $\theta > \pi$, the mode $\mathcal{F}^{+}\mathrm{e}^{-\mathrm{i}kx}$ is deactivated, and $\mathcal{F}^{-}\mathrm{e}^{-\mathrm{i}kx}$ is activated in its place. The sizes of the discontinuities that occur at the shadow boundaries are consistent with the mode amplitudes shown in Table 5.1. Note that the plane wave terms are, in general, of greater amplitude than the circular wave; this is why the latter is not particularly visible in Figure 5.1.

5.2. Inward resonance. The inward resonance case in which $\psi_m = 0$ is more interesting and presents more of a challenge. The extra difficulty in handling inward resonance was noted by Hills [29], who attempted to analyze this case for a semi-infinite array of isotropic point scatterers. As before, we require the coefficients B_n from the infinite array problem, and these can be obtained using the method in [17]. There is now an additional obstacle, caused by the divergence of the series in the

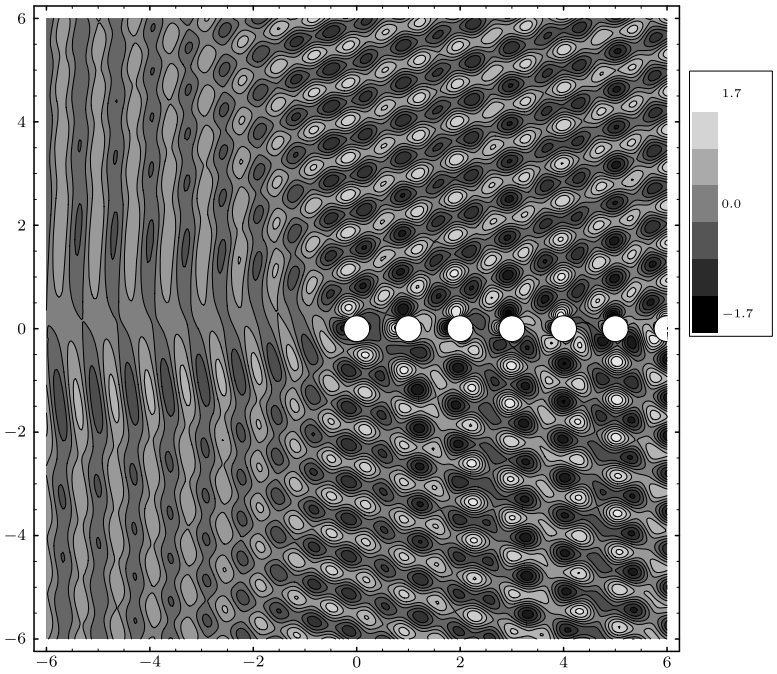


FIG. 5.1. *Contour plot of* $\mathrm{Re}[\phi_{\mathrm{sc}}]$*, with* $a = 0.25$*,* $\psi_0 = 0.6\pi$*, and the wavenumber chosen so that mode* -1 *is (outward) resonant (*$k \approx 9.1$*).*

TABLE 5.1
Propagating mode directions and amplitudes for the contour plot shown in Figure 5.1.

j	ψ_j	$\lvert\mathcal{F}_j^+\rvert$	$\lvert\mathcal{F}_j^-\rvert$
-1	π	1.46	0.680
0	0.6π	0.620	0.589
1	0.38π	0.180	0.336

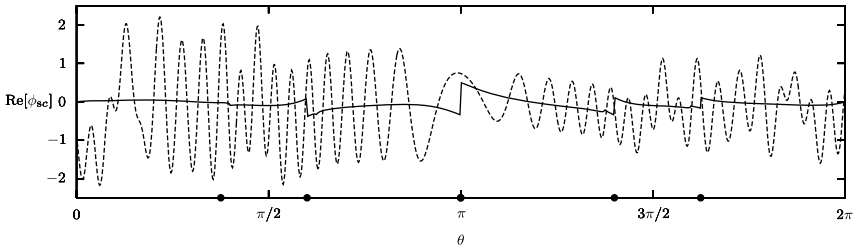


FIG. 5.2. *Far field plots for the parameters used in Figure* 5.1*, with* $r = 5$*. Correction terms are included for all shadow boundaries; the dashed line includes plane wave contributions, whereas the solid line does not.*

right-hand side of (3.11). In fact, it can be deduced from equations in [23] that

$$S_n^p(\lambda) = \hat{S}_n^p(\lambda) + 2(-\mathrm{i})^n \mathrm{e}^{\mathrm{i}pk}/(k\psi_m), \tag{5.1}$$

where $\hat{S}_n^p(\lambda)$ remains bounded as $\psi_m \to 0$, and we have used the fact that, since mode m is resonant, $\cos\psi_0 = (1 - 2m\pi)/k$. In order for the solution to remain bounded,

we must have

$$\frac{2}{k}\sum_{n=-\infty}^{\infty}(-\mathrm{i})^n Z_n B_n = a_1\psi_m + O(\psi_m^2), \tag{5.2}$$

and the value of the constant a_1 is obtained as a by-product of the procedure used in computing B_n; see [17]. Equation (3.11) now becomes

$$C_m^p + \sum_{n=-\infty}^{\infty} Z_n \sum_{\substack{j=0\\ \neq p}}^{\infty} C_n^j X_{n-m}^{jp}\, H_{n-m}(k|j-p|) = a_1 \mathrm{i}^m \mathrm{e}^{\mathrm{i}pk} + \sum_{n=-\infty}^{\infty} Z_n B_n \hat{S}_{n-m}^p(\lambda). \tag{5.3}$$

The solution to this linear system for an inward resonant case is composed of a sum of two components, each corresponding to one of the terms on the right-hand side of (5.3). The first component is a constant multiple of the solution to the head-on incidence problem; therefore the case where $\psi_0 = 0$ is canonical to all inward resonances at the same frequency. In particular, this means that the most interesting features of the scattered field at inward resonance are purely symmetric, and this is a significant simplification, as we shall see. The second component is akin to the solution of an ordinary (nonresonant) scattering problem; its computation presents no special difficulty beyond those already discussed.

5.3. Head-on incidence ($\psi_0 = 0$). We now consider the head-on incidence case in detail. Note that subtraction of the infinite array solution is not required here, since $B_n = 0$ for all n [17]. Indeed, we also have $a_1 = -1$; therefore (5.3) is identical to (2.5). Also, the integrand in (4.3) no longer has poles at the points $t = \cos\psi_m$, $m \neq 0$, and therefore the scattered plane waves disappear in this case.

It turns out that the coefficients $\hat{C}_n^p$ decay more slowly as p increases than in the nonresonant cases discussed above. To see this, we must consider the boundary condition on the surface of the scatterers for large p. Therefore, we may leave aside for the present the possibility of Rayleigh–Bloch waves, since these independently satisfy the boundary conditions in the far field. Next, we impose the restriction $y < a$ and take the limit $x \to \infty$. Since the incident wave $\mathrm{e}^{\mathrm{i}kx}$ is present everywhere, it follows that there must be a simple pole above the path of integration in (4.3); otherwise $\phi_{\mathrm{sc}} \to 0$ and the boundary condition cannot be satisfied. In order for its contribution to possess the correct x dependence, this singularity must be located at $t = 1$. Indeed, from (4.14)–(4.16), we must have

$$\hat{C}_n^p = C_n^p \sim C_n p^{-1/2}\mathrm{e}^{\mathrm{i}kp} \tag{5.4}$$

as $p \to \infty$, so that $\hat{c}_n(t)$ has a branch point at $t = 1$ and the ratio $\hat{c}_n(t)/\gamma(t)$ has the required simple pole. The residue can also be deduced; thus from (4.3), (4.16), and (5.4),

$$\sum_{n=-\infty}^{\infty}(-\mathrm{i})^n Z_n C_n = -\sqrt{k/(2\pi)}\mathrm{e}^{\mathrm{i}\pi/4}, \tag{5.5}$$

so that

$$\phi \sim \mathrm{e}^{\mathrm{i}kx} - \mathrm{e}^{\mathrm{i}kr\cos\theta} \to 0 \tag{5.6}$$

as $x \to \infty$.

If the frequency is sufficiently low to permit the excitation of Rayleigh–Bloch waves, double filtering can be applied to (5.3); only the right-hand side differs from (3.11). Indeed, of the methods described in section 3, only double filtering yields accurate results in the head-on incidence case. This is due to the rate of decay of the coefficients $\hat{C}_n^p$ as $p \to \infty$ given by (5.4) being slower than that which occurs in other cases (cf. (3.7)).

Results in the appendix show that the coefficients C_n can be approximated via

$$C_n = \lim_{p\to\infty} \mathrm{e}^{-\mathrm{i}kp} C_n^p; \tag{5.7}$$

the computed values can then be checked using (5.5). At higher frequencies, i.e., for $k > \pi$, single filtering can be used with k in place of $\tilde{\beta}$. The values for C_n can then be approximated using the limit

$$C_n = \lim_{p\to\infty} \sqrt{p} \sum_{j=0}^{p} \mathrm{e}^{-\mathrm{i}kj} D_n^j. \tag{5.8}$$

Since $\hat{c}_n(t)$ is $2\pi/k$ periodic, it is evident that there are now branch points located at

$$t = \cos\psi_m = 1 + 2m\pi/k, \quad m \neq 0. \tag{5.9}$$

The path of integration in (4.3) is indented so as to pass below these singularities. We can deduce from (4.16) and (5.4) that

$$\hat{c}_n(t) = \frac{C_n \mathrm{e}^{-\mathrm{i}\pi/4}\sqrt{\pi/k}}{(t-\cos\psi_m)^{1/2}} + f_n(t), \tag{5.10}$$

where $f_n(t)$ is regular in the vicinity of the point $t = \cos\psi_m$ and the branch of the fractional power is chosen so that $(t-\cos\psi_m)^{1/2} = \sqrt{t-\cos\psi_m}$ for $t > \cos\psi_m$. Now the asymptotic behavior of an integral with branch points (that are not branch points of the exponent) is far more complicated than that of an integral with poles. The essential reason for this is that a rational function can easily be split into partial fractions, whereas a product of square roots cannot. Therefore we make the assumption that the neighborhood of the point $t = \cos\psi_m$ in which $f_n(t)$ is analytic is of sufficient size to permit the branch points of $\hat{c}_n(t)$ to be treated separately. This is valid if k is not too large. Of course, if $k < \pi$, the branch points of $\hat{c}_n(t)$ given by (5.9) lie outside the interval $[-1, 1]$ and are of no concern since their contribution to the far field is evanescent. Otherwise, if $\cos\psi_m \in (-1, 1)$, then we shall write $m \in \mathcal{M}$, as before. Note that this requires $m < 0$ and $k > -m\pi$.

When the saddle point in (4.3) lies to the right of a branch point (other than $t = -1$), then the steepest descent path is diverted in a counterclockwise loop around the cut, and an extra contribution must be included in the far field. The contribution from $t = \cos\psi_m$ is therefore present only in the regions where $\theta < \psi_m$ and $\theta > 2\pi - \psi_m$, which is the behavior exhibited by the scattered plane waves in the nonresonant case. Since the field is symmetric at head-on incidence, we give results for $y \geq 0$ only. Provided that we are not close to the array, we have, from (4.3) and (5.10),

$$\phi_{\mathrm{sc}} \sim \sqrt{\frac{2}{\pi kr}} \mathrm{e}^{-\mathrm{i}\pi/4} \left[h(\theta)\mathrm{e}^{\mathrm{i}kr} + \sum_{\substack{m\in\mathcal{M} \\ \psi_m > \theta}} \frac{\mathcal{G}(\psi_m)\mathrm{e}^{\mathrm{i}kr\cos(\theta-\psi_m)}}{\sqrt{\sin(\psi_m - \theta)}} \right], \tag{5.11}$$

in which

$$h(\theta) = \sum_{n=-\infty}^{\infty} (-\mathrm{i})^n Z_n \hat{c}_n(\cos\theta)\mathrm{e}^{\mathrm{i}n\theta}, \tag{5.12}$$

and

$$\mathcal{G}(\psi) = \sqrt{\frac{2\pi}{k\sin\psi}}\mathrm{e}^{\mathrm{i}\pi/4} \sum_{n=-\infty}^{\infty} Z_n C_n (-\mathrm{i})^n \mathrm{e}^{\mathrm{i}n\psi}. \tag{5.13}$$

Thus, the branch point contribution is not a circular wave, as its crests are linear, perpendicular to the line $\theta = \psi_m$. Noting from (5.9) that $\mathrm{e}^{\mathrm{i}n\psi_m} = 1 + O(k^{-1})$, it then follows from (5.5) that a multiplicative factor $\mathcal{G}(\psi_m)$ has no bearing on the asymptotic dependence upon k, to leading order.

5.4. Uniform asymptotics. The approximation (5.11) is nonuniform in the sense that it is singular at the shadow boundaries, where $\psi_m = \theta$. The uniform counterpart to (5.11) is given by

$$\begin{aligned}\phi_{\mathrm{sc}} \sim \mathrm{e}^{\mathrm{i}kr}\Bigg\{ & \mathrm{e}^{-\mathrm{i}\pi/4}\sqrt{\frac{2}{\pi kr}}h(\theta) - w\left(\mathrm{e}^{\mathrm{i}\pi/4}\zeta_0\right) + \frac{\mathrm{e}^{\mathrm{i}\pi/4}}{\zeta_0\sqrt{\pi}} - \sum_{m\in\mathcal{M}} \frac{\mathcal{G}(\psi_m)\mathrm{e}^{\mathrm{i}\pi/4}}{\sqrt{\pi}\sqrt[4]{kr}}\sqrt{\sec\frac{\psi_m-\theta}{2}} \\ & \times \left[\mathrm{e}^{\mathrm{i}(3\pi/8-\zeta_m^2/2)} D_{-1/2}\left(\sqrt{2}\mathrm{e}^{-\mathrm{i}\pi/4}\zeta_m\right) - \frac{1}{\sqrt[4]{2}(-\zeta_m)^{1/2}}\right]\Bigg\},\end{aligned} \tag{5.14}$$

in which $\zeta_m = \sqrt{2kr}\sin(\frac{1}{2}(\theta - \psi_m))$. Here, the term involving ζ_0 removes the singularity at $\theta = 0$. Note that the resonant mode actually has *no* region of existence, since $\zeta_0 \geq 0$. Nevertheless, its influence can be felt as $\theta \to 0$, since for small z, $w(z) = \mathrm{e}^{-z^2}[1 + O(z)]$. As in the nonresonant case, in the limit $\theta \to 0$, (5.14) accurately represents a contribution to the far field, though this is not necessarily the most significant. The final term involves the parabolic cylinder function $D_{-1/2}(\cdot)$ and can be obtained using methods outlined in [30], although this is by no means an easy procedure. However, it is relatively straightforward to check that (5.14) is correct. First, it is continuous across all of the shadow boundaries. To see this, we use (5.10) and (5.12) to show that the singular behavior of the term involving $h(\theta)$ at $\theta = \psi_m$ is cancelled by the series. Also, for large $|\zeta|$, we have from [31, section 9.246]

$$\mathrm{e}^{\mathrm{i}(3\pi/8-\zeta^2/2)} D_{-1/2}(\sqrt{2}\mathrm{e}^{-\mathrm{i}\pi/4}\zeta) = \frac{1 + \mathrm{i}\sqrt{2}\,\mathrm{H}(-\zeta)\mathrm{e}^{-\mathrm{i}\zeta^2}}{\sqrt[4]{2}(-\zeta)^{1/2}} + O(\zeta^{-5/2}), \tag{5.15}$$

where $\mathrm{H}(\cdot)$ is the Heaviside unit function and $(-\zeta_m)^{1/2}$ is either positive real or negative imaginary. If we now use this expansion, along with (4.9), in (5.14), we retrieve (5.11), as we should expect. Note that $D_{-1/2}(0) \approx 1.216$; therefore in the vicinity of the shadow boundary, the scattered field is $O((kr)^{-1/4})$, as in the case of point scatterers [29].

Figure 5.3 shows a contour plot of the real part of the total field, at head-on incidence with $a = 0.25$ and $k = 2.0$. The cancellation of the incident field close to the array is clearly apparent, so that the symmetric Rayleigh–Bloch wave (for which $|\alpha| \approx 0.542$) is clearly visible.

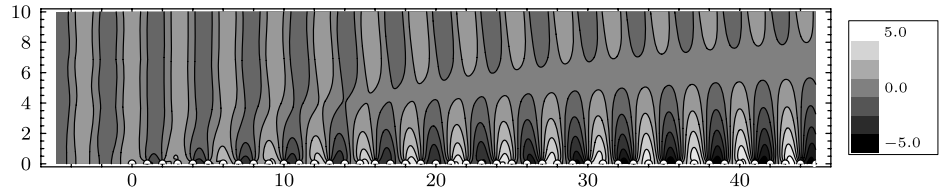


FIG. 5.3. *Contour plot of the real part of the total field,* $\mathrm{Re}[\phi]$ *at head-on incidence with* $a = 0.25$ *and* $k = 2.0$.

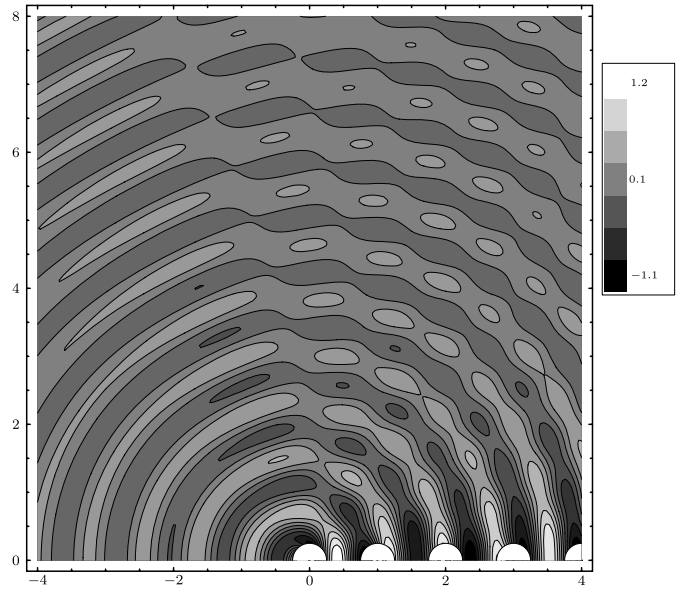


FIG. 5.4. *Contour plot of* $\mathrm{Re}[\phi_{\mathrm{sc}}]$*, at head-on incidence with* $a = 0.25$ *and* $k = 8.0$*, using Dirichlet boundary conditions.*

The branch point contributions that are significant at higher frequencies can most easily be observed in a plot of the scattered field. Thus, Figure 5.4 shows a contour plot of $\mathrm{Re}[\phi_{\mathrm{sc}}]$, with $a = 0.25$ and $k = 8.0$, using Dirichlet boundary conditions. Values of ψ_m for which $m \in \mathcal{M}$ and the associated values of $|\mathcal{G}(\psi_m)|$ are shown in Table 5.2. Far field plots with $r = 8$ are shown in Figure 5.5, with shadow boundaries indicated by black disks. The dashed line is computed from (5.14), whereas for the solid line, $D_{-1/2}(\sqrt{2}\mathrm{e}^{-\mathrm{i}\pi/4}\zeta)$ is replaced by $-\mathrm{i}D_{-1/2}(\sqrt{2}\mathrm{e}^{-\mathrm{i}\pi/4}|\zeta|)$ for $\zeta < 0$ so as to deactivate the branch point contributions. This plot is therefore discontinuous at the shadow boundaries, and the sizes of the discontinuities are consistent with the values of $|\mathcal{G}(\psi_m)|$ in Table 5.2. There are three regions to consider. For $\theta \gtrsim 0.69\pi$, no branch point contributions are present in the field, and the circular wave dominates. For smaller observation angles, a branch point contribution is activated, causing interference. Notice in particular the strong field close to the shadow boundary, where (5.14) predicts $O((kr)^{-1/4})$ behavior. A second branch point contribution is active when $\theta \lesssim 0.43\pi$. This is somewhat weaker and has a more limited effect on the field pattern.

TABLE 5.2
Shadow boundary locations and values of $\mathcal{G}(\psi_m)$ for the plot shown in Figure 5.1.

m	ψ_m	$\lvert\mathcal{G}(\psi_m)\rvert$
-2	0.69π	0.39
-1	0.43π	0.12

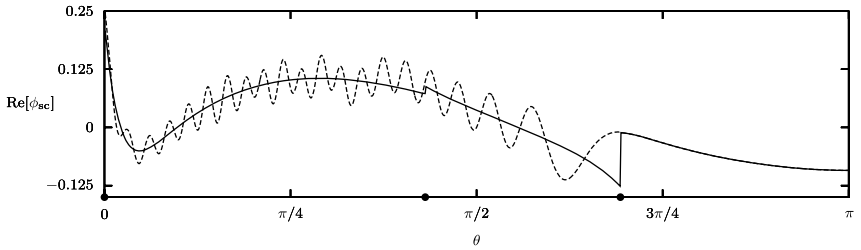


FIG. 5.5. *Far field plots for the parameters used in Figure* 5.4*, with $r = 8$. Correction terms are included for both shadow boundaries; the dashed line includes branch point contributions, whereas the solid line does not.*

A final possibility is that of *double resonance*, which requires that $k = n\pi$, $n \in \mathbb{N}$. In this case, if $\cos\psi_m = 1$, then $\cos\psi_{m-n} = -1$, so that modes m and $m - n$ are inward and outward resonant, respectively. Once the coefficients B_n for the infinite array problem have been obtained using the method in [17], the computation of the coefficients C_m^p for this case presents no special difficulty beyond those already discussed. However, the determination of the far field pattern involves a significant additional complication and will therefore appear in a future paper.

6. Conclusion. Problems involving semi-infinite arrays are notoriously difficult to solve accurately because the inevitable spatial truncation that has to be made can introduce significant errors. We have shown how infinite array subtraction, together with a novel filtering approach, can be used to obtain accurate solutions which can be computed efficiently for two-dimensional acoustic scattering of a plane wave by a semi-infinite array of rigid or soft circles. Unlike the case of isotropic point scatterers solved previously by one of the authors [8], this case is made considerably more complicated by the presence of Rayleigh–Bloch surface waves which can be excited along the array. We have presented methods which enable the amplitude of these modes to be computed accurately for the first time.

In nonresonant and outward resonant cases, the far field away from the array has been shown to be composed of the sum of a finite number of plane waves propagating in different directions and a circular wave emanating from the edge of the array. At inward resonance, the field can include additional terms that are neither circular waves nor plane waves. Uniform asymptotic expansions that vary continuously across all shadow boundaries have been derived.

If the incident field is generated by a line source rather than a plane wave, then the problem is much easier from a computational point of view. There is now no need to subtract an infinite array solution, since the decay of the source potential with distance means that the errors introduced by spatial truncation are small, and there is no need for double filtering because there are no resonances. Once the Rayleigh–Bloch waves are removed by single filtering, the remaining contribution in the coefficients D_n^p will decay like $p^{-3/2}$ as $p \to \infty$. The solution to the source-excitation problem has been

implemented by the authors; the results offer no extra insight, though the problem may be important for practical applications.

One of the main reasons for studying semi-infinite arrays is that they provide a tool with which to analyze large finite arrays. Thus we intend to use the techniques presented in this paper to study scattering by a long finite array under the assumption that the ends of the array are far enough apart to be treated separately.

Appendix. If we define

$$\Delta^p_m = \Gamma^p_m - \mathrm{e}^{\mathrm{i}\tilde{\beta}}\Gamma^{p-1}_m, \qquad p = 2, 3, \ldots, \tag{A.1}$$

then from (3.16) and (3.19) we obtain

$$E^p_m + \sum_{n=-\infty}^{\infty} Z_n \sum_{\substack{j=0\\ \neq p}}^{\infty} E^j_n X^{jp}_{n-m}\, H_{n-m}(k|j-p|) = \Delta^p_m - \mathrm{e}^{\mathrm{i}k}\Delta^{p-1}_m, \tag{A.2}$$

$p = 2, 3, \ldots, m \in \mathbb{Z}$. The equations in which $p = 0$ and $p = 1$ now require special treatment. From (3.19), we have

$$D^p_m = \sum_{j=0}^{p} \mathrm{e}^{\mathrm{i}(p-j)k} E^j_m, \qquad p = 0, 1, 2, \ldots. \tag{A.3}$$

Setting $p = 1$ in (3.16) we find that

$$\begin{aligned} \mathrm{e}^{\mathrm{i}k}E^0_m + E^1_m + \sum_{n=-\infty}^{\infty} E^0_n Z_n\, H_{n-m}(k) \\ + \sum_{n=-\infty}^{\infty} (-1)^{n-m} Z_n \sum_{j=0}^{\infty} E^j_n\, \mathrm{e}^{\mathrm{i}(1-j)k} \sum_{l=\max(j-1,1)}^{\infty} \mathrm{e}^{\mathrm{i}kl}\, H_{n-m}(kl) = \Delta^1_m. \end{aligned} \tag{A.4}$$

Finally from (3.17), for $p = 0$,

$$E^0_m + \sum_{n=-\infty}^{\infty} Z_n (-1)^{n-m} T_{nm} = \Gamma^0_m, \tag{A.5}$$

where

$$T_{nm} = \sum_{j=0}^{\infty}\sum_{q=0}^{j} E^q_n \mathrm{e}^{\mathrm{i}(j-q)k} \sum_{l=\max(j,1)}^{\infty} \mathrm{e}^{\mathrm{i}\tilde{\beta}(l-j)}\, H_{n-m}(kl). \tag{A.6}$$

This expression can be rearranged so that the sum over j becomes innermost. After evaluating the finite geometric series that appears, we obtain

$$T_{nm} = \frac{1}{\mathrm{e}^{\mathrm{i}\tilde{\beta}} - \mathrm{e}^{\mathrm{i}k}} \sum_{q=0}^{\infty} E^q_n \sum_{l=\max(q,1)}^{\infty} \left(\mathrm{e}^{\mathrm{i}\tilde{\beta}(1+l-q)} - \mathrm{e}^{\mathrm{i}k(1+l-q)}\right) H_{n-m}(kl). \tag{A.7}$$

The sum over l is again easily expressed in terms of the sums S^p_n defined in (3.4).

To reconstruct the coefficients C_m^p, we substitute (A.3) into (3.14), reverse the order of the summations, and evaluate the resulting geometric series to obtain

$$C_m^p = \frac{\mathrm{e}^{\mathrm{i}p\tilde{\beta}}}{1-\mathrm{e}^{\mathrm{i}(k-\tilde{\beta})}} \sum_{q=0}^{p} E_m^q \mathrm{e}^{-\mathrm{i}q\tilde{\beta}} + \frac{\mathrm{e}^{\mathrm{i}pk}}{1-\mathrm{e}^{-\mathrm{i}(k-\tilde{\beta})}} \sum_{q=0}^{p} E_m^q \mathrm{e}^{-\mathrm{i}qk} \tag{A.8}$$

$$= \frac{\mathrm{e}^{\mathrm{i}p\tilde{\beta}}}{1-\mathrm{e}^{\mathrm{i}(k-\tilde{\beta})}} \sum_{q=0}^{p} E_m^q \mathrm{e}^{-\mathrm{i}q\tilde{\beta}} + \frac{D_m^p}{1-\mathrm{e}^{-\mathrm{i}(k-\tilde{\beta})}}, \tag{A.9}$$

where we have used (3.19) to evaluate the second (telescopic) series.

The Rayleigh–Bloch amplitude α can then be retrieved by letting $p \to \infty$ and using (3.18):

$$\alpha \widetilde{B}_m = \frac{1}{1-\mathrm{e}^{\mathrm{i}(k-\tilde{\beta})}} \sum_{q=0}^{\infty} E_m^q \mathrm{e}^{-\mathrm{i}q\tilde{\beta}}. \tag{A.10}$$

Once again, asymptotic acceleration can be used in the approximate evaluation of this series, once spatial truncation has been applied.

REFERENCES

[1] B. Tomasic and A. Hessel, *Analysis of finite arrays—a new approach*, IEEE Trans. Antennas and Propagation, 47 (1999), pp. 555–565.

[2] D. S. Janning and B. A. Munk, *Effects of surface waves on the currents of truncated periodic arrays*, IEEE Trans. Antennas and Propagation, 50 (2002), pp. 1254–1265.

[3] R. W. Ziolkowski and N. Engheta, *Metamaterials special issue introduction*, IEEE Trans. Antennas and Propagation, 51 (2003), pp. 2546–2549.

[4] M. Kashiwagi, *Hydrodynamic interactions among a great number of columns supporting a very large flexible structure*, J. Fluids and Structures, 14 (2000), pp. 1013–1034.

[5] H. Kagemoto, M. Murai, M. Saito, B. Molin, and Š. Malenica, *Experimental and theoretical analysis of the wave decay along a long array of vertical cylinders*, J. Fluid Mech., 456 (2002), pp. 113–135.

[6] N. L. Hills and S. N. Karp, *Semi-infinite diffraction gratings.* I, Comm. Pure Appl. Math., 18 (1965), pp. 203–233.

[7] R. F. Millar, *Plane wave spectra in grating theory.* III. *Scattering by a semiinfinite grating of identical cylinders*, Canad. J. Phys., 42 (1964), pp. 1149–1184.

[8] C. M. Linton and P. A. Martin, *Semi-infinite arrays of isotropic point scatterers. A unified approach*, SIAM J. Appl. Math., 64 (2004), pp. 1035–1056.

[9] R. Porter and D. V. Evans, *Rayleigh-Bloch surface waves along periodic gratings and their connection with trapped modes in waveguides*, J. Fluid Mech., 386 (1999), pp. 233–258.

[10] C. M. Linton and M. McIver, *The existence of Rayleigh-Bloch surface waves*, J. Fluid Mech., 470 (2002), pp. 85–90.

[11] A. N. Norris and Z. Wang, *Bending-wave diffraction from strips and cracks on thin plates*, Q. J. Mech. Appl. Math., 47 (1994), pp. 607–627.

[12] H. D. Maniar and J. N. Newman, *Wave diffraction by a long array of cylinders*, J. Fluid Mech., 339 (1997), pp. 309–330.

[13] C. M. Linton and D. V. Evans, *The interaction of waves with arrays of vertical circular cylinders*, J. Fluid Mech., 215 (1990), pp. 549–569.

[14] V. Twersky, *Elementary function representation of Schlömilch series*, Arch. Rational Mech. Anal., 8 (1961), pp. 323–332.

[15] C. M. Linton, *The Green's function for the two-dimensional Helmholtz equation in periodic domains*, J. Engrg. Math., 33 (1998), pp. 377–402.

[16] P. A. Martin, *Multiple Scattering. Interaction of Time-Harmonic Waves with N Obstacles*, Cambridge University Press, Cambridge, UK, 2006.

[17] C. M. Linton and I. Thompson, *Resonant effects in scattering by periodic arrays*, Wave Motion, 44 (2007), pp. 167–175.

[18] P. McIver, C. M. Linton, and M. McIver, *Construction of trapped modes for wave guides and diffraction gratings*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 454 (1998), pp. 2593–2616.

[19] D. V. Evans and R. Porter, *Trapping and near-trapping by arrays of cylinders in waves*, J. Engrg. Math., 35 (1999), pp. 149–179.

[20] A.-S. Bonnet-Bendhia and F. Starling, *Guided waves by electromagnetic gratings and non-uniqueness examples for the diffraction problem*, Math. Methods Appl. Sci., 17 (1994), pp. 305–338.

[21] M. Nishimoto and H. Ikuno, *Space-wavenumber analysis of field scattered from a semi-infinite strip grating*, Electr. Eng. Japan, 132 (2000), pp. 1–8.

[22] R. Porter and D. V. Evans, *Scattering of flexural waves by multiple narrow cracks in ice sheets floating on water*, Wave Motion, 43 (2006), pp. 425–443.

[23] C. M. Linton, *Schlömilch series that arise in diffraction theory and their efficient computation*, J. Phys. A, 39 (2006), pp. 3325–3339.

[24] R. J. Mailloux, *Excitation of a surface wave along an infinite Yagi–Uda array*, IEEE Trans. Antennas and Propagation, 13 (1965), pp. 719–724.

[25] J. P. Skinner and P. J. Collins, *A one-sided version of the Poisson sum formula for semi-infinite array Green's functions*, IEEE Trans. Antennas and Propagation, 45 (1997), pp. 601–607.

[26] M. A. Sumbatyan and A. Scalia, *Equations of Mathematical Diffraction Theory*, Chapman & Hall/CRC, Boca Raton, FL, 2005.

[27] I. Thompson, *An improved uniform approximation for diffraction integrals*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 462 (2006), pp. 1341–1353.

[28] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[29] N. L. Hills, *Semi-infinite diffraction gratings.* II. *Inward resonance*, Comm. Pure Appl. Math., 18 (1965), pp. 389–395.

[30] N. Bleistein and R. A. Handelsman, *Asymptotic Expansions of Integrals*, 2nd ed., Dover, New York, 1986.

[31] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series, and Products*, 6th ed., Academic Press, San Diego, 2000.

SIAM J. APPL. MATH.
Vol. 67, No. 5, pp. 1259–1282

ON THE CONVERGENCE OF THE HARMONIC B_Z ALGORITHM IN MAGNETIC RESONANCE ELECTRICAL IMPEDANCE TOMOGRAPHY*

J. J. LIU†, J. K. SEO‡, M. SINI§, AND E. J. WOO¶

Abstract. Magnetic resonance electrical impedance tomography (MREIT) is a new medical imaging technique that aims to provide electrical conductivity images with sufficiently high spatial resolution and accuracy. A new MREIT image reconstruction method called the harmonic B_z algorithm was proposed in 2002, and it is based on the measurement of B_z that is a single component of an induced magnetic flux density $\mathbf{B} = (B_x, B_y, B_z)$ subject to an injection current. Since then, MREIT imaging techniques have made significant progress, and recent published numerical simulations and phantom experiments show that we can produce high-quality conductivity images when the conductivity contrast is not very high. Though numerical simulations can explain why we could successfully distinguish different tissues with small conductivity differences, a rigorous mathematical analysis is required to better understand the underlying physical and mathematical principle. The purpose of this paper is to provide such a mathematical analysis of those numerical simulations and experimental results. By using a uniform a priori estimate for the solution of the elliptic equation in the divergent form and an induction argument, we show that, for a relatively small contrast of the target conductivity, the iterative harmonic B_z algorithm with a good initial guess is stable and exponentially convergent in the continuous norm. Both two- and three-dimensional versions of the algorithm are considered, and the difference in the convergence property of these two cases is analyzed. Some numerical results are also given to show the expected exponential convergence behavior.

Key words. MREIT, conductivity, image reconstruction, convergence

AMS subject classifications. 35R30, 35J05, 76Q05

DOI. 10.1137/060661892

1. Introduction.

Magnetic resonance electrical impedance tomography (MREIT) is an electrical conductivity imaging technique using a magnetic resonance imaging (MRI) scanner with a current injection apparatus. Since the early 1980s, there have been significant efforts to produce cross-sectional images of a conductivity distribution σ inside a three-dimensional body Ω using boundary measurements of current-voltage data (Neumann-to-Dirichlet data) satisfying the elliptic equation $\nabla \cdot (\sigma \nabla u) = 0$ in Ω, and this technique has been called electrical impedance tomography (EIT) [4, 14, 21]. Here u denotes the electric potential inside Ω. It is well known that EIT has suffered from the ill-posedness of the corresponding inverse problem related with the insensitivity of Cauchy data on the boundary $\partial\Omega$ to any internal local change of σ. Acquisition of accurate Cauchy data on $\partial\Omega$ requires a sophisticated EIT instrument and a large number of surface electrodes. In practice, however, the cum-

*Received by the editors June 2, 2006; accepted for publication (in revised form) March 15, 2007; published electronically June 15, 2007. This work was supported by the SRC/ERC program of MOST/ KOSEF (R11-2002-103 and R01-2005-000-10339-0).

http://www.siam.org/journals/siap/67-5/66189.html

†Department of Mathematics, Southeast University, Nanjing, 210096, People's Republic of China (jjliu@seu.edu.cn). This author's work was supported by NSFC grant 10371018 and the Impedance Imaging Research Center (IIRC) at Kyung Hee University.

‡Department of Mathematics, Yonsei University, Seoul, 120-749, Korea (seoj@yonsei.ac.kr).

§Department of Mathematics, Yonsei University, Seoul, 120-749, Korea. Current address: RICAM, Austrian Academy of Sciences, A-4040 Linz, Austria (mourad.sini@oeaw.ac.at). This author's work was supported by the BK21 project at Yonsei University and the IIRC at Kyung Hee University.

¶College of Electronics and Information, Kyung Hee University, Kyungki, 449-701, Korea (ejwoo@khu.ac.kr).

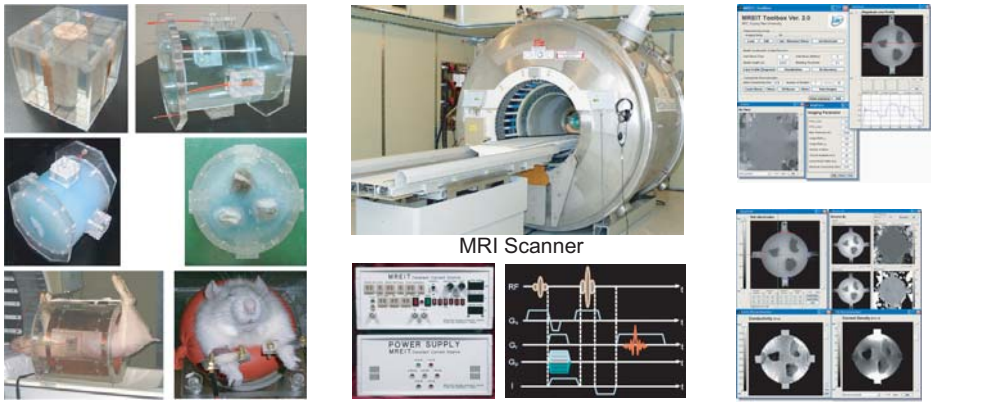


FIG. 1. *MREIT system at Impedance Imaging Research Center in Korea and image reconstruction software.*

bersome procedure to attach many electrodes is prone to increase measurement errors in addition to electronic noise and various artifacts. Furthermore, there exist uncertainties in terms of electrode positions and boundary shape of an imaging subject. Due to the ill-posedness and the errors originating from these practical difficulties, the spatial resolution and accuracy of EIT images are relatively poor, and therefore its applicability has been limited in the clinical environment.

On the other hand, MREIT takes advantage of the internal information of B_z, the z-component of the magnetic flux density distribution $\mathbf{B} = (B_x, B_y, B_z)$ induced by the internal current density $\mathbf{J} = -\sigma \nabla u$ subject to an injection current through a pair of surface electrodes. The B_z data can be measured by using an MRI scanner as illustrated in Figure 1. Here the z-axis is the direction of the main magnetic field of the MRI scanner. MREIT utilizes the fact that the data B_z convey the information about any local change of σ via the Biot–Savart law:

$$B_z(x,y,z) = \frac{\mu_0}{4\pi} \int_\Omega \frac{\sigma(\mathbf{r}) \left[(x-x') \frac{\partial u}{\partial y}(\mathbf{r}') - (y-y') \frac{\partial u}{\partial x}(\mathbf{r}') \right]}{|\mathbf{r}-\mathbf{r}'|^3} \, d\mathbf{r}', \ \mathbf{r} = (x,y,z) \in \Omega.$$

This supplementary use of the internal B_z data enables MREIT to bypass the ill-posedness problem in EIT. In early 2002, the first constructive B_z-based MREIT algorithm called the harmonic B_z algorithm was proposed in [18]. Since then, MREIT has advanced rapidly and now is at the stage of animal experiments [13].

The harmonic B_z algorithm is based on the following curl of the Ampere law:

$$\frac{1}{\mu_0} \Delta B_z = \left\langle \nabla \sigma \, , \, \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \nabla u \right\rangle, \tag{1.1}$$

where μ_0 is the magnetic permeability of the free space, $\langle \cdot, \cdot \rangle$ is the inner product, and Δ denotes the Laplacian. Recent published numerical simulations and phantom experiments show that conductivity images with high spatial resolution are achievable [9, 10, 15, 16, 17, 19]. Figure 2 shows a schematic diagram of the harmonic B_z algorithm and typical MREIT images of a conductivity phantom including three chunks of biological tissues having different conductivity values inside a cylindrical container Ω.

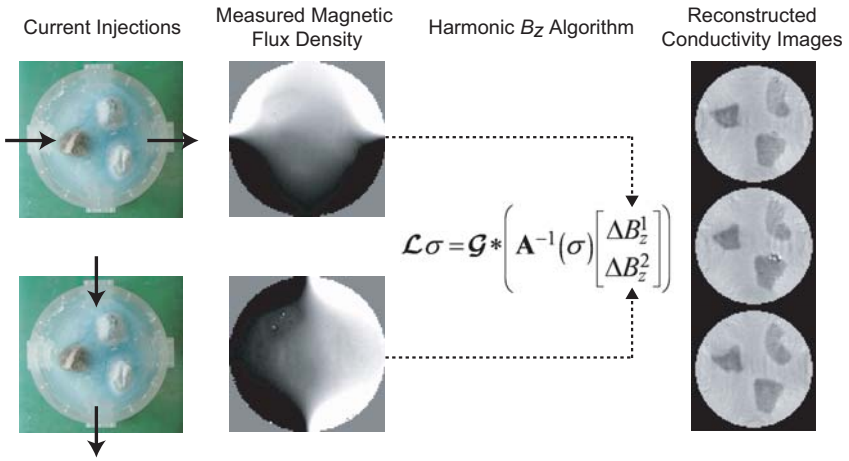


FIG. 2. *Overview of the harmonic B_z algorithm.*

Although the harmonic B_z algorithm shows a remarkable performance in various numerical simulations and phantom experiments, rigorous mathematical theories regarding its convergence behavior have not been supported yet. The purpose of this paper is to deal with this convergence analysis rigorously. For a suitably constructed admissible iteration set in terms of a priori information about the target conductivity, we can prove that the sequence $\{\sigma^n\}$ is uniformly bounded by a uniform estimate on the solution to the elliptic equation in the divergent form. Using this a priori estimate and an induction argument, we show that, in both two- and three-dimensional cases, the harmonic B_z algorithm is stable and exponentially convergent, provided that the contrast of the target conductivity distribution is not very high. It is impossible to get the C^1 convergence in the three-dimensional problem even when the harmonic B_z algorithm is applied to a target conductivity distribution with a small contrast. This mathematical difficulty comes from the algorithm itself. With this theoretical result, we partially answer the question on the applicable scope of the harmonic B_z algorithm and explain the fast convergent phenomena arising in numerical tests. We refer to our recent article [12], which briefly discusses this convergence issue using special examples of two-dimensional conductivity distributions.

It seems that the small contrast in the target conductivity is necessary for the convergence of the harmonic B_z iteration scheme, provided that the input current is not so large. This phenomenon can be explained physically. For a given input current from the boundary, if the conductivity has a large jump inside the medium, then the current going through will be small, and therefore the magnetic flux density will also be weak. For a relatively high contrast of the conductivity distribution, the algorithm needs to be adapted to control the representation of this contrast in the iteration process. This issue should be considered in the future.

2. Exact mathematical model of MREIT. Since our goal is to use the MREIT technique in practical clinical applications, we must set up an exact mathematical model of MREIT that agrees with a planned medical imaging system. To simplify our study, let us make several assumptions which should not go astray from the practical model. Let the subject to be imaged occupy a three-dimensional bounded domain $\Omega \subset \mathbb{R}^3$ with a smooth connected boundary $\partial\Omega$, and each $\Omega_{z_0} := \Omega \cap \{z =$

$z_0\} \subset \mathbb{R}^2$, the slice of Ω cut by the plane $\{z = z_0\}$, has a smooth connected boundary. We assume that the conductivity distribution σ of the subject Ω is isotropic, $C^1(\overline{\Omega})$, and $0 < \sigma_- < \sigma < \sigma_+$ with two known constants $\sigma_\pm$. Though σ is usually piecewise-smooth in practice, this can be approximated by the $C^1(\overline{\Omega})$ function, and so it is a matter of how big $\|\sigma\|_{C^1(\Omega)}$ is. We attach a pair of copper electrodes $\mathcal{E}^+$ and $\mathcal{E}^-$ on $\partial\Omega$ in order to inject current, and let $\mathcal{E}^+ \cup \mathcal{E}^-$ be the portion of the surface $\partial\Omega$ where electrodes are attached; see Figure 2.

The injection current I produces an internal current density $\mathbf{J} = (J_x, J_y, J_z)$ inside the subject Ω satisfying the following problem:

$$\begin{cases} \nabla \cdot \mathbf{J} = 0 \quad \text{in } \Omega, \\ I = -\displaystyle\int_{\mathcal{E}^+} \mathbf{J} \cdot \mathbf{n} ds = \int_{\mathcal{E}^-} \mathbf{J} \cdot \mathbf{n} ds, \quad \mathbf{J} \times \mathbf{n} = 0 \ \text{ on } \mathcal{E}^+ \cup \mathcal{E}^-, \\ \mathbf{J} \cdot \mathbf{n} = 0 \quad \text{on } \partial\Omega \setminus \overline{\mathcal{E}^+ \cup \mathcal{E}^-}, \end{cases} \tag{2.1}$$

where $\mathbf{n}$ is the outward unit normal vector on $\partial\Omega$ and ds the surface area element. The condition of $\mathbf{J} \times \mathbf{n} = 0$ on $\mathcal{E}^+ \cup \mathcal{E}^-$ comes from the fact that copper electrodes are considered as perfect conductors. Since $\mathbf{J}$ is expressed as $\mathbf{J} = -\sigma\nabla u$, where u is the corresponding electrical potential, (2.1) can be converted to

$$\begin{cases} \nabla \cdot (\sigma \nabla u) = 0 \quad \text{in } \Omega, \\ I = \displaystyle\int_{\mathcal{E}^+} \sigma \frac{\partial u}{\partial \mathbf{n}} \, ds = -\int_{\mathcal{E}^-} \sigma \frac{\partial u}{\partial \mathbf{n}} ds, \quad \nabla u \times \mathbf{n} = 0 \ \text{ on } \mathcal{E}^+ \cup \mathcal{E}^-, \\ \sigma \dfrac{\partial u}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega \setminus \overline{\mathcal{E}^+ \cup \mathcal{E}^-}, \end{cases} \tag{2.2}$$

where $\frac{\partial u}{\partial \mathbf{n}} = \nabla u \cdot \mathbf{n}$. The above nonstandard boundary value problem (2.2) is well-posed and has a unique solution up to a constant. We omit the proof of the uniqueness (up to a constant) within the class $W^{1,2}(\Omega)$ since it follows from the standard argument in the PDE.

Let us briefly discuss the boundary conditions that are essentially related with the size of electrodes. The condition $\nabla u \times \mathbf{n}|_{\mathcal{E}^\pm} = 0$ ensures that each of $u|_{\mathcal{E}^+}$ and $u|_{\mathcal{E}^-}$ is a constant, since ∇u is normal to its level surface. The term $\pm I = \int_{\mathcal{E}^\pm} \sigma \frac{\partial u}{\partial \mathbf{n}} \, ds$ means that the total amount of injection current through the electrodes is I mA. Let us denote $g := -\sigma \frac{\partial u}{\partial \mathbf{n}}|_{\partial\Omega}$. In practice, it is difficult to specify the Neumann data g in a pointwise sense because only the total amount of injection current I is known. It should be noticed that the boundary condition in (2.2) leads to $|g| = \infty$ on $\partial\mathcal{E}^\pm$, singularity along the boundary of electrodes, and $g \notin L^2(\partial\Omega)$. But fortunately $g \in H^{-1/2}(\partial\Omega)$, which also can be proven by the standard regularity theory in the PDE.

The exact model (2.2) can be converted into the following standard problem of an elliptic equation with mixed boundary conditions.

LEMMA 2.1. *Assume that $\tilde{u}$ solves*

$$\begin{cases} \nabla \cdot (\sigma \nabla \tilde{u}) = 0 \quad \textit{in} \quad \Omega, \\ \tilde{u}|_{\mathcal{E}^+} = 1, \ \tilde{u}|_{\mathcal{E}^-} = 0, \\ -\sigma \dfrac{\partial \tilde{u}}{\partial \mathbf{n}} = 0 \quad \textit{on} \quad \partial\Omega \setminus (\mathcal{E}^+ \bigcup \mathcal{E}^-). \end{cases} \tag{2.3}$$

If u is a solution of the mixed boundary value problem (2.2)*, then*

$$u = \frac{I}{\int_{\partial\mathcal{E}^+} \sigma \frac{\partial \tilde{u}}{\partial \mathbf{n}}\, ds}\tilde{u} \quad \text{in } \Omega \qquad \textit{(up to a constant).} \tag{2.4}$$

Proof. The proof is elementary by looking at the energy of $w = u - c\tilde{u}$ for a constant c:

$$\begin{aligned}\int_\Omega \sigma|\nabla w|^2\, d\mathbf{r} &= \int_{\partial\Omega} \sigma \frac{\partial w}{\partial \mathbf{n}}\, w\, ds \\ &= \int_{\mathcal{E}^+} \sigma \frac{\partial w}{\partial \mathbf{n}}\, ds\, (u|_{\mathcal{E}^+} - c) + \left(\int_{\mathcal{E}^-} \sigma \frac{\partial w}{\partial \mathbf{n}}\, ds\right) u|_{\mathcal{E}^-} \\ &= (u|_{\mathcal{E}^+} - u|_{\mathcal{E}^-} - c)\left(I - c\int_{\partial\mathcal{E}^+} \sigma \frac{\partial \tilde{u}}{\partial \mathbf{n}} ds\right).\end{aligned}$$

Hence, for $c = \frac{I}{\int_{\partial\mathcal{E}^+} \sigma \frac{\partial \tilde{u}}{\partial \mathbf{n}}\, ds}$, the above relation generates $|\nabla w| = 0$ in Ω, which means w is a constant in Ω. ∎

Now we explain the inverse problem for the MREIT model, in which we try to reconstruct σ. The presence of the internal current density $\mathbf{J} = -\sigma\nabla u$ generates a magnetic flux density $\mathbf{B} = (B_x, B_y, B_z)$ such that the Ampere law $\mathbf{J} = \nabla \times \mathbf{B}/\mu_0$ holds in Ω. With the z-axis pointing to the direction of the main magnetic field of the MRI scanner, the relation between the measurable quantity B_z and the unknown σ is governed by the Biot–Savart law:

$$B_z(\mathbf{r}) = \frac{\mu_0}{4\pi}\int_\Omega \frac{\langle \mathbf{r} - \mathbf{r}',\ \sigma(\mathbf{r}')\mathbb{L}\nabla u(\mathbf{r}')\rangle}{|\mathbf{r} - \mathbf{r}'|^3}\, d\mathbf{r}' \qquad \text{for } \mathbf{r} \in \Omega, \tag{2.5}$$

where

$$\mathbb{L} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Here we must read u as a nonlinear function of σ. The following lemma is crucial to understand why we need at least two injection currents with the requirement (2.11) in what follows.

LEMMA 2.2. *Suppose u is a solution of* (2.2) *and the pair (σ, u) satisfies* (2.5)*. Then B_z in* (2.5) *can be expressed as*

$$B_z = \frac{\mu_0}{4\pi}\int_\Omega \frac{-1}{|\mathbf{r} - \mathbf{r}'|}\begin{vmatrix} \frac{\partial\sigma}{\partial x} & \frac{\partial\sigma}{\partial y} \\ \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \end{vmatrix} d\mathbf{r}' + \frac{\mu_0}{4\pi}\int_{\partial\Omega} \frac{1}{|\mathbf{r} - \mathbf{r}'|}\, \mathbf{n}\cdot(\sigma\mathbb{L}\nabla u)\, ds. \tag{2.6}$$

Moreover, there exist infinitely many pairs $(\tilde{\sigma}, \tilde{u})$ such that

$$\begin{vmatrix} \frac{\partial\sigma}{\partial x} & \frac{\partial\sigma}{\partial y} \\ \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \end{vmatrix} = \begin{vmatrix} \frac{\partial\tilde{\sigma}}{\partial x} & \frac{\partial\tilde{\sigma}}{\partial y} \\ \frac{\partial\tilde{u}}{\partial x} & \frac{\partial\tilde{u}}{\partial y} \end{vmatrix}$$

in Ω and $\mathbf{n}\cdot(\sigma\mathbb{L}\nabla u) = \mathbf{n}\cdot(\tilde{\sigma}\mathbb{L}\nabla\tilde{u})$ on $\partial\Omega$.

Proof. From (2.5), we have

$$\begin{aligned} B_z &= \frac{\mu_0}{4\pi}\int_\Omega \nabla_{\mathbf{r}'}\frac{1}{|\mathbf{r} - \mathbf{r}'|}\cdot(\sigma(\mathbf{r}')\mathbb{L}\nabla u(\mathbf{r}'))\, d\mathbf{r}' \\ &= \frac{\mu_0}{4\pi}\int_\Omega \frac{-1}{|\mathbf{r} - \mathbf{r}'|}\nabla\cdot(\sigma\mathbb{L}\nabla u)\, d\mathbf{r}' + \frac{\mu_0}{4\pi}\int_{\partial\Omega}\frac{1}{|\mathbf{r} - \mathbf{r}'|}\, \mathbf{n}\cdot(\sigma\mathbb{L}\nabla u)\, ds.\end{aligned}$$

Then (2.6) follows from

$$\nabla \cdot (\sigma \mathbb{L} \nabla u) = \mathbf{e}_z \cdot [\nabla \sigma \times \nabla u] = \begin{vmatrix} \frac{\partial \sigma}{\partial x} & \frac{\partial \sigma}{\partial y} \\ \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \end{vmatrix},$$

where $\mathbf{e}_z = (0, 0, 1)$.

Now we will show that there are infinitely many pairs $(\tilde{\sigma}, \tilde{u})$ such that $\mathbf{e}_z \cdot [\nabla \sigma \times \nabla u] = \mathbf{e}_z \cdot [\nabla \tilde{\sigma} \times \nabla \tilde{u}]$ and $\tilde{u}$ is a solution of (2.2) with σ replaced by $\tilde{\sigma}$. Indeed, we can construct infinitely many pairs $(\tilde{\sigma}, \tilde{u})$ satisfying the much stronger condition $\sigma \nabla u = \tilde{\sigma} \nabla \tilde{u}$. From the maximum-minimum principle for an elliptic equation, $u|_{\mathcal{E}^+}$ and $u|_{\mathcal{E}^-}$ are the maximum and minimum values of u in $\overline{\Omega}$, respectively. Choose a and b such that $\inf_\Omega u = u|_{\mathcal{E}^-} < a < b < u|_{\mathcal{E}^+} = \sup_\Omega u$. For any increasing function $\phi \in C^2([a, b])$ satisfying

$$\phi'(a) = \phi'(b) = 1, \quad \phi''(a) = \phi''(b) = 0, \quad \phi(a) = a, \ \phi(b) = b, \tag{2.7}$$

we define

$$\tilde{u}(\mathbf{r}) = \begin{cases} \phi(u(\mathbf{r})) & \text{if } \mathbf{r} \in \hat{\Omega}, \\ u(\mathbf{r}) & \text{if } \mathbf{r} \in \Omega \setminus \hat{\Omega}, \end{cases} \qquad \tilde{\sigma}(\mathbf{r}) = \begin{cases} \frac{\sigma(\mathbf{r})}{\phi'(u(\mathbf{r}))} & \text{if } \mathbf{r} \in \hat{\Omega}, \\ \sigma(\mathbf{r}) & \text{if } \mathbf{r} \in \Omega \setminus \hat{\Omega}, \end{cases}$$

where $\hat{\Omega} := \{\mathbf{r} \in \Omega \ : \ a \le u(\mathbf{r}) \le b\}$. The conditions of ϕ guarantee $\tilde{\sigma} \in C^1(\Omega)$ and $\tilde{\sigma} > 0$ in Ω. Since $\nabla \tilde{u} = \phi'(u) \nabla u$, we have $\tilde{\sigma} \nabla \tilde{u} = \frac{\sigma}{\phi'(u)} \nabla \hat{u} = \sigma \nabla u$. So it is clear that

$$\begin{vmatrix} \frac{\partial \sigma}{\partial x} & \frac{\partial \sigma}{\partial y} \\ \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \end{vmatrix} = \begin{vmatrix} \frac{\partial \hat{\sigma}}{\partial x} & \frac{\partial \hat{\sigma}}{\partial y} \\ \frac{\partial \hat{u}}{\partial x} & \frac{\partial \hat{u}}{\partial y} \end{vmatrix}$$

and $\mathbf{n} \cdot (\sigma \mathbb{L} \nabla u) = \mathbf{n} \cdot (\tilde{\sigma} \mathbb{L} \nabla \tilde{u})$ on $\partial\Omega$. Since $\tilde{u} = u$ near the electrodes $\mathcal{E}^+$ and $\mathcal{E}^-$, $\tilde{u}$ has the same boundary condition on the electrodes as u. Therefore, $\tilde{u}$ is a solution of (2.2) with σ replaced by $\tilde{\sigma}$. This completes the proof since ϕ can be chosen arbitrarily under the constraint (2.7). ∎

According to Lemma 2.2, the unique determination of σ requires us to inject at least two input currents I_1 and I_2. Now we are ready to explain the exact MREIT model. We inject electrical currents I_1 and I_2 through two pairs of surface electrodes $\mathcal{E}_1^\pm$ and $\mathcal{E}_2^\pm$, respectively. Let u_j and B_z^j be the potential and magnetic flux density, respectively, corresponding to I_j, with $j = 1, 2$.

For the measured data B_z^1, B_z^2 corresponding to two input currents I_1, I_2 and a given constant $\alpha > 0$, we try to reconstruct σ satisfying the following conditions for $j = 1, 2$:

$$\begin{cases} \nabla \cdot (\sigma \nabla u_j) = 0 \quad \text{in } \Omega, \\ I_j = \displaystyle\int_{\mathcal{E}_j^+} \sigma \frac{\partial u_j}{\partial \mathbf{n}} \, ds = -\int_{\mathcal{E}_j^-} \sigma \frac{\partial u_j}{\partial \mathbf{n}}, \quad \nabla u_j \times \mathbf{n}|_{\mathcal{E}_j^+ \cup \mathcal{E}_j^-} = 0, \\ \sigma \dfrac{\partial u_j}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega \setminus \overline{\mathcal{E}_j^+ \cup \mathcal{E}_j^-}, \\ B_z^j(\mathbf{r}) = \dfrac{\mu_0}{4\pi} \displaystyle\int_\Omega \frac{\langle \mathbf{r} - \mathbf{r}' \, , \, \sigma \mathbb{L} \nabla u_j \rangle}{|\mathbf{r} - \mathbf{r}'|^3} \, d\mathbf{r}', \quad \mathbf{r} \in \Omega, \\ \left| u_1|_{\mathcal{E}_2^+} - u_1|_{\mathcal{E}_2^-} \right| = \alpha. \end{cases} \tag{2.8}$$

Remark 2.3. The last condition regarding α is necessary for fixing the scaling uncertainty of σ. Without this condition, whenever σ and u_j satisfy the other four relations in (2.8), so do $c\sigma$ and $\frac{u_j}{c}$ for any positive constant c. Here we should avoid measuring the voltage difference between the pair of electrodes used for current injection since any electrode contact impedance may cause measurement errors. Therefore, in practice, we usually use the other pair of electrodes for the voltage measurement.

To explain the MREIT image reconstruction algorithm, we define

$$\mathcal{L}_{z_0}\sigma(x,y) := \sigma(x,y,z_0) + \frac{1}{2\pi}\int_{\partial\Omega_{z_0}} \frac{(x-x',\ y-y')\ \cdot\ \nu(x',y')}{|x-x'|^2+|y-y'|^2}\ \sigma(x',y',z_0)\ dl, \tag{2.9}$$

where ν is the unit outward normal vector to $\partial\Omega_{z_0}$ and $\Omega_{z_0} := \Omega \cap \{z = z_0\} \subset \mathbb{R}^2$. For a vector-valued function $F = (F_1, F_2)$ defined on Ω, we define

$$\mathcal{G}_{z_0} * F(x,y) := \frac{1}{2\pi\mu_0}\int_{\Omega_{z_0}} \frac{(x-x',\ y-y')}{|x-x'|^2+|y-y'|^2}\cdot F(x',y',z_0)\ dx'dy'. \tag{2.10}$$

Let each $u_j[\sigma]$ be a solution to the direct problem (2.2) corresponding to I_j satisfying

$$\begin{vmatrix} \frac{\partial u_1}{\partial x} & \frac{\partial u_1}{\partial y} \\ \frac{\partial u_2}{\partial x} & \frac{\partial u_2}{\partial y} \end{vmatrix} \neq 0 \quad \text{in } \Omega, \tag{2.11}$$

and set

$$\mathbb{A}[\sigma] := \begin{bmatrix} \frac{\partial u_1[\sigma]}{\partial y} & -\frac{\partial u_1[\sigma]}{\partial x} \\ \frac{\partial u_2[\sigma]}{\partial y} & -\frac{\partial u_2[\sigma]}{\partial x} \end{bmatrix}. \tag{2.12}$$

Now let us state the implicit relation between σ and B_z^j, on which the harmonic B_z algorithm is based.

LEMMA 2.4. *Suppose that $|\nabla\sigma|$ is compactly supported in Ω and $u_j[\sigma]$ for $j = 1, 2$ satisfy* (2.11). *Then the following identity:*

$$\mathcal{L}_z\sigma(x,y) = \mathcal{G}_z * \left(\mathbb{A}[\sigma]^{-1}\begin{bmatrix} \nabla^2 B_z^1 \\ \nabla^2 B_z^2 \end{bmatrix}\right)(x,y), \qquad (x,y)\in\Omega_z \tag{2.13}$$

holds for each z. Moreover, $\mathcal{L}_z : H_*^{1/2}(\Omega_z) \to H_*^{1/2}(\Omega_z)$ is invertible where $H_*^{1/2}(\Omega_z) := \{\eta \in H^{1/2}(\Omega_z) : \int_{\partial\Omega_z}\eta = 0\}$.

Proof. The proof of (2.13) is based on the fact that $\nabla\cdot\mathbf{B} = 0$ and the Ampere law $\mathbf{J} = \frac{1}{\mu_0}\nabla\times\mathbf{B}$. Direct computation yields $\nabla u_j \times \nabla\sigma = \frac{1}{\mu_0}\Delta\mathbf{B}^j$, and we have

$$\begin{bmatrix} \frac{\partial\sigma}{\partial x} \\ \frac{\partial\sigma}{\partial y} \end{bmatrix} = \frac{1}{\mu_0}\mathbb{A}[\sigma]^{-1}\begin{bmatrix} \Delta B_z^1 \\ \Delta B_z^2 \end{bmatrix}. \tag{2.14}$$

Since $\nabla\sigma = 0$ near $\partial\Omega$, so does $\Delta B_z = 0$. Hence, the right-hand side of (2.13) is well defined. Now the result follows from the formal identity

$$\sigma(x,y,z) = \int_{\Omega_z}\frac{1}{2\pi}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\log\sqrt{(x-x')^2+(y-y')^2}\,\sigma(x',y',z)dx'dy'$$

and integration by parts.

The invertibility of $\mathcal{L}$ can be proved by the standard layer potential theory [7, 20]. Let $w \in H_*^{1/2}(\Omega_z)$. We will find $v \in H_*^{1/2}(\Omega_z)$ such that $\mathcal{L}_z v = w$. Note that $w|_{\partial\Omega_z} \in L_*^2(\partial\Omega_z) := \{\phi \in L^2(\Omega_z) : \int_{\partial\Omega_z} \phi = 0\}$. It is well known that there exists a unique $\psi \in L_*^2(\partial\Omega_z)$ such that $\frac{1}{2}\psi - \mathcal{K}\psi = w|_{\partial\Omega_z}$ on $\partial\Omega_z$, where $\mathcal{K}\psi(x,y) = \frac{-1}{2\pi}\int_{\partial\Omega_{z_0}} \frac{(x-x',\, y-y')\cdot\nu}{|x-x'|^2+|y-y'|^2}\,\psi(x',y')\,dl$ for $(x,y) \in \partial\Omega_z$. Now we define

$$v(x,y) = w(x,y) - \frac{1}{2\pi}\int_{\partial\Omega_{z_0}} \frac{(x-x',\, y-y')\,\cdot\,\nu(x',y')}{|x-x'|^2+|y-y'|^2}\,\psi(x',y')\,dl \tag{2.15}$$

for $(x,y) \in \Omega_z$. Due to the trace formula of the double layer potential, $v = \psi$ on $\partial\Omega_z$. By replacing ψ in (2.15) with v, we have $w = \mathcal{L}_z v$, and this completes the proof. ∎

Remark 2.5. The condition (2.11) is necessary for the harmonic B_z algorithm. However, we still do not have a rigorous theory for the issue related to (2.11) in a three-dimensional domain. In the two-dimensional case, the validity of condition (2.11), under suitably chosen boundary data, is proved in [2] when σ is smooth. When σ is just measurable, (2.11) holds in the a.e. sense [1]. In the three-dimensional case, there are examples [3, 11] which suggest that it may be difficult, if not impossible, to find boundary data such that (2.11) holds independently of σ, even assuming smoothness of σ.

In this paper, we will consider the convergence result for the harmonic B_z method based on the governing problem (2.3), since the solution u to the standard governing problem (2.2) can be expressed as a constant multiple in terms of Lemma 2.1. Let σ^* be the target conductivity to be determined. Based on Lemma 2.4, the harmonic B_z algorithm approximating σ^* at each slice Ω_{z_0}, for given σ^* on $\partial\Omega_{z_0}$, can be stated as follows. Notice here that we also use σ^* to represent the known boundary value of unknown conductivity σ^* defined in the whole domain, which can be distinguished from the context. Given an initial guess $\sigma^0(x,y,z)$ in Ω with exact boundary values, the harmonic B_z iteration algorithm constructs an approximation sequence $\{\sigma^n(x,y,z_0)\}$ by

$$\begin{cases} \nabla\sigma^{n+1}(x,y,z_0) := \dfrac{1}{\mu_0}\mathbb{A}[\sigma^n]^{-1}\begin{bmatrix} \Delta B_z^1 \\ \Delta B_z^2 \end{bmatrix}, \\ \sigma^{n+1}(x,y,z_0) := H(\sigma^*) - \dfrac{1}{2\pi}\displaystyle\int_{\Omega_{z_0}} \frac{(x-x',\, y-y')}{|x-x'|^2+|y-y'|^2}\cdot\nabla\sigma^{n+1}(x',y',z_0)dx'dy' \end{cases} \tag{2.16}$$

for $n = 0, 1, 2, \ldots$ at each slice Ω_{z_0}, where $\mathbb{A}[\sigma]$ is defined in (2.12) and

$$H(\sigma^*) := \frac{1}{2\pi}\int_{\partial\Omega_{z_0}} \frac{(x-x',\, y-y')\,\cdot\,\nu(x',y')}{|x-x'|^2+|y-y'|^2}\,\sigma^*(x',y',z_0)\,dl.$$

Notice that, to compute $\mathbb{A}[\sigma^n]^{-1}$ at each slice $\Omega_{z_0} \subset \mathbb{R}^2$, we need the value σ^n in the whole domain $\Omega \subset \mathbb{R}^3$. This fact will cause some difficulties when we do the iteration for the full three-dimensional model (see the convergence proof in subsection 3.2 of this paper). On the other hand, in the above scheme, the value of σ^n on the boundary of the slice Ω_{z_0} is specified as the exact value $\sigma^*(x,y,z_0)$ for all n.

3. Convergence for the harmonic B_z algorithm. It should be noticed that $\mathbb{A}[\sigma^n]^{-1}(x,y,z)$ may be large near $\partial\Omega$ due to the fact that two induced current densities $\sigma^n\nabla u_1^n, \sigma^n\nabla u_2^n$ are probably almost parallel for some configuration. To avoid

this case, let us assume that σ^* is constant in $\Omega \setminus \tilde{\Omega}$ for some interior domain $\tilde{\Omega}$. Then it is easy to show that $\nabla^2 B_z^1 \equiv \nabla^2 B_z^2 \equiv 0$ in $\Omega \setminus \tilde{\Omega}$. So the original iteration scheme (2.16) at each slice becomes

(3.1)
$$\begin{cases} \nabla\sigma^{n+1}(x,y,z_0) := \dfrac{1}{\mu_0}\mathbb{A}[\sigma^n]^{-1}\begin{bmatrix} \Delta B_z^1 \\ \Delta B_z^2 \end{bmatrix}, \\ \sigma^{n+1}(x,y,z_0) := \tilde{H}(\sigma^*) - \dfrac{1}{2\pi}\displaystyle\int_{\tilde{\Omega}_{z_0}} \frac{(x-x',\, y-y')}{|x-x'|^2+|y-y'|^2} \cdot \nabla\sigma^{n+1}(x',y',z_0)dx'dy' \end{cases}$$

for $(x,y) \in \tilde{\Omega}_{z_0}$, where $\tilde{\Omega}_{z_0} = \tilde{\Omega} \bigcap \{(x,y,z) : z = z_0\} \subset \mathbb{R}^2$ and $\tilde{H}(\sigma^*)$ is $H(\sigma^*)$ with $\partial\Omega_{z_0}$ replaced by $\partial\tilde{\Omega}_{z_0}$. For $(x,y) \in \Omega_{z_0} \setminus \tilde{\Omega}_{z_0}$, it is obvious that $\sigma^n(x,y,z_0) \equiv \sigma^*(x,y,z_0)$ from this iteration since $\nabla\sigma^n \equiv 0$ in $\Omega_{z_0} \setminus \tilde{\Omega}_{z_0}$.

The major difficulty dealing with the convergence comes from the uniform upper bound of the inverse matrix $\mathbb{A}[\sigma^n]^{-1}$ in the iteration procedure. Without some uniform bound on the iteration conductivity σ^n, it is quite difficult to estimate $\mathbb{A}[\sigma^n]^{-1}$. The key ideas taken in this paper to overcome this difficulty are some assumptions on the target conductivity and the initial guess. With these conditions, we can establish the convergence. Rather than the general way of the convergence proof, which sets the uniform bound for the sequence $\{\sigma^n\}$ and then obtains the convergence of the sequence, we should establish the uniform bound on σ^n and estimate the error $\|\sigma^n - \sigma^*\|$ at each step simultaneously by the induction argument.

We first give the following estimates, which will be used in the convergence proof.

LEMMA 3.1. *Denote by $\mathcal{E}$ any regular open subsurface of the boundary $\partial\Omega$ of $\Omega \subset \mathbb{R}^m$, with $m = 2, 3$. Then for the boundary value problem*

$$\begin{cases} \nabla \cdot (\sigma\nabla u) = \nabla \cdot f & \text{in } \Omega, \\ u|_{\mathcal{E}} = h & \text{on } \mathcal{E}, \\ -\sigma\nabla u \cdot \mathbf{n} = g & \text{on } \partial\Omega \setminus \overline{\mathcal{E}}, \end{cases} \tag{3.2}$$

with $\sigma \in L^\infty(\Omega)$ satisfying $\inf_\Omega \sigma > 0$, $h \in H^{1/2}(\mathcal{E})$, and $g \in H^{-1/2}(\partial\Omega \setminus \overline{\mathcal{E}})$, the following estimates hold:

If $f \in (L^2(\Omega))^m$ and $\sigma \in C(\Omega)$, then $u \in H^1(\Omega)$ and

$$\|u\|_{H^1(\Omega)} \le C_1(\sigma)[\|f\|_{L^2(\Omega)} + \|h\|_{H^{1/2}(\mathcal{E})} + \|g\|_{H^{-1/2}(\partial\Omega\setminus\overline{\mathcal{E}})}]; \tag{3.3}$$

if $f \in (H^1(\Omega))^m$ and $\sigma \in C^1(\Omega)$, then $u \in H^2(\tilde{\Omega})$ and

$$\|u\|_{H^2(\tilde{\Omega})} \le C_2(\sigma)[\|u\|_{H^1(\Omega)} + \|\nabla \cdot f\|_{L^2(\Omega)}]; \tag{3.4}$$

if $f \in (C^{0,\alpha}(\Omega))^m$, with $\alpha \in (0,1)$ and $\sigma \in C^1(\Omega)$, then $u \in C^{1,\alpha}(\tilde{\tilde{\Omega}})$ and

$$\|\nabla u\|_{C^{0,\alpha}(\tilde{\Omega})} \le C_3(\sigma)[\|u\|_{C^{0,\alpha}(\tilde{\tilde{\Omega}})} + \|f\|_{C^{0,\alpha}(\tilde{\tilde{\Omega}})}]; \tag{3.5}$$

if $f \in (L^p(\Omega))^m$, with $p > 1$ and $\sigma \in C(\Omega)$, then $u \in W^{1,p}(\tilde{\tilde{\Omega}})$ and

$$\|\nabla u\|_{L^p(\tilde{\Omega})} \le C_4(\sigma)[\|u\|_{L^p(\tilde{\tilde{\Omega}})} + \|f\|_{L^p(\tilde{\tilde{\Omega}})}], \tag{3.6}$$

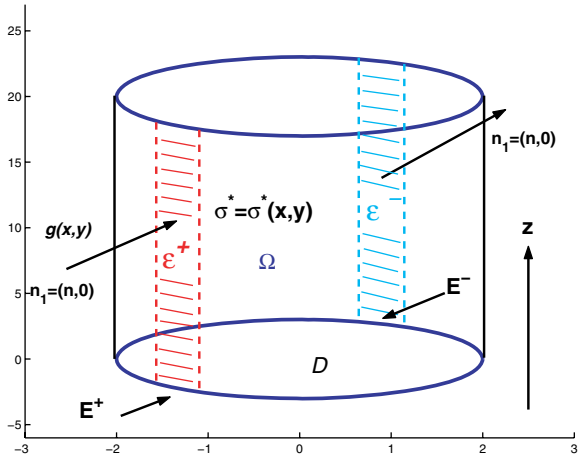


FIG. 3. *Axially symmetric cylindrical configuration.*

where $\tilde{\Omega} \subset\subset \tilde{\tilde{\Omega}} \subset\subset \Omega$ *are regular domains, and the functions* $C_i(\sigma)$ *have the following forms:*

$$C_i(\sigma) = F_i\left(\|\sigma\|_{C(\Omega)}, \|\nabla\sigma\|_{C(\Omega)}, \frac{1}{\inf_\Omega \sigma}\right), \quad i = 2, 3, \tag{3.7}$$

$$C_i(\sigma) = F_i\left(\|\sigma\|_{C(\Omega)}, \frac{1}{\inf_\Omega \sigma}\right), \quad i = 1, 4. \tag{3.8}$$

The functions F_i $(i = 1, 2, 3, 4)$ *are known bounded functions with respect to the arguments.*

For the proof of these estimates, we refer to [6] and Theorems 8.8 and 8.32 and Corollary 8.36 in [8]. The form of the constant $C_i(\sigma)$ is of importance in our convergence proof.

3.1. Convergence in axially symmetric cylindrical sections. Let $\Omega := D \times \mathbb{R}^1 \subset \mathbb{R}^3$ be a cylinder along the z direction with infinite length and the fixed cross section $D \subset \mathbb{R}^2$. We assume that the conductivity σ^* in Ω does not change along the z direction. That is, $\sigma^*(x, y, z) \equiv \sigma^*(x, y)$, $(x, y) \in D$.

Consider the electrodes $\mathcal{E}^{\pm}$ on $\partial\Omega$ where the input currents are specified. For an ideal electrode pair $\mathcal{E}^{\pm}$ parallel to the z direction with infinite length, we assume that the current density is independent of z; see Figure 3. In this case, it follows from (2.2) that the potential u is also independent of z in Ω due to the causality, and $u(x, y)$ meets

$$\begin{cases} \nabla \cdot (\sigma \nabla u) = 0 \quad \text{in } D, \\ \tilde{I} = \displaystyle\int_{E^+} \sigma \frac{\partial u}{\partial \mathbf{n}}\, ds = -\int_{E^-} \sigma \frac{\partial u}{\partial \mathbf{n}} ds, \quad \nabla u \times \mathbf{n} = 0 \ \text{ on } E^+ \cup E^-, \\ \sigma \dfrac{\partial u}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega \setminus \overline{E^+ \cup E^-}, \end{cases} \tag{3.9}$$

where $E^{\pm} := \mathcal{E}^{\pm} \bigcap \overline{D} \subset \partial D$, $\tilde{I}$ is the total input current in $E^{\pm}$, and $\mathbf{n} \in \mathbb{R}^2$ is the outward normal direction of ∂D. The equation (3.9) is the governing model for potential $u(x, y)$, which is in fact defined in the two-dimensional domain D.

To unify the notations in our proof for convergence in both the axially cylindrical case and the three-dimensional case, we will use $\Omega, \mathcal{E}^{\pm}$ to represent the geometry in (3.9) instead of $D, E^{\pm}$ in this subsection. Now we can state our convergence result of the harmonic B_z method based on the model (3.9) for the two-dimensional domain $\Omega \subset \mathbb{R}^2$.

THEOREM 3.2. *Assume that the target conductivity $\sigma^*(x,y) \in C^1(\overline{\Omega})$ meets the following:*

H1. $0 < \sigma_-^* \le \sigma^*(x) \le \sigma_+^*$ *for known constants* $\sigma_\pm^*$;

H2. *there exists* $\tilde{\Omega} \subset\subset \Omega$ *such that* σ^* *is a known constant in* $\Omega \setminus \tilde{\Omega}$;

H3. $|\det A[\sigma^*](x,y)| \ge d_-^* > 0$ *in* $\tilde{\Omega}$, *where* d_-^* *is a known constant.*

Under these hypotheses, there exist constants $\epsilon = \epsilon(\sigma_\pm^*, d_-^*) > 0$ small enough and $\theta = \theta(\epsilon, \sigma_\pm^*, d_-^*) \in (0,1)$ such that if we take the initial guess σ^0 as the constant $\sigma^*|_{\Omega\setminus\tilde{\Omega}}$, then the sequence $\{\sigma^n\}$ given by the harmonic B_z iteration scheme holds for $\|\nabla\sigma^*\|_{C(\tilde{\Omega})} \le \epsilon$ that

$$\sigma^n \equiv \sigma^* \text{ in } \Omega\setminus\tilde{\Omega}, \quad \|\sigma^n - \sigma^*\|_{C^1(\tilde{\Omega})} \le K\theta^n\epsilon, \quad n = 1,2,\ldots, \tag{3.10}$$

where $K := diam(\Omega) + 1$.

Remark 3.3. We obtain a local convergence for the target conductivity σ^* with a small contrast. At the present stage, we do not know how to remove the smallness requirement on ϵ.

Remark 3.4. The convergence property holds only in the interior domain $\tilde{\Omega}$. The reason is as follows. First, the regularity property of an elliptic equation with the mixed boundary condition will fail at the boundary. Second, for some geometry configuration, the induced internal current densities near the boundary should be almost parallel, so it is very hard to get the uniform bound on $\mathbb{A}[\sigma^n]^{-1}$ near the boundary.

Proof. Let us take $0 < \epsilon < \frac{1}{2K}\sigma_-^*$. Denote by u_j^n and u_j^* the solutions of the direct problem

$$\begin{cases} \nabla\cdot(\sigma\nabla u_j) = 0 \quad \text{in} \quad \Omega, \\ u_j|_{\mathcal{E}_j^+} = 1, \qquad u_j|_{\mathcal{E}_j^-} = 0, \\ -\sigma\nabla u_j \cdot \mathbf{n} = 0 \quad \text{on} \quad \partial\Omega \setminus \mathcal{E}_j^+ \cup \mathcal{E}_j^-, \end{cases} \tag{3.11}$$

with $\sigma = \sigma^n$ and σ^*, respectively. This is a special case of (3.2) with $\mathcal{E} := \mathcal{E}_j^+ \bigcup \mathcal{E}_j^-$ in Lemma 3.1. For a given interior domain $\tilde{\Omega}$, there exists a constant $\overline{C}_* = \overline{C}_*(\sigma_\pm^*)$ such that

$$\left\|\nabla u_j^*\right\|_{C(\tilde{\Omega})} + \left\|u_j^*\right\|_{H^2(\tilde{\Omega})} \le \overline{C}_*. \tag{3.12}$$

This fact can be deduced from Lemma 3.1 as follows. Indeed, from the first and second points of this Lemma, we have $\left\|u_j^*\right\|_{H^2(\tilde{\Omega})} \le C_1(\sigma^*)C_2(\sigma^*)$. By the Sobolev imbedding theorem, we have $\left\|u_j^*\right\|_{C^{0,\alpha}(\tilde{\Omega})} \le C_s \left\|u_j^*\right\|_{H^2(\tilde{\Omega})}$ for every $\alpha \in (0,1)$. Finally, combining these last two estimates with the third point of Lemma 3.1, we have

$$\left\|\nabla u_j^*\right\|_{C(\tilde{\Omega})} + \left\|u_j^*\right\|_{H^2(\tilde{\Omega})} \le C_s C_1(\sigma^*)C_2(\sigma^*)C_3(\sigma^*),$$

which leads to (3.12) with

$$\overline{C}_* := C_s \sup_{(t_1,t_2,t_3)\in\mathbb{S}_1} F_1(t_1,t_3)F_2(t_1,t_2,t_3)F_3(t_1,t_2,t_3), \tag{3.13}$$

where $\mathbb{S}_1 := [\sigma_-^*, \sigma_+^*] \times [0, \frac{1}{2K}\sigma_-^*] \times [\frac{1}{\sigma_+^*}, \frac{1}{\sigma_-^*}]$, and each F_i is uniformly bounded with respect to their arguments.

Step 1. Expand the initial guess σ^0 at σ^*.

Expand σ^0 as $\sigma^0 = \sigma^* + e^0$. Since $\|\nabla\sigma^*\|_{C(\Omega)} < \epsilon$ and $\sigma^* = \sigma^0$ in $\Omega \setminus \tilde{\Omega}$,

$$\left\|e^0\right\|_{C(\Omega)} \leq diam(\Omega)| \left\|\nabla e^0\right\|_{C(\Omega)} \leq diam(\Omega)\epsilon.$$

Hence, $\left\|e^0\right\|_{C^1(\Omega)} \leq (diam(\Omega)+1)\epsilon = K\epsilon$. We expand u_j^0 at u_j^* as

$$u_j^0 = u_j^* + \epsilon w_j^0. \tag{3.14}$$

Noticing that $\sigma^0 = \sigma^*$ in $\Omega \setminus \tilde{\Omega}$, ϵw_j^0 meets

$$\begin{cases} \nabla \cdot \left(\sigma^0 \nabla \epsilon w_j^0\right) = -\nabla \cdot (e^0 \nabla u_j^*) \quad \text{in} \quad \Omega, \\ \epsilon w_j^0|_{\mathcal{E}_j^+} = 0, \qquad \epsilon w_j^0|_{\mathcal{E}_j^-} = 0, \\ -\sigma^0 \nabla \epsilon w_j^0 \cdot \ \mathbf{n} = (\sigma^0 - \sigma^*)\nabla u_j^* \cdot \ \mathbf{n} = 0 \quad \text{on} \quad \partial\Omega \setminus \mathcal{E}_j^+ \cup \mathcal{E}_j^-. \end{cases} \tag{3.15}$$

Since $\|e^0\|_{C^1(\tilde{\Omega})} \leq K\epsilon$ and $e^0 = 0$ in $\Omega \setminus \tilde{\Omega}$, it follows from (3.12) that the right-hand side of the first equation in (3.15) satisfies

$$\left\|\nabla \cdot (e^0 \nabla u_j^*)\right\|_{L^2(\Omega)} \leq \overline{C}_* \left\|e^0\right\|_{C^1(\tilde{\Omega})}. \tag{3.16}$$

Therefore it follows from Lemma 3.1 and the Sobolev imbedding theorem that

$$\begin{aligned} \left\|\epsilon w_j^0\right\|_{C(\tilde{\tilde{\Omega}})} &\leq C_s \left\|\epsilon w_j^0\right\|_{H^2(\tilde{\tilde{\Omega}})} \leq C_s C_2(\sigma^0)[\left\|\epsilon w_j^0\right\|_{H^1(\tilde{\Omega})} + \left\|\nabla \cdot e^0 \nabla u_j^*\right\|_{L^2(\tilde{\Omega})}] \\ &\leq C_s C_2(\sigma^0)[C_1(\sigma^0)\left\|e^0 \nabla u_j^*\right\|_{L^2(\Omega)} + \overline{C}_* \left\|e^0\right\|_{C^1(\Omega)}] \\ &\leq C_s C_2(\sigma^0)[C_1(\sigma^0)C_1(\sigma^*) + \overline{C}_*] \left\|e^0\right\|_{C^1(\Omega)}. \end{aligned}$$

Hence

$$\begin{aligned} \left\|\epsilon \nabla w_j^0\right\|_{C(\tilde{\Omega})} &\leq C_3(\sigma^0)[\left\|\epsilon w_j^0\right\|_{C(\tilde{\tilde{\Omega}})} + \left\|e^0 \nabla u_j^*\right\|_{C(\tilde{\Omega})}] \\ &\leq C_3(\sigma^0)[C_s C_2(\sigma^0)[C_1(\sigma^0)C_1(\sigma^*) + \overline{C}_*] + \overline{C}_*] \left\|e^0\right\|_{C^1(\Omega)}. \end{aligned} \tag{3.17}$$

We denote by

$$\overline{F}(\sigma) := C_3(\sigma)\left[C_s C_2(\sigma)\left(C_1(\sigma) \sup_{[\sigma_-^*, \sigma_+^*] \times [\frac{1}{\sigma_+^*}, \frac{1}{\sigma_-^*}]} F_1(t_1, t_3) + \overline{C}_*\right) + \overline{C}_*\right] \tag{3.18}$$

a known function due to the Lemma 3.1. For $\epsilon \in (0, \frac{1}{2K}\sigma_-^*)$, we introduce the constant

$$\overline{C}_\epsilon(\sigma^*) := \sup_{\|\sigma - \sigma^*\|_{C^1(\Omega)} \leq K\epsilon} \overline{F}(\sigma), \tag{3.19}$$

which is well defined. Noticing that $\|\sigma - \sigma^*\|_{C^1(\Omega)} \leq K\epsilon$, we have $\sigma > \frac{1}{2}\sigma_-^* > 0$ for $0 < \epsilon < \frac{1}{2K}\sigma_-^*$ due to H1. Moreover, this constant can be estimated by a known constant as

$$\overline{C}_\epsilon(\sigma^*) \leq \sup_{\mathbb{S}_2} \overline{F}(\sigma) =: \overline{G}(\sigma_\pm^*) \tag{3.20}$$

for $\epsilon \in (0, \frac{1}{2K}\sigma_-^*)$ from the a priori information about σ^*, where

$$\mathbb{S}_2 := \left\{ \sigma(x,y) : \frac{1}{2}\sigma_-^* < \sigma < \frac{1}{2}\sigma_-^* + \sigma_+^*,\ \|\nabla\sigma\|_C \le \frac{K+1}{2K}\sigma_-^* \right\}.$$

Now it follows from (3.17)–(3.20) that

$$\left\|\epsilon\nabla w_j^0\right\|_{C(\tilde{\Omega})} \le \overline{G}(\sigma_\pm^*)\left\|e^0\right\|_{C^1(\Omega)}. \tag{3.21}$$

Since $\|e^0\|_{C^1(\tilde{\Omega})} \le K\epsilon$, it follows that

$$\left\|\nabla w_j^0\right\|_{C(\tilde{\Omega})} \le K\overline{G}(\sigma_\pm^*). \tag{3.22}$$

Step 2. Estimate $\left\|\sigma^1 - \sigma^*\right\|_{C^1(\tilde{\Omega})}$.

First, $\mathbb{A}[\sigma^*]^{-1}$ exists from H3. From the definition, we know that $\nabla\sigma^1$ satisfies

$$\mathbb{A}[\sigma^0]\nabla\sigma^1 = \frac{1}{\mu_0}\begin{bmatrix} \Delta B_z^1 \\ \Delta B_z^2 \end{bmatrix},$$

which can be written as

$$\left(I + \epsilon\mathbb{A}[\sigma^*]^{-1}\begin{bmatrix} \frac{\partial w_1^0}{\partial y}, & -\frac{\partial w_1^0}{\partial x} \\ \frac{\partial w_2^0}{\partial y}, & -\frac{\partial w_2^0}{\partial x} \end{bmatrix}\right)\nabla\sigma^1 = \frac{1}{\mu_0}\mathbb{A}[\sigma^*]^{-1}\begin{bmatrix} \Delta B_z^1 \\ \Delta B_z^2 \end{bmatrix} = \nabla\sigma^*$$

due to the definition of the matrix $\mathbb{A}[\sigma^0]$ and (3.14). So we have

$$\left(I + \epsilon\mathbb{A}[\sigma^*]^{-1}\begin{bmatrix} \frac{\partial w_1^0}{\partial y}, & -\frac{\partial w_1^0}{\partial x} \\ \frac{\partial w_2^0}{\partial y}, & -\frac{\partial w_2^0}{\partial x} \end{bmatrix}\right)\nabla(\sigma^1 - \sigma^*) = -\epsilon\mathbb{A}[\sigma^*]^{-1}\begin{bmatrix} \frac{\partial w_1^0}{\partial y}, & -\frac{\partial w_1^0}{\partial x} \\ \frac{\partial w_2^0}{\partial y}, & -\frac{\partial w_2^0}{\partial x} \end{bmatrix}\nabla\sigma^*. \tag{3.23}$$

On the other hand, it is obvious from (3.22) that

$$\begin{aligned}\left\|\epsilon\mathbb{A}[\sigma^*]^{-1}\begin{bmatrix} \frac{\partial w_1^0}{\partial y}, & -\frac{\partial w_1^0}{\partial x} \\ \frac{\partial w_2^0}{\partial y}, & -\frac{\partial w_2^0}{\partial x} \end{bmatrix}\right\|_{C(\tilde{\Omega})} &\le \epsilon\left\|\mathbb{A}[\sigma^*]^{-1}\right\|_{C(\tilde{\Omega})}\sqrt{2}\max_{j=1,2}\left\|\nabla w_j^0\right\|_{C(\tilde{\Omega})} \\ &\le \epsilon\left\|\mathbb{A}[\sigma^*]^{-1}\right\|_{C(\tilde{\Omega})}\sqrt{2\overline{G}}(\sigma_\pm^*)K. \end{aligned} \tag{3.24}$$

A direct computation leads to $\left\|\mathbb{A}[\sigma^*]^{-1}\right\|_{C(\tilde{\Omega})} \le \frac{\sqrt{2}}{d_-^*}\max_{j=1,2}\|\nabla u_j(\sigma^*)\|_{C(\tilde{\Omega})}$, from which we deduce

$$\left\|\mathbb{A}[\sigma^*]^{-1}\right\|_{C(\tilde{\Omega})} \le \frac{\sqrt{2}}{d_-^*}\overline{C}_* \tag{3.25}$$

due to (3.12). Now we take $\epsilon \in (0, \frac{1}{2K}\sigma_-^*)$ small enough such that

$$\epsilon\frac{2}{d_-^*}\overline{G}(\sigma_\pm^*)\overline{C}_*K < \frac{1}{2}; \tag{3.26}$$

then it follows from (3.25) that

$$\epsilon\left\|\mathbb{A}[\sigma^*]^{-1}\right\|_{C(\tilde{\Omega})}\sqrt{2\overline{G}}(\sigma_\pm^*)K < \frac{1}{2}. \tag{3.27}$$

It follows from (3.22), (3.23), and (3.27) that

$$\begin{aligned}\left\|\nabla(\sigma^1-\sigma^*)\right\|_{C(\tilde{\Omega})} &\le 2\epsilon \left\|\mathbb{A}[\sigma^*]^{-1}\right\|_{C(\tilde{\Omega})} \sqrt{2} \max_{j=1,2} \left\|\nabla w_j^0\right\| \|\nabla\sigma^*\|_{C(\tilde{\Omega})} \\ &\le 2\sqrt{2} \left\|\mathbb{A}[\sigma^*]^{-1}\right\|_{C(\tilde{\Omega})} \overline{G}(\sigma_\pm^*) K\epsilon^2. \end{aligned} \tag{3.28}$$

This last estimate generates

$$\left\|\sigma^1-\sigma^*\right\|_{C^1(\tilde{\Omega})} \le K \left\|\nabla(\sigma^1-\sigma^*)\right\|_{C(\tilde{\Omega})} \le K2\sqrt{2} \left\|\mathbb{A}[\sigma^*]^{-1}\right\|_{C(\tilde{\Omega})} \overline{G}(\sigma_\pm^*) K\epsilon^2.$$

Introducing a new constant

$$\overline{D}_* := K\frac{4}{d_-^*}\overline{C}_*, \tag{3.29}$$

the above estimate becomes

$$\left\|\sigma^1-\sigma^*\right\|_{C^1(\tilde{\Omega})} \le \overline{D}_*\overline{G}(\sigma_\pm^*) K\epsilon^2 \tag{3.30}$$

due to (3.25). For $\epsilon \in (0, \frac{1}{2K}\sigma_-^*)$ satisfying (3.26), it follows from the definition of $\overline{D}_*$ that

$$\epsilon\overline{D}_*\overline{G}(\sigma_\pm^*) := \theta \in (0,1) \tag{3.31}$$

and then

$$\left\|\sigma^1-\sigma^*\right\|_{C^1(\tilde{\Omega})} \le K\theta\epsilon, \tag{3.32}$$

which implies via (3.19) that

$$\overline{F}(\sigma^1) \le \overline{C}_\epsilon(\sigma^*). \tag{3.33}$$

From (3.26) and (3.31), we assert that $\epsilon := \epsilon(\sigma_\pm^*, d_-^*)$ and $\theta := \theta(\epsilon, \sigma_\pm^*, d_-^*)$. Since σ^* is constant in $\Omega \setminus \tilde{\Omega}$, we deduce from (2.14) that $\nabla^2 B_z^j = 0$ in $\Omega \setminus \tilde{\Omega}$, $j = 1, 2$. Hence from (2.16) we deduce that $\nabla\sigma^1 = 0$ in $\Omega \setminus \tilde{\Omega}$, and therefore $\sigma^1 = \sigma^*$ in $\Omega \setminus \tilde{\Omega}$, since $\sigma^1 = \sigma^*$ on $\partial\Omega$. Now we can apply the induction argument to prove the theorem. That is, assume that the following properties:

$$\left\|\sigma^k-\sigma^*\right\|_{C^1(\tilde{\Omega})} \le K\theta^k\epsilon \text{ and } \sigma^k = \sigma^* \text{ in } \Omega\setminus\tilde{\Omega} \tag{3.34}$$

are true for $k = n$. We shall prove that it is also true for $k = n + 1$.

Step 3. Expand σ^n at σ^*.

For the expansion $\sigma^n = \sigma^* + e^n$, it follows that

$$\|e^n\|_{C^1(\tilde{\Omega})} \le (\overline{D}_*\overline{G}(\sigma_\pm^*))^n K\epsilon^{n+1} \tag{3.35}$$

from (3.34) with $k = n$. Correspondingly, we expand the solution u_j^n at u_j^* as

$$u_j^n = u_j^* + \epsilon^{n+1} w_j^n. \tag{3.36}$$

Noticing $\sigma^n = \sigma^*$ in $\Omega \setminus \tilde{\Omega}$, it is easy to see that $\epsilon^{n+1}w_j^n$ satisfies

$$\begin{cases} \nabla\cdot\left(\sigma^n\nabla\epsilon^{n+1}w_j^n\right) = -\nabla\cdot(e^n\nabla u_j^*) \quad \text{in} \quad \Omega, \\ \epsilon^{n+1}w_j^n|_{\mathcal{E}_j^+} = 0, \qquad \epsilon^{n+1}w_j^n|_{\mathcal{E}_j^-} = 0, \\ -\sigma^n\nabla\epsilon^{n+1}w_j^n\cdot\ \mathbf{n} = (\sigma^n-\sigma^*)\nabla u_j^*\cdot\ \mathbf{n} = 0 \quad \text{on} \quad \partial\Omega\setminus\mathcal{E}_j^+\cup\mathcal{E}_j^-. \end{cases} \tag{3.37}$$

Similarly as in (3.21), we have

$$\left\|\nabla \epsilon^{n+1} w_j^n\right\|_{C(\tilde{\Omega})} \le \overline{F}(\sigma^n) \left\|e^n\right\|_{C^1(\Omega)} \le \overline{F}(\sigma^n)(\overline{D}_* \overline{G}(\sigma_\pm^*))^n K \epsilon^{n+1}.$$

That is,

$$\left\|\nabla w_j^n\right\|_{C(\tilde{\Omega})} \le \overline{F}(\sigma^n)(\overline{D}_* \overline{G}(\sigma_\pm^*))^n K \le \overline{D}_*^n (\overline{G}(\sigma_\pm^*))^{n+1} K, \tag{3.38}$$

since $\overline{F}(\sigma^n) \le \overline{C}_\epsilon(\sigma^*) \le \overline{G}(\sigma_\pm^*)$ due to the fact that $\|\sigma^n - \sigma^*\|_{C^1(\tilde{\Omega})} \le K\theta^n \epsilon < K\epsilon$.

Step 4. Estimate $\left\|\sigma^{n+1} - \sigma^*\right\|_{C^1}$.

From (2.16) we get $\nabla \sigma^{n+1} = 0$ in $\Omega \setminus \tilde{\Omega}$ and then $\sigma^{n+1} = \sigma^*$ in $\Omega \setminus \tilde{\Omega}$. By the same argument as in Step 2, we have

$$\left(I + \epsilon^{n+1} \mathbb{A}[\sigma^*]^{-1} \begin{bmatrix} \frac{\partial w_1^n}{\partial y} & -\frac{\partial w_1^n}{\partial x} \\ \frac{\partial w_2^n}{\partial y} & -\frac{\partial w_2^n}{\partial x} \end{bmatrix}\right) \nabla(\sigma^{n+1} - \sigma^*) = -\epsilon^{n+1} \mathbb{A}[\sigma^*]^{-1} \begin{bmatrix} \frac{\partial w_1^n}{\partial y} & -\frac{\partial w_1^n}{\partial x} \\ \frac{\partial w_2^n}{\partial y} & -\frac{\partial w_2^n}{\partial x} \end{bmatrix} \nabla \sigma^*. \tag{3.39}$$

With the condition (3.26) for ϵ, we have

$$\begin{aligned} \left\| \epsilon^{n+1} \mathbb{A}[\sigma^*]^{-1} \begin{bmatrix} \frac{\partial w_1^n}{\partial y}, & -\frac{\partial w_1^n}{\partial x} \\ \frac{\partial w_2^n}{\partial y}, & -\frac{\partial w_2^n}{\partial x} \end{bmatrix} \right\|_{C(\tilde{\Omega})} &\le \epsilon^{n+1} \left\|\mathbb{A}[\sigma^*]^{-1}\right\|_{C(\tilde{\Omega})} \sqrt{2} \max_{j=1,2} \left\|\nabla w_j^n\right\|_{C(\tilde{\Omega})} \\ &\le \epsilon^{n+1} \left\|\mathbb{A}[\sigma^*]^{-1}\right\|_{C(\Omega)} \sqrt{2} \overline{D}_*^n (\overline{G}(\sigma_\pm^*))^{n+1} K \\ &\le \frac{1}{2} (\epsilon \overline{D}_* \overline{G}(\sigma_\pm^*))^n < \frac{1}{2} \end{aligned} \tag{3.40}$$

due to (3.26), (3.31), and (3.38). So it follows from (3.38)–(3.40) that

$$\begin{aligned} \left\|\nabla(\sigma^{n+1} - \sigma^*)\right\|_{C(\tilde{\Omega})} &\le 2\epsilon^{n+1} \left\|\mathbb{A}[\sigma^*]^{-1}\right\|_{C(\tilde{\Omega})} \sqrt{2} \max_{j=1,2} \left\|\nabla w_j^n\right\|_{C(\tilde{\Omega})} \|\nabla \sigma^*\|_{C(\tilde{\Omega})} \\ &\le 2\sqrt{2} K \epsilon^{n+1} \left\|\mathbb{A}[\sigma^*]^{-1}\right\|_{C(\tilde{\Omega})} \overline{D}_*^n (\overline{G}(\sigma_\pm^*))^{n+1} \epsilon. \end{aligned} \tag{3.41}$$

This estimate together with $\left\|\sigma^{n+1} - \sigma^*\right\|_{C^1(\Omega)} \le K \left\|\nabla \sigma^{n+1} - \nabla \sigma^*\right\|_{C(\Omega)}$ generates

$$\left\|\sigma^{n+1} - \sigma^*\right\|_{C^1(\tilde{\Omega})} \le K2 \left\|\mathbb{A}^{-1}[\sigma^*]\right\|_{C(\tilde{\Omega})} \sqrt{2} \overline{D}_*^n (\overline{G}(\sigma_\pm^*))^{n+1} \epsilon^{n+1} K\epsilon \le (\overline{D}_* \overline{G}(\sigma_\pm^*)\epsilon)^{n+1} K\epsilon$$

from (3.25) and (3.29). Now under (3.31) for ϵ, the above estimate leads to

$$\left\|\sigma^{n+1} - \sigma^*\right\|_{C^1(\tilde{\Omega})} \le \theta^{n+1} K\epsilon. \tag{3.42}$$

It is obvious that $\sigma^{n+1} = \sigma^*$ in $\Omega \setminus \tilde{\Omega}$. So (3.34) is also true for $k = n+1$. The proof is complete. ∎

3.2. Convergence for a three-dimensional medium. Now we consider the convergence in the three-dimensional case. The essence of the proof is almost the same as that in the two-dimensional case, but we need some modifications due to the fact that the harmonic B_z algorithm computes a conductivity distribution in each two-dimensional slice of the three-dimensional medium. More explanations are given in Remark 3.6.

THEOREM 3.5. *Assume the target conductivity $\sigma^* \in C^1(\overline{\Omega})$, with $\Omega \subset \mathbb{R}^3$, which satisfies the following conditions:*

A1. $0 < \sigma_-^* \le \sigma^* \le \sigma_+^*$ *with known constants* $\sigma_\pm^*$*;*

A2. *there exists* $\tilde{\Omega} \subset\subset \Omega$ *such that* σ^* *is a known constant in* $\Omega \setminus \tilde{\Omega}$*;*

A3. $|\det \mathbb{A}[\sigma^*](x,y,z)| \ge d_-^* > 0$ *in* Ω*, where* d_-^* *is a known constant.*

Under these hypotheses, there exist constants $\epsilon = \epsilon(\sigma_\pm^*, d_-^*) > 0$ *small enough and* $\theta = \theta(\epsilon, \sigma_\pm^*, d_-^*) \in (0,1)$ *such that if we take the initial guess* σ^0 *as the constant* $\sigma^*|_{\Omega\setminus\tilde{\Omega}}$*, then it holds that the sequence* $\{\sigma^n := \sigma^n(x,y,z_0)$ *for all* $z_0\}$ *defined in* Ω*, where* $\sigma^n(x,y,z_0)$ *is constructed by the harmonic* B_z *iteration* (2.16) *for every* z_0*, converges to the true conductivity* σ^* *in* Ω *for* σ^* *satisfying*

$$\|\nabla \sigma^*\|_{C(\tilde{\Omega})} \le \epsilon. \tag{3.43}$$

More precisely, it holds that

$$\sigma^n = \sigma^* \text{ in } \Omega \setminus \tilde{\Omega}, \tag{3.44}$$

$$\|\sigma^n - \sigma^*\|_{C(\tilde{\Omega})} \le K\theta^n \epsilon, \quad \|\nabla_{x,y}(\sigma^n - \sigma^*)\|_{C(\tilde{\Omega})} \le K\theta^n \epsilon, \quad n = 1, 2, \dots, \tag{3.45}$$

where $K := diam(\Omega) + 1$.

Remark 3.6. In this three-dimensional setting, the estimate (3.45) is given by the C-norm, while the one in the two-dimensional case in Theorem 3.2 is given by the C^1-norm. We cannot improve the derivative estimate $\nabla_{x,y}(\sigma^n - \sigma^*)$ in (3.45) by $\nabla(\sigma^n - \sigma^*)$ since we do not know $\partial_z(\sigma^n - \sigma^*)$, although we have the full three gradient estimates (3.43) for σ^*. The main difficulty in this case is due to the fact that, in the iteration process (2.16), we get $\nabla_{x,y}\sigma^{n+1}$ at each slice with no information about $\partial_z \sigma^{n+1}$. That is, the harmonic B_z method approximates the three-dimensional conductivity function $\sigma^*(x,y,z)$ in $\Omega \subset \mathbb{R}^3$ at each two-dimensional slice $\Omega_{z_0} := \Omega \bigcap \{(x,y,z) : z = z_0\} \subset \mathbb{R}^2$. Then σ^n in $\Omega \subset \mathbb{R}^3$, at each iteration, is constructed as $\sigma^n := \bigcup_{z_0} \sigma^n(x,y,z_0)$.

Proof. We also take $\epsilon \in (0, \frac{1}{2K}\sigma_-^*)$. Similarly to the two-dimensional case, we denote by u_j^n and u_j^* the solutions to

$$\begin{cases} \nabla \cdot (\sigma \nabla u_j) = 0 \quad \text{in} \quad \Omega, \\ u_j|_{\mathcal{E}_j^+} = 1, \qquad u_j|_{\mathcal{E}_j^-} = 0, \\ -\sigma \nabla u_j \cdot \mathbf{n} = 0 \quad \text{on} \quad \partial\Omega \setminus \mathcal{E}_j^+ \cup \mathcal{E}_j^-, \end{cases} \tag{3.46}$$

with $\sigma = \sigma^n$ and σ^*, respectively, and we write σ^0 as $\sigma^0 = \sigma^* + e^0$.

In every slice Ω_{z_0}, we have

$$|e^0(x,y,z_0)| \le diam(\Omega_{z_0}) \left\|\nabla_{x,y} e^0\right\|_{C(\Omega_{z_0})} \le diam(\Omega)\epsilon;$$

hence $\left\|e^0\right\|_{C(\tilde{\Omega})} \le K\epsilon$. Correspondingly, we expand u_j^0 at u_j^* as

$$u_j^0 = u_j^* + \epsilon w_j^0. \tag{3.47}$$

Hence ϵw_j^0 satisfies

$$\begin{cases} \nabla \cdot \left(\sigma^0 \nabla \epsilon w_j^0\right) = -\nabla \cdot (e^0 \nabla u_j^*) \quad \text{in} \quad \Omega, \\ \epsilon w_j^0|_{\mathcal{E}_j^+} = 0, \qquad \epsilon w_j^0|_{\mathcal{E}_j^-} = 0, \\ -\sigma^0 \nabla \epsilon w_j^0 \cdot \mathbf{n} = (\sigma^0 - \sigma^*)\nabla u_j^* \cdot \mathbf{n} = 0 \quad \text{on} \quad \partial\Omega \setminus \mathcal{E}_j^+ \cup \mathcal{E}_j^-. \end{cases} \tag{3.48}$$

In the two-dimensional case, we can estimate the L^2-norm of $(\nabla \cdot e^n \nabla u_j^*)$ by (3.16). This is due to the fact that we can estimate $\|\nabla e^n\|_{C(\Omega)}$. But in the three-dimensional case, it is impossible to estimate $\|\partial_z e^n\|_{C(\Omega)}$. To overcome this difficulty, instead of using the L^2 and the Holder estimates of elliptic problems, we use the L^p estimate with $p > 1$.

First, by applying Lemma 3.1 to (3.46) with $\sigma = \sigma^*$ and the Sobolev imbedding theorem, we deduce that

$$\left\|u_j^*\right\|_{H^2(\tilde{\tilde{\Omega}})} + \left\|\nabla u_j^*\right\|_{C(\tilde{\Omega})} \leq C_* = C_*(\sigma_\pm^*) \tag{3.49}$$

due to (3.43) for $\epsilon \in (0, \frac{1}{2K}\sigma_-^*)$, where $\tilde{\Omega} \subset \tilde{\tilde{\Omega}} \subset \Omega$ and C_* in this three-dimensional case is constructed in the same way as constant $\overline{C}_*$ in (3.13), which implies that the right-hand side of the equation in (3.48) satisfies

$$\|e^0 \nabla u_j^*\|_{L^p(\Omega)} \leq C_* \left\|e^0\right\|_{C(\tilde{\Omega})} \quad \forall p > 1 \tag{3.50}$$

due to $\sigma^0 = \sigma^1$ on $\Omega \setminus \tilde{\Omega}$. The L^p interior estimates of the problem (3.48) give

$$\|\nabla \epsilon w_j^0\|_{L^p(\tilde{\tilde{\Omega}})} \leq C_4(\sigma^0)[\|\epsilon w_j^0\|_{L^p(\Omega)} + \|e^0 \nabla u_j^*\|_{L^p(\Omega)}]. \tag{3.51}$$

Again by the Sobolev imbedding theorem $H^1(\Omega) \subset L^p(\Omega)$, with $1 < p \leq 6$, we have

$$\|\epsilon w_j^0\|_{L^p(\Omega)} \leq C_s \|\epsilon w_j^0\|_{H^1(\Omega)} \leq C_s C_1(\sigma^0) \|e^0 \nabla u_j^*\|_{L^2(\Omega)} \leq C_s C_1(\sigma^0) C_* \left\|e^0\right\|_{C(\tilde{\Omega})}.$$

Hence combining this last estimate with (3.49) and (3.51) gives

$$\|\nabla \epsilon w_j^0\|_{L^p(\tilde{\tilde{\Omega}})} \leq C_4(\sigma^0)[C_s C_4(\sigma^0) C_* + C_*] \left\|e^0\right\|_{C(\tilde{\Omega})}. \tag{3.52}$$

Using Nirenberg's difference quotient method with respect to x in (3.48) in $\tilde{\tilde{\Omega}}$ yields

$$\nabla \cdot \sigma^0 \nabla \epsilon D_{x,h} w_j^0 = -\nabla \cdot e^0 \nabla D_{x,h} u_j^* - \nabla \cdot D_{x,h} e^0 \nabla u_{j,x,h}^* - \nabla \cdot D_{x,h} \sigma^0 \nabla \epsilon w_{j,x,h}^0, \tag{3.53}$$

where $D_{x,h} u := \frac{u(x+h,y,z) - u(x,y,z)}{h}$, $u_{x,h}(x,y,z) := u(x+h,y,z)$, with $h < dist(\partial\Omega, \tilde{\tilde{\Omega}})$.

The term in the right-hand side of (3.53) satisfies

$$\begin{cases} \|e^0 \nabla D_{x,h} u_j^*\|_{L^p(\tilde{\tilde{\Omega}})} \leq (1+\epsilon) C_4(\sigma^*) C_* \left\|e^0\right\|_{C(\tilde{\Omega})}, \\ \|D_{x,h} e^0 \nabla u_{j,x,h}^*\|_{L^p(\tilde{\tilde{\Omega}})} \leq C_* \left\|D_{x,h} e^0\right\|_{C(\tilde{\Omega})}, \\ \|D_{x,h} \sigma^0 \nabla \epsilon w_{j,x,h}^0\|_{L^p(\tilde{\tilde{\Omega}})} \leq \left\|D_{x,h} \sigma^0\right\|_{C(\tilde{\tilde{\Omega}})} \|\nabla \epsilon w_j^0\|_{L^p(\tilde{\tilde{\Omega}})}. \end{cases} \tag{3.54}$$

Indeed, $D_{x,h} u_j^*$ satisfies $\nabla \cdot \sigma^* \nabla (D_{x,h} u_j^*) = -\nabla \cdot (D_{x,h} \sigma^*) \nabla u_{j,x,h}^*$ in $\tilde{\tilde{\Omega}}$ from (3.46). Applying point 4 of Lemma 3.1, we deduce

$$\|\nabla (D_{x,h} u_j^*)\|_{L^p(\tilde{\Omega})} \leq C_4(\sigma^*)[\|D_{x,h} u_j^*\|_{L^p(\tilde{\tilde{\Omega}})} + \|(D_{x,h} \sigma^*) u_{j,x,h}^*\|_{L^p(\tilde{\tilde{\Omega}})}].$$

The estimate $\|D_{x,h} u_j^*\|_{L^p(\tilde{\tilde{\Omega}})} \leq \|\nabla u_j^*\|_{L^p(\tilde{\tilde{\Omega}})}$ and (3.50) give

$$\|\nabla (D_{x,h} u_j^*)\|_{L^p(\tilde{\Omega})} \leq (1+\epsilon) C_4(\sigma^*) C_*,$$

which is the first estimate in (3.54). The second term in (3.54) comes from (3.49).

Again from the interior L^p estimates applied for (3.53), we deduce that $D_{x,h}w_j^0$ is in $W^{1,p}_{loc}(\tilde{\Omega})$ and

$$\begin{aligned}\left\|\epsilon D_{x,h}w_j^0\right\|_{W^{1,p}(\tilde{\Omega})} \leq (1+C_4(\sigma^0))[&\left\|\epsilon D_{x,h}w_j^0\right\|_{L^p(\tilde{\tilde{\Omega}})} + \left\|e^0\nabla D_{x,h}u_j^*\right\|_{L^p(\tilde{\tilde{\Omega}})} \\ &+ \left\|D_{x,h}e^0\nabla u^*_{j,x,h}\right\|_{L^p(\tilde{\tilde{\Omega}})} + \left\|D_{x,h}\sigma^0\nabla\epsilon w^0_{j,x,h}\right\|_{L^p(\tilde{\tilde{\Omega}})}].\end{aligned}$$

(3.55)

Notice that $\left\|\epsilon D_{x,h}w_j^0\right\|_{L^p(\tilde{\tilde{\Omega}})} \leq \left\|\nabla\epsilon w_j^0\right\|_{L^p(\tilde{\tilde{\Omega}})}$ and

$$\left\|D_{x,h}\sigma^0\right\|_{C(\tilde{\tilde{\Omega}})} \leq \left\|\partial_x\sigma^0\right\|_{C(\tilde{\tilde{\Omega}})} \leq \left\|\nabla_{x,y}\sigma^0\right\|_{C(\tilde{\Omega})}. \tag{3.56}$$

The estimate (3.56) is trivial for σ^0 since it is a constant. However, we need this kind of estimate for the iterated sequence $\{\sigma^n\}$ with $\nabla_{x,y}\sigma^n$ continuous in $\tilde{\tilde{\Omega}}$. Hence the estimate (3.55) generates from (3.52), (3.54), and (3.56) that

$$\|\epsilon D_{x,h}w_j^0\|_{W^{1,p}(\tilde{\Omega})} \leq G(\sigma^0)[\left\|e^0\right\|_{C(\tilde{\Omega})} + \left\|D_{x,h}e^0\right\|_{C(\tilde{\Omega})}]$$

for $\epsilon \in (0, \frac{1}{2K}\sigma^*_-)$, where

$$G(\sigma) := (1+C_4(\sigma))C_* \times \max\left\{[1+\|\nabla_{x,y}\sigma\|_{C(\tilde{\Omega})}]C_4(\sigma)[C_sC_4(\sigma)+1] + \left(1+\frac{\sigma^*_-}{2K}\right)\sup_{[\sigma^*_-,\sigma^*_+]\times[\frac{1}{\sigma^*_+},\frac{1}{\sigma^*_-}]} F_4(t_1,t_3), 1\right\}. \tag{3.57}$$

Similarly, we know that $D_{y,h}w_j^0 \in W^{1,p}_{loc}(\tilde{\Omega})$ and also have the estimate

$$\|\epsilon D_{y,h}w_j^0\|_{W^{1,p}(\tilde{\Omega})} \leq G(\sigma^0)[\left\|e^0\right\|_{C(\tilde{\Omega})} + \left\|D_{y,h}e^0\right\|_{C(\tilde{\Omega})}].$$

Taking the limit with respect to $h \to 0$, we deduce that $\partial_x w_j^0,\ \partial_y w_j^0 \in W^{1,p}_{loc}(\tilde{\Omega})$ and

$$\|\epsilon\partial_x w_j^0\|_{W^{1,p}(\tilde{\Omega})},\ \ \|\epsilon\partial_y w_j^0\|_{W^{1,p}(\tilde{\Omega})} \leq G(\sigma^0)[\left\|e^0\right\|_{C(\tilde{\Omega})} + \left\|\nabla_{x,y}e^0\right\|_{C(\tilde{\Omega})}],$$

from which the Sobolev imbedding theorem $W^{1,p}(\tilde{\Omega}) \subset C(\tilde{\Omega})$ for $p>3$ implies

$$\|\epsilon\partial_x w_j^0\|_{C(\tilde{\Omega})},\ \ \|\epsilon\partial_y w_j^0\|_{C(\tilde{\Omega})} \leq C_sG(\sigma^0)[\left\|e^0\right\|_{C(\tilde{\Omega})} + \left\|\nabla_{x,y}e^0\right\|_{C(\tilde{\Omega})}].$$

We set $F(\sigma) := C_sG(\sigma)$, with $G(\sigma)$ defined in (3.57) and the constant $C_\epsilon(\sigma^*) := \sup_{\mathbb{S}_3} F(\sigma)$, with

$$\mathbb{S}_3 := \{\sigma(x,y,z) : \|\sigma-\sigma^*\|_{C(\Omega)} \leq K\epsilon,\ \|\nabla_{x,y}\sigma\|_{C(\tilde{\Omega})} \leq (K+1)\epsilon\}.$$

Again, noticing that $F(\sigma)$ contains only F_1 and F_4, the constant $C_\epsilon(\sigma^*)$ can be estimated by

$$C_\epsilon(\sigma^*) \leq \sup_{\mathbb{S}_4} F(\sigma) =: G(\sigma^*_\pm) \tag{3.58}$$

for $\epsilon \in (0, \frac{1}{2K}\sigma^*_-)$, where

$$\mathbb{S}_4 := \left\{\sigma(x,y,z) : \frac{1}{2}\sigma^*_- < \sigma < \frac{1}{2}\sigma^*_- + \sigma^*_+,\ \|\nabla_{x,y}\sigma\|_{C(\tilde{\Omega})} \leq \frac{K+1}{2K}\sigma^*_-\right\}.$$

In particular, we have $C_s G(\sigma^0) \le G(\sigma_\pm^*)$. Then the above estimate reads as

$$\|\epsilon\partial_x w_j^0\|_{C(\tilde\Omega)},\ \|\epsilon\partial_y w_j^0\|_{C(\tilde\Omega)} \le G(\sigma_\pm^*)[\|e^0\|_{C(\tilde\Omega)} + \|\nabla_{x,y} e^0\|_{C(\tilde\Omega)}]. \tag{3.59}$$

Obviously, (3.59) is also true in $\tilde\Omega_{z_0} := \tilde\Omega \bigcap \{(x,y,z) : z = z_0\} \subset \mathbb{R}^2$ for any z_0, that is,

$$\|\epsilon\partial_x w_j^0\|_{C(\tilde\Omega_{z_0})},\ \|\epsilon\partial_y w_j^0\|_{C(\tilde\Omega_{z_0})} \le G(\sigma_\pm^*)[\|e^0\|_{C(\tilde\Omega)} + \|\nabla_{x,y} e^0\|_{C(\tilde\Omega)}],$$

which corresponds to (3.21) in the two-dimensional case. As for (3.25) in the two-dimensional case, we have

$$\|\mathbb{A}[\sigma^*]^{-1}\|_{C(\tilde\Omega_{z_0})} \le \|\mathbb{A}[\sigma^*]^{-1}\|_{C(\tilde\Omega)} \le \frac{\sqrt{2}}{d_-^*} C_* \tag{3.60}$$

for any z_0 due to A2. We choose $\epsilon \in (0, \frac{1}{2K}\sigma_-^*)$ small enough such that

$$\epsilon \|\mathbb{A}[\sigma^*]^{-1}\|_{C(\tilde\Omega)} \sqrt{2} G(\sigma_\pm^*) K \le \frac{2\epsilon C_* G(\sigma_\pm^*) K}{d_-^*} < \frac{1}{2}; \tag{3.61}$$

then we get from the same argument as that in subsection 3.1 that

$$\begin{aligned} \|\nabla_{x,y}(\sigma^* - \sigma^1)\|_{C(\tilde\Omega_{z_0})} &\le 2\|\mathbb{A}[\sigma^*]^{-1}\|_{C(\tilde\Omega_{z_0})} \sqrt{2} \|\epsilon \nabla_{x,y} w_j\|_{C(\tilde\Omega_{z_0})} \|\nabla_{x,y}\sigma^*\|_{C(\tilde\Omega_{z_0})} \\ &\le 2\sqrt{2}\|\mathbb{A}[\sigma^*]^{-1}\|_{C(\tilde\Omega)} G(\sigma_\pm^*) K\epsilon^2. \end{aligned} \tag{3.62}$$

As for $\sigma^1 - \sigma^*$, we have the estimate $\|\sigma^1 - \sigma^*\|_{C(\tilde\Omega_{z_0})} \le K \|\nabla_{x,y}(\sigma^1 - \sigma^*)\|_{C(\tilde\Omega_{z_0})}$, and hence

$$\|\sigma^1 - \sigma^*\|_{C(\tilde\Omega)} \le K 2\sqrt{2} \|\mathbb{A}[\sigma^*]^{-1}\|_{C(\tilde\Omega)} G(\sigma_\pm^*) K\epsilon^2. \tag{3.63}$$

For each $\sigma^1(x,y,z_0)$ generated by the harmonic B_z method at each slice $\tilde\Omega_{z_0}$, we generate σ^1 in $\tilde\Omega \subset \mathbb{R}^3$ by $\sigma^1 := \bigcup_{z_0} \sigma^1(x,y,z_0)$. Now for $\epsilon \in (0, \frac{1}{2K}\sigma_-^*)$ satisfying (3.61),

$$K2\sqrt{2}\|\mathbb{A}[\sigma^*]^{-1}\|_{C(\tilde\Omega)} G(\sigma_\pm^*)\epsilon \le \frac{4}{d_-^*} C_* K G(\sigma_\pm^*)\epsilon := D_* G(\sigma_\pm^*)\epsilon := \theta \in (0,1); \tag{3.64}$$

then it follows that

$$\|\nabla_{x,y}(\sigma^1 - \sigma^*)\|_{C(\tilde\Omega)},\ \|\sigma^1 - \sigma^*\|_{C(\tilde\Omega)} \le K\theta\epsilon,$$

which means $F(\sigma^1) \le C_\epsilon(\sigma^*) \le G(\sigma_\pm^*)$. As in the two-dimensional case, we have $\sigma^1 = \sigma^*$ in $\Omega \setminus \tilde\Omega$. So the theorem is true for $n = 1$. It follows from (3.61) and (3.64) that $\epsilon = \epsilon(\sigma_\pm^*, d_-^*), \theta = \theta(\epsilon, \sigma_\pm^*, d_-^*)$.

Now we can apply the induction argument to prove the theorem. That is, assume that $\sigma^k \equiv \sigma^*$ in $\Omega \setminus \tilde\Omega$ and the following estimates:

$$\|\nabla_{x,y}(\sigma^k - \sigma^*)\|_{C(\tilde\Omega)},\ \|\sigma^k - \sigma^*\|_{C(\tilde\Omega)} \le K(D_* G(\sigma_\pm^*)\epsilon)^k \epsilon = K\theta^k\epsilon$$

are true for $k = n$. We shall prove that this is also true for $k = n + 1$.

This can be done by the same way as in the two-dimensional case, with the same modifications for the three-dimensional case as given in the proof of this theorem for the step $n = 1$. Indeed, $\nabla_{x,y}\sigma^n$ is continuous in every slice Ω_{z_0} from its definition (2.16) for $\sigma^* \in C^1$. This fact implies that (3.56) is true with σ^0 replaced by σ^n. Moreover, we have $C_s G(\sigma^n) \le C_\epsilon(\sigma^*) \le G(\sigma_\pm^*)$ from the assumption of the induction argument. So we omit the details. ∎

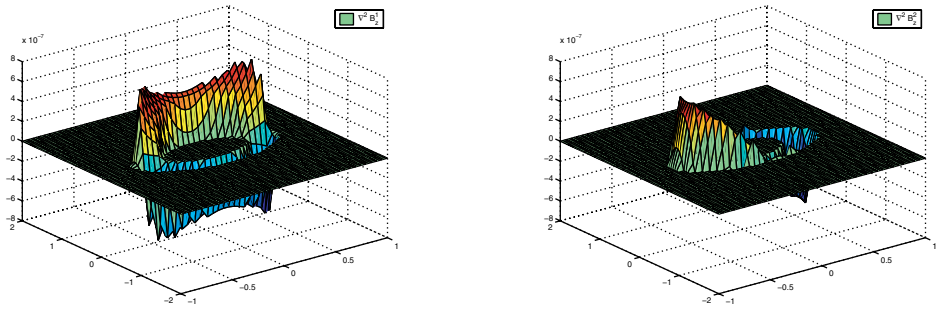


FIG. 4. *Distributions of* $\nabla^2 B_z^1$, $\nabla^2 B_z^2$.

4. Numerical performance. We present some numerics to show our theoretical convergence property. For simplicity, we deal with the axially symmetric case where input current densities in the electrodes are specified. That is, we consider the following two-dimensional problem for potential $u = u(x, y)$:

$$\begin{cases} \nabla \cdot (\sigma \nabla u) = 0 & \text{in } \Omega, \\ -\sigma \nabla u \cdot \mathbf{n} = g(x, y) & \text{on } \partial\Omega, \end{cases} \tag{4.1}$$

with $u(0, 0) = 0$ at the reference point $(0, 0)$. Consider the target conductivity in $\Omega := [-1, 1] \times [-2, 2]$ of the form

$$\sigma^*(r) = \begin{cases} 3 & \text{if} \quad 0 \le r \le 0.4, \\ -10r^2 + 4.6 & \text{if} \quad 0.4 \le r \le 0.6, \\ 1 & \text{otherwise}, \end{cases} \tag{4.2}$$

where $r = \sqrt{x^2 + y^2}$. Electrodes are specified as $\mathcal{E}_1^{\pm} := \{(\pm 1, y) : |y| < 0.1\}$ and $\mathcal{E}_2^{\pm} := \{(x, \pm 2) : |x| < 0.1\}$ on $\partial\Omega$. The input current densities $g^j, j = 1, 2$ are given by

$$g^j|_{\mathcal{E}_j^{\pm}} = \pm 1 \quad \text{and} \quad g^j = 0 \quad \text{on } \partial\Omega \setminus [\mathcal{E}_j^+ \cup \mathcal{E}_j^-]. \tag{4.3}$$

The corresponding $\nabla^2 B_z^j = \mu_0(\sigma_x u_y^{j,*} - \sigma_y u_x^{j,*})$ are shown in Figure 4, where $u^{j,*}$ is the solution corresponding to (σ^*, g_j) for $j = 1, 2$.

We introduce $\Phi(\mathbf{r}, \mathbf{r}') = \frac{1}{2\pi} \ln|\mathbf{r} - \mathbf{r}'|$ and divide Ω into $N \times M$ small rectangles. Denote by $\mathbf{r}_{k,l}$ the center of each rectangle $e(k, l)$. Then a simple computation generates the following discrete iteration formula:

$$\sigma^{n+1}(\mathbf{r}_{k,l}) = f(\mathbf{r}_{k,l}) - \frac{1}{\mu_0} \sum_{i,j=1}^{N,M} \int_{e(i,j)} \nabla\Phi(\mathbf{r}', \mathbf{r}_{k,l}) \cdot \mathbb{A}[\sigma^n]^{-1} \begin{pmatrix} \nabla^2 B_z^1 \\ \nabla^2 B_z^2 \end{pmatrix} (\mathbf{r}') d\mathbf{r}', \tag{4.4}$$

where $f(\mathbf{r}) = \int_{\partial\Omega} \nabla_{\mathbf{r}'}\Phi(\mathbf{r}, \mathbf{r}') \cdot \mathbf{n}_{\mathbf{r}'}\, \sigma^*(\mathbf{r}') ds(\mathbf{r}')$.

For the integrals in the elements, where we take $\mathbb{A}[\sigma^n]^{-1}, \nabla^2 B_z$ as constants at each element, they are zero for $(i, j) = (k, l)$ due to the symmetric property (noticing that $\mathbf{r}_{k,l}$ is the center of $e(k, l)$) and regular for $(i, j) \neq (k, l)$. So we can construct the sequence $\{\sigma^n\}$ inside Ω for a given initial guess $\sigma^0(\mathbf{r})$.

In our numerical test, the finite element method with bilinear base functions $\phi^{i,j}(x, y) = (a + by)(c + dy)$ at each nodal point are used to solve the direct problem

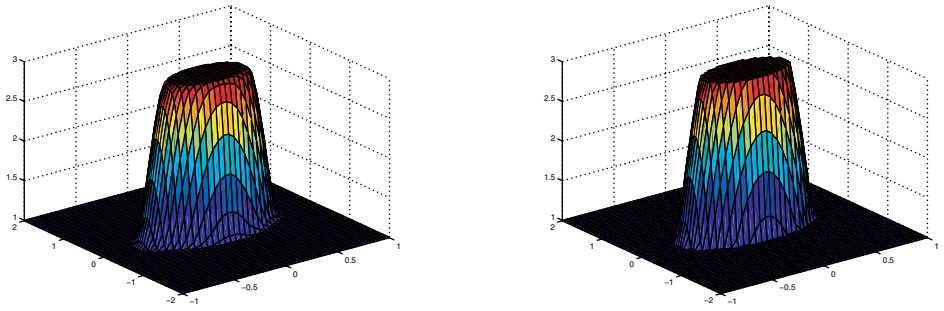


FIG. 5. *Reconstruction after six iterations (left) and the exact one (right).*

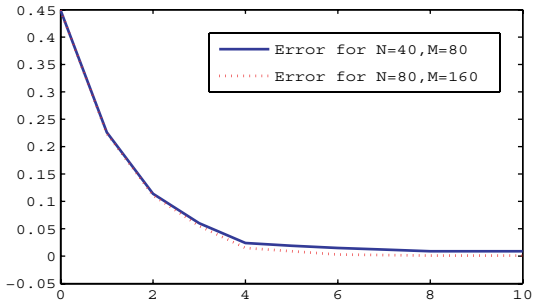


FIG. 6. *$E(n)$ with respect to the iteration number n for two different meshes.*

for $u[\sigma^n]$ at each iteration step. Then $\nabla u[\sigma^n]$ at the center of each element can be computed by the difference method. We also use this scheme to simulate $\nabla^2 B_z$ from (4.1)–(4.3) for our inversion input. To avoid the well-known *inverse crime* in the numerical tests [5], we use different grids in simulating the input data from those used in the inversion algorithm.

First, we take $N = 40, M = 80$ (case 1) and the initial guess function $\sigma_0(x, y) \equiv 1$. The recovering result after six iterations as well as the exact one are shown in Figure 5. Now we choose a finer mesh with $N = 80, M = 160$ (case 2). Denote by $E(n)$ the relative L^2-error between the target conductivity and the reconstructed one after the nth iteration. The numerical values of $E(n)$ in these two cases are given in Table 4.1, while curves are plotted in Figure 6 for a bigger iteration number $n = 10$.

From the error distributions in Figure 6 and Table 4.1, we can observe that the iteration algorithm converges very quickly. In fact, the error is almost unchanged after six iterations. This phenomenon matches very well with our theoretical result, which assumes a convergence rate of the order θ^n, with $\theta \in (0, 1)$. The excellent convergence performance comes from our good initial guess $\sigma^0 \equiv 1$. This means that the error of ϵ for the initial guess is not so large, and θ will be also small for small ϵ. We can also observe that the finer mesh can improve the inversion result at the expense of an increased number of computations.

Now let us consider an inferior initial guess function

$$\sigma^0(x, y) = \frac{4}{5} - \frac{1}{5}\cos\frac{(x^2 + y^2)\pi}{5}, \tag{4.5}$$

TABLE 4.1
Relative L^2-error $E(n)$ in two cases.

n	Case 1	Case 2	n	Case 1	Case 2
0	0.449118	0.449185	4	0.035278	0.028515
1	0.226044	0.224892	5	0.024009	0.015632
2	0.114281	0.111283	6	0.019200	0.009711
3	0.060467	0.055545	7	0.017172	0.007149

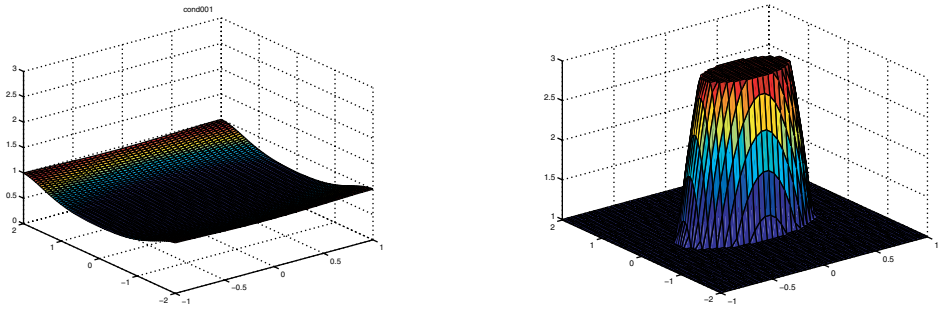


FIG. 7. *Inferior initial guess σ^0 (left) and the exact σ^* (right).*

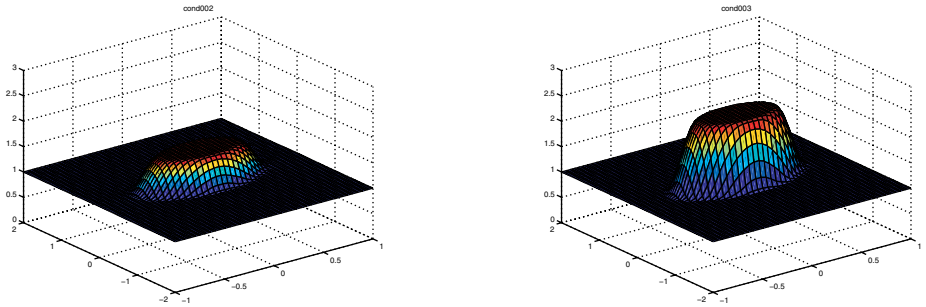


FIG. 8. *Reconstruction for $n = 1$ (left) and $n = 2$ (right) using the inferior initial guess.*

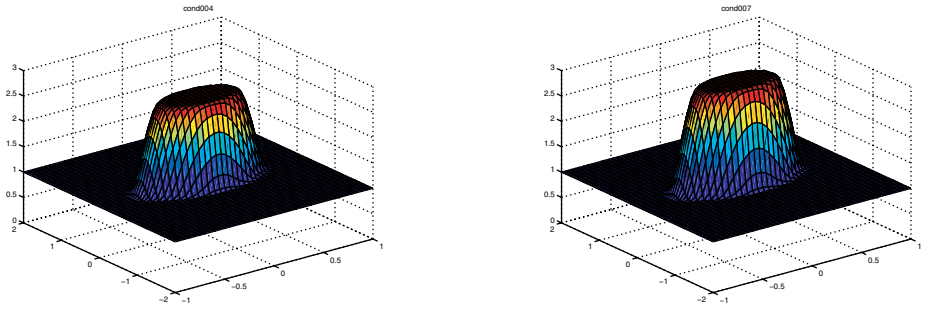


FIG. 9. *Reconstruction for $n = 3$ (left) and $n = 6$ (right) using the inferior initial guess.*

which is quite different from the exact $\sigma^*(x, y)$; see Figure 7. The reconstruction results for $n = 1, 2, 3, 6$ are given in Figures 8 and 9 with its error distribution illustrated in Figure 10. Even for this case with the undesirable initial guess, we can see that the algorithm still catches the target σ^* in the whole domain.

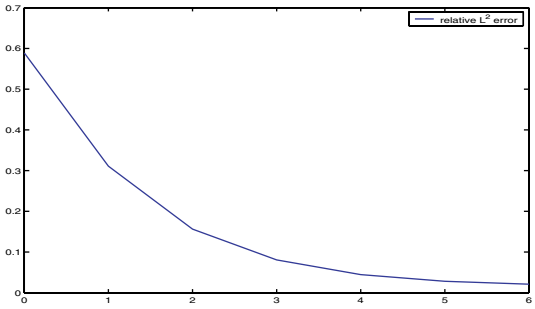


FIG. 10. *Relative error distribution for the inferior initial guess.*

It should be noticed that we used the simulation data $\nabla^2 B_z^j, j = 1, 2$, directly for the harmonic B_z algorithm rather than computing $\nabla^2 B_z^j$ from B_z^j. Obviously, if we use noisy measured B_z data as the inversion input, a suitable denoising technique must be used. Noticing the expression of the iteration (2.16), the harmonic B_z algorithm uses in fact the first derivative of B_z. A similar inversion scheme using the first derivative of B_z named the gradient B_z method can be found in [16, 17].

Acknowledgment. We thank a referee for pointing out results about the condition (2.11) in Remark 2.5.

REFERENCES

[1] G. ALESSANDRINI AND V. NESI, *Univalent σ-harmonic mappings*, Arch. Ration. Mech. Anal., 50 (2001), pp. 747–757.

[2] P. BAUMAN, A. MARINI, AND V. NESI, *Univalent solutions of an elliptic system of partial differential equations arising in homogenization*, Indiana Univ. Math. J., 128 (2000), pp. 53–64.

[3] M. BRIANE, G. W. MILTON, AND V. NESI, *Change of sign of the correctors determinant for homogenization in three-dimensional conductivity*, Arch. Ration. Mech. Anal., 173 (2004), pp. 133–150.

[4] M. CHENEY, D. ISAACSON, AND J. C. NEWELL, *Electrical impedance tomography*, SIAM Rev., 41 (1999), pp. 85–101.

[5] D. COLTON AND R. KRESS, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer-Verlag, Berlin, 1998.

[6] G. DIFAZIO, *L^p estimates for divergence form elliptic equations with discontinuous coefficients*, Boll. Unione Mat. Ital. Sez. A Mat. Soc. Cult., 10 (1996), pp. 409–420.

[7] G. B. FOLLAND, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976.

[8] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 2001.

[9] Y. Z. IDER, S. ONART, AND W. LIONHEART, *Uniqueness and reconstruction in magnetic resonance-electrical impedance tomography (MR-EIT)*, Physiol. Meas., 24 (2003), pp. 591–604.

[10] O. KWON, C. PARK, E. J. PARK, J. K. SEO, AND E. J. WOO, *Electrical conductivity imaging using a variational method in B_z-based MREIT*, Inverse Problems, 21 (2005), pp. 969–980.

[11] R. S. LAUGESEN, *Injectivity can fail for higher-dimensional harmonic extensions*, Complex Var., 28 (1996), pp. 357–369.

[12] J. J. LIU, H. C. PYO, J. K. SEO, AND E. J. WOO, *Convergence properties and stability issues in the MREIT algorithm*, Contemp. Math., 408 (2006), pp. 201–218.

[13] S. LEE, J. K. SEO, C. J. PARK, B. I. LEE, E. J. WOO, S. Y. LEE, O. KWON, AND J.HAHN, *Conductivity image reconstruction from defective data in MREIT: Numerical simulation and animal experiment*, IEEE Trans. Med. Imag., 25 (2006), pp. 168–176.

[14] P. METHERALL, D. C. BARBER, R. H. SMALLWOOD, AND B. H. BROWN, *Three-dimensional electrical impedance tomography*, Nature, 380 (1996), pp. 509–512.

[15] S. H. Oh, B. I. Lee, E. J. Woo, S. Y. Lee, T. S. Kim, O. Kwon, and J. K. Seo, *Electrical conductivity images of biological tissue phantoms in MREIT*, Physiol. Meas., 26 (2005), pp. 279–288.

[16] C. Park, E. J. Park, E. J. Woo, O. Kwon, and J. K. Seo, *Static conductivity imaging using variational gradient B_z algorithm in magnetic resonance electrical impedance tomography*, Physiol. Meas., 25 (2004), pp. 257–269.

[17] C. Park, O. Kwon, E. J. Woo, and J. K. Seo, *Electrical conductivity imaging using gradient B_z decomposition algorithm in magnetic resonance electrical impedance tomography (MREIT)*, IEEE Trans. Med. Imag., 23 (2004), pp. 388–394.

[18] J. K. Seo, J. R. Yoon, E. J. Woo, and O. Kwon, *Reconstruction of conductivity and current density images using only one component of magnetic field measurements*, IEEE Trans. Biomed. Eng., 50 (2003), pp. 1121–1124.

[19] J. K. Seo, H. C. Pyo, C. Park, O. Kwon, and E. J. Woo, *Image reconstruction of anisotropic conductivity tensor distribution in MREIT: Computer simulation study*, Phys. Med. Biol., 49 (2004), pp. 4371–4382.

[20] G. Verchota, *Layer potentials and boundary value problems for Laplace's equation in Lipschitz domains*, J. Funct. Anal., 59 (1984), pp. 572–611.

[21] J. G. Webster, *Electrical Impedance Tomography*, Adam Hilger, Bristol, UK, 1990.

SIAM J. APPL. MATH.
Vol. 67, No. 5, pp. 1283–1309

ASYMPTOTIC PROFILES OF THE STEADY STATES FOR AN SIS EPIDEMIC PATCH MODEL*

L. J. S. ALLEN†, B. M. BOLKER‡, Y. LOU§, AND A. L. NEVAI¶

Abstract. Spatial heterogeneity, habitat connectivity, and rates of movement can have large impacts on the persistence and extinction of infectious diseases. These factors are shown to determine the asymptotic profile of the steady states in a frequency-dependent SIS (susceptible-infected-susceptible) epidemic model with n patches in which susceptible and infected individuals can both move between patches. Patch differences in local disease transmission and recovery rates characterize whether patches are low-risk or high-risk, and these differences collectively determine whether the spatial domain, or habitat, is low-risk or high-risk. The basic reproduction number $\mathcal{R}_0$ for the model is determined. It is then shown that when the disease-free equilibrium is stable ($\mathcal{R}_0 < 1$) it is globally asymptotically stable, and that when the disease-free equilibrium is unstable ($\mathcal{R}_0 > 1$) there exists a unique endemic equilibrium. Two main theorems link spatial heterogeneity, habitat connectivity, and rates of movement to disease persistence and extinction. The first theorem relates the basic reproduction number to the heterogeneity of the spatial domain. For low-risk domains, the disease-free equilibrium is stable ($\mathcal{R}_0 < 1$) if and only if the mobility of infected individuals lies above a threshold value, but for high-risk domains, the disease-free equilibrium is always unstable ($\mathcal{R}_0 > 1$). The second theorem states that when the endemic equilibrium exists, it tends to a spatially inhomogeneous disease-free equilibrium as the mobility of susceptible individuals tends to zero. This limiting disease-free equilibrium has a positive number of susceptible individuals on all low-risk patches and can also have a positive number of susceptible individuals on some, but not all, high-risk patches. Sufficient conditions for whether high-risk patches in the limiting disease-free equilibrium have susceptible individuals or not are given in terms of habitat connectivity, and these conditions are illustrated using numerical examples. These results have important implications for disease control.

Key words. spatial heterogeneity, dispersal, habitat connectivity, basic reproduction number, disease-free equilibrium, endemic equilibrium

AMS subject classifications. 92D30, 92D40, 92D50, 91D25, 34C60, 37N25

DOI. 10.1137/060672522

1. Introduction.

Spatial heterogeneity, habitat connectivity, and rates of movement play important roles in disease persistence and extinction. Movement of susceptible or infected individuals can enhance or suppress the spread of disease, depending on the heterogeneity and connectivity of the spatial environment (see, e.g., Castillo-Chavez and Yakubu (2001, 2002); Bolker and Grenfell (1995); Hess (1996); Lloyd and May (1996); Salmani and van den Driessche (2006); Ruan (2006); Allen et al. (2007)). For example, the compartmental patch model of Ruan, Wang, and

*Received by the editors October 16, 2006; accepted for publication (in revised form) March 19, 2007; published electronically June 19, 2007. This research was initiated during the visit of the first and second authors to the Mathematical Biosciences Institute, The Ohio State University. This material is based upon work supported by the National Science Foundation under agreement 0112050.
http://www.siam.org/journals/siap/67-5/67252.html

†Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409-1042 (linda.j.allen@ttu.edu). This author acknowledges support from NSF grant DMS-0201105.

‡Department of Zoology, University of Florida, P.O. Box 118525, Gainesville, FL 32611-8525 (bolker@zoo.ufl.edu). This author acknowledges support from NSF Integrated Research Challenges in Environmental Biology grant IBN 9977063.

§Department of Mathematics, Ohio State University, Columbus, OH 43210 (lou@math.ohio-state.edu). This author acknowledges support from NSF grant DMS-0615845.

¶Corresponding author. Mathematical Biosciences Institute, Ohio State University, Columbus, OH 43210 (anevai@mbi.osu.edu).

Levin (2006) demonstrates that diseases such as SARS can be contained by screening for infection at borders and barring residents of disease hot spots from travel. Spatial heterogeneity can also give rise to complex and surprising disease dynamics (Allen, Kirupaharan, and Wilson (2004); Castillo-Chavez and Yakubu (2001, 2002); Hess (1996); Lloyd and Jansen (2004); Wang and Zhao (2004)). In numerical investigations of a discrete-time two-patch SIS (susceptible-infected-susceptible) epidemic model, Allen, Kirupaharan, and Wilson (2004) considered a case where, in the absence of movement, the disease persists in only one of the two patches—a high-risk patch, where the patch reproduction number is greater than one. When the patches are connected by susceptible and infective movement, an endemic equilibrium is reached in both patches. But if the movement pattern is changed so that only infected individuals disperse between the two patches, a surprising result occurs. The disease does not persist in either patch; the high-risk patch becomes empty, and all susceptible individuals eventually reside in the low-risk patch, where the patch reproduction number is less than one. We investigate this latter phenomenon in a continuous-time SIS metapopulation model with n patches that includes both high-risk and low-risk patches.

Disease spread in metapopulation models involving discrete patches has been investigated in a variety of settings (Arino and van den Driessche (2006, 2003a, 2003b); Arino et al. (2005); Jin and Wang (2005); Rvachev and Longini (1985); Salmani and van den Driessche (2006); Sattenspiel and Dietz (1995); Wang and Mulone (2003); Wang and Zhao (2004)). In a review article, Arino and van den Driessche (2006) summarize some known results on disease dynamics in metapopulation models with regard to existence and stability of disease-free and endemic equilibria. They develop a general framework for movement of susceptible, exposed, infected, and recovered individuals (SEIRS model) and define a *mobility matrix*, an irreducible matrix that defines the spatial arrangement of patches and rates of movement between patches (see also Arino and van den Driessche (2003a, 2003b)). Wang and colleagues studied uniform persistence and global stability of disease-free and endemic equilibria in SIS metapopulation models (Jin and Wang (2005); Wang and Mulone (2003); Wang and Zhao (2004)).

Here, we formulate a frequency-dependent SIS metapopulation model consisting of n patches. The spatial arrangement of patches, and rates of movement between patches, are defined by an irreducible matrix. The spatial domain is characterized as *low-risk* or *high-risk* if the spatial average of the patch transmission rates is less than or greater than, respectively, the spatial average of the recovery rates. Individual patches are also characterized as *low-risk* or *high-risk* if the patch transmission rate is less than or greater than the patch recovery rate, which is equivalent to the patch reproduction number being less than or greater than one, respectively. A unique disease-free equilibrium is shown to exist, and a basic reproduction number $\mathcal{R}_0$ is determined. If $\mathcal{R}_0 < 1$, the disease-free equilibrium is shown to be globally asymptotically stable, and if $\mathcal{R}_0 > 1$, a unique endemic equilibrium is shown to exist.

Our two main theorems link spatial heterogeneity, habitat connectivity, and rates of movement to disease persistence and extinction. The first theorem relates the basic reproduction number to the heterogeneity of the spatial domain. It is shown that for low-risk domains, the disease-free equilibrium is stable ($\mathcal{R}_0 < 1$) if and only if the mobility of infected individuals lies above a threshold value. For high-risk domains the disease-free equilibrium is always unstable ($\mathcal{R}_0 > 1$). The second theorem concerns the spatial heterogeneity in the limiting case where the mobility of susceptible individuals approaches zero. We show that if $\mathcal{R}_0 > 1$, then the endemic

equilibrium approaches a spatially inhomogeneous disease-free equilibrium which has a positive number of susceptible individuals on all low-risk patches and no susceptibles on at least one of the high-risk patches. These results have important implications for disease control. If the spatial environment can be modified to include low-risk patches (i.e., low transmission rates or high recovery rates) and if the movement of susceptible individuals can be restricted (e.g., quarantine), then it may be possible to eliminate the disease.

1.1. The model. Let $n \geq 2$ be the number of patches and $\Omega = \{1, 2, \ldots, n\}$. Consider the SIS patch model

$$\frac{d\bar{S}_j}{dt} = d_S \sum_{k\in\Omega}(L_{jk}\bar{S}_k - L_{kj}\bar{S}_j) - \frac{\beta_j \bar{S}_j \bar{I}_j}{\bar{S}_j + \bar{I}_j} + \gamma_j \bar{I}_j, \quad j \in \Omega, \tag{1.1a}$$

$$\frac{d\bar{I}_j}{dt} = d_I \sum_{k\in\Omega}(L_{jk}\bar{I}_k - L_{kj}\bar{I}_j) + \frac{\beta_j \bar{S}_j \bar{I}_j}{\bar{S}_j + \bar{I}_j} - \gamma_j \bar{I}_j, \quad j \in \Omega, \tag{1.1b}$$

where $\bar{S}_j(t)$ and $\bar{I}_j(t)$ denote the number of susceptible and infected individuals in patch j at time $t \geq 0$; d_S and d_I are positive diffusion coefficients for the susceptible and infected subpopulations; L_{jk} represents the degree of movement from patch k into patch j; and β_j and γ_j are nonnegative constants that express the rate of disease transmission and recovery in patch j. Because $\bar{S}_j\bar{I}_j/(\bar{S}_j+\bar{I}_j)$ is a Lipschitz continuous function of $\bar{S}_j$ and $\bar{I}_j$ in the open first quadrant, we extend its definition to the entire first quadrant by defining it to be zero when at least one of $\bar{S}_j = 0$ or $\bar{I}_j = 0$ holds. We assume that

(A1) $\bar{S}_j(0) \geq 0$ and $\bar{I}_j(0) \geq 0$ for $j \in \Omega$, and $\sum_{j\in\Omega}\left[\bar{S}_j(0) + \bar{I}_j(0)\right] > 0$.

Let $\bar{S} = (\bar{S}_j)$ and $\bar{I} = (\bar{I}_j)$. Brauer and Nohel's work (1989) implies that a unique solution $(\bar{S}, \bar{I})$ of (1.1) exists for all time. Let

$$N = \sum_{j\in\Omega}\left[\bar{S}_j(0) + \bar{I}_j(0)\right] \tag{1.2}$$

be the total number of individuals in all patches at $t = 0$. By (A1), N is positive. Summing the $2n$ equations in (1.1) makes it clear that

$$\sum_{j\in\Omega}\left[\bar{S}_j(t) + \bar{I}_j(t)\right] = N, \quad t \geq 0. \tag{1.3}$$

We will assume that the *connectivity matrix* $L = (L_{jk})$ satisfies

(A2) L is nonnegative, irreducible, and symmetric.

We shall say that a matrix $A = (A_{jk})$ is *nonnegative* (or *positive*) if all its elements are nonnegative (or positive), in which case we will write $A \geq 0$ (or $A > 0$). Similar comments apply to vectors $u = (u_j)$. The symmetry assumption ensures that the per capita rates of susceptible and infected individuals entering patch j from patch k ($d_S L_{jk}$ and $d_I L_{jk}$) are equal to the per capita rates of individuals moving in the other direction ($d_S L_{kj}$ and $d_I L_{kj}$). Hence, in (1.1)

$$L_{jk}\bar{S}_k - L_{kj}\bar{S}_j = L_{jk}(\bar{S}_k - \bar{S}_j) \quad \text{and} \quad L_{jk}\bar{I}_k - L_{kj}\bar{I}_j = L_{jk}(\bar{I}_k - \bar{I}_j).$$

The irreducibility assumption implies that the system of patches considered as a directed graph with patches as the vertices is strongly connected (Ortega (1987)).

Other characterizations of irreducibility are given in Appendix A, and we will make use of these additional facts as needed.

We say that in a *low-risk patch* disease transmission occurs at a lower rate than disease recovery when the number of infected individuals in that patch is very small. A *high-risk patch* is defined in a similar manner. Let

$$H^- = \{j \in \Omega : \beta_j < \gamma_j\} \quad \text{and} \quad H^+ = \{j \in \Omega : \beta_j > \gamma_j\}$$

denote the set of these low-risk and high-risk patches, respectively. We assume that

(A3) H^- and H^+ are nonempty and $H^- \cup H^+ = \Omega$.

Let $\mathcal{R}_0^{[j]} = \beta_j/\gamma_j$ be the *patch reproduction number* for patch $j \in \Omega$ (we set $\mathcal{R}_0^{[j]} = \infty$ when $\gamma_j = 0$). Then $\mathcal{R}_0^{[j]} < 1$ for low-risk patches ($j \in H^-$), and $\mathcal{R}_0^{[j]} > 1$ for high-risk patches ($j \in H^+$). It is well known that the disease can persist in isolated high-risk patches but not in isolated low-risk patches.

Let

$$\Sigma_\beta = \sum_{j\in\Omega} \beta_j \quad \text{and} \quad \Sigma_\gamma = \sum_{j\in\Omega} \gamma_j.$$

We say that Ω is a *low-risk domain* if $\Sigma_\beta < \Sigma_\gamma$, but a *high-risk domain* if $\Sigma_\beta \geq \Sigma_\gamma$.

For an arbitrary patch $j \in \Omega$, it will be convenient to define

$$L_j = \sum_{k\in\Omega} L_{jk}, \quad L_j^- = \sum_{k\in H^-} L_{jk}, \quad \text{and} \quad L_j^+ = \sum_{k\in H^+} L_{jk}.$$

These sums denote the connectivity between patch j and all patches, all low-risk patches, and all high-risk patches, respectively. The irreducibility of L implies that $L_j > 0$ for all $j \in \Omega$.

1.2. The equilibrium problem. We will be primarily interested in equilibrium solutions of (1.1), i.e., solutions of

$$d_S \sum_{k\in\Omega} L_{jk}(\tilde{S}_k - \tilde{S}_j) - \frac{\beta_j \tilde{S}_j \tilde{I}_j}{\tilde{S}_j + \tilde{I}_j} + \gamma_j \tilde{I}_j = 0, \quad j \in \Omega, \tag{1.4a}$$

$$d_I \sum_{k\in\Omega} L_{jk}(\tilde{I}_k - \tilde{I}_j) + \frac{\beta_j \tilde{S}_j \tilde{I}_j}{\tilde{S}_j + \tilde{I}_j} - \gamma_j \tilde{I}_j = 0, \quad j \in \Omega, \tag{1.4b}$$

where $\tilde{S}_j$ and $\tilde{I}_j$ denote the number of susceptible and infected individuals, respectively, in patch $j \in \Omega$ at equilibrium. In view of (1.3), we impose the condition

$$\sum_{j\in\Omega} \left(\tilde{S}_j + \tilde{I}_j\right) = N. \tag{1.4c}$$

Let $\tilde{S} = (\tilde{S}_j)$ and $\tilde{I} = (\tilde{I}_j)$. We are interested only in solutions $(\tilde{S}, \tilde{I})$ of (1.4) which satisfy $\tilde{S} \geq 0$ and $\tilde{I} \geq 0$. A *disease-free equilibrium* (DFE) is a solution in which $\tilde{I}_j = 0$ for all $j \in \Omega$. An *endemic equilibrium* (EE) is a solution in which $\tilde{I}_j > 0$ for some $j \in \Omega$. To distinguish between these two types of equilibria, we will for notational convenience denote a DFE by $(\hat{S}, 0)$ and an EE by $(\tilde{S}, \tilde{I})$.

1.3. Statement of the main results. We consider in section 2 properties of the DFE, including its existence, uniqueness, and stability. We first show that there exists a unique DFE $(\hat{S}, 0)$, and it is given by $\hat{S}_j = N/n$ for $j \in \Omega$. We then calculate the basic reproduction number $\mathcal{R}_0$ for (1.1) using the next generation approach (Diekmann, Heesterbeek, and Metz (1990); Diekmann and Heesterbeek (2000); van den Driessche and Watmough (2002)) for which it is known that if $\mathcal{R}_0 < 1$, then the DFE is locally asymptotically stable, but if $\mathcal{R}_0 > 1$, then it is unstable. Our calculation will show that $\mathcal{R}_0$ does not depend on the diffusion coefficient d_S. Finally, we show that if $\mathcal{R}_0 < 1$, then the DFE is globally asymptotically stable.

In section 3, we find an equivalent characterization for the stability of the DFE in terms of d_I rather than $\mathcal{R}_0$. In particular, we show that the DFE in a low-risk domain is stable if and only if the diffusion coefficient for infected individuals lies above a certain threshold value, but in a high-risk domain, the DFE is always unstable. We also show that when the DFE is unstable, then there exists a unique EE. Moreover, the disease persists in every patch.

THEOREM 1. *Suppose that* (A1)–(A3) *hold and N is fixed.*

(a) *In a low-risk domain ($\Sigma_\beta < \Sigma_\gamma$), there exists a threshold value $d_I^* \in (0, \infty)$ such that $\mathcal{R}_0 > 1$ for $d_I < d_I^*$ and $\mathcal{R}_0 < 1$ for $d_I > d_I^*$.*

(b) *In a high-risk domain ($\Sigma_\beta \geq \Sigma_\gamma$), we have $\mathcal{R}_0 > 1$ for all d_I.*

(c) *If $\mathcal{R}_0 > 1$, then an EE exists, it is unique, and $\tilde{I} > 0$.*

Observe from (1.4) that in the limiting case $d_S = 0$ there also exists a family of infinitely many spatially inhomogeneous DFEs $(\hat{S}, 0)$, each of which satisfies

$$\hat{S} \geq 0 \quad \text{and} \quad \sum_{j \in \Omega} \hat{S}_j = N. \tag{1.5}$$

In section 4, we show that if $\mathcal{R}_0 > 1$, then the EE approaches such a spatially inhomogeneous DFE as the mobility of susceptible individuals becomes very small. We write this limiting DFE as $(S^*, 0)$ and also consider the distribution of patches for which S^* is either positive or zero.

THEOREM 2. *Suppose that* (A1)–(A3) *hold, N is fixed, and $\mathcal{R}_0 > 1$.*

(a) *$(\tilde{S}, \tilde{I}) \to (S^*, 0)$ as $d_S \to 0$ for some S^* satisfying* (1.5)*;*

(b) *$S^* > 0$ on H^- and $S_j^* = 0$ for some $j \in H^+$;*

(c) *if*

$$\frac{1}{d_I} > \max_{k \in H^+} \left[\frac{L_k^-}{\beta_k - \gamma_k} \right] + \max_{k \in H^-} \left[\frac{L_k^+}{\beta_k - \gamma_k} \right], \tag{1.6}$$

then $S^* \equiv 0$ on H^+;

(d) *if*

$$\frac{1}{d_I} < \frac{L_j^-}{\beta_j - \gamma_j} + \min_{k \in H^-} \left[\frac{L_k^+}{\beta_k - \gamma_k} \right] \tag{1.7}$$

for some $j \in H^+$, then $S_j^* > 0$.

We now make several remarks concerning Theorem 2, which connects spatial heterogeneity, habitat connectivity, and rates of movement. First, condition (1.6) will be satisfied whenever d_I is sufficiently small.

Second, Theorem 2(c) immediately implies that if

$$\frac{1}{d_I} > \max_{k \in H^+} \left[\frac{L_k^-}{\beta_k - \gamma_k} \right], \tag{1.8}$$

then $S^* \equiv 0$ on H^+ because $L_k^+/(\beta_k - \gamma_k)$ is nonpositive for every $k \in H^-$. Although condition (1.6) is more inclusive than condition (1.8), the latter is usually easier to verify. Furthermore, if some low-risk patch ($k \in H^-$) is not directly connected to any high-risk patches ($L_k^+ = 0$), then conditions (1.6) and (1.8) are in fact equivalent.

Third, Theorem 2(d) implies that if

$$\frac{1}{d_I} < \max_{k\in H^+}\left[\frac{L_k^-}{\beta_k - \gamma_k}\right] + \min_{k\in H^-}\left[\frac{L_k^+}{\beta_k - \gamma_k}\right], \tag{1.9}$$

then $S^* \not\equiv 0$ on H^+.

1.4. Examples. Before proving Theorems 1 and 2, we first illustrate the second theorem with some examples of metapopulations occupying different distributions of low-risk and high-risk patches.

Example 1. If $H^- = \{1, 2, \ldots, n-1\}$ and $H^+ = \{n\}$, then Theorem 2(b) implies that $S^* > 0$ on H^- and $S^* = 0$ on H^+. For this case, condition (1.6) in Theorem 2 (c) may or may not hold, but condition (1.7) in Theorem 2(d) cannot. That is, for this particular configuration of patches, the assumption that $\mathcal{R}_0 > 1$ in Theorem 2 excludes the possibility that condition (1.7) can be satisfied.

Example 2. If $H^- = \{1\}$ and $H^+ = \{2, 3, \ldots, n\}$, then

$$\max_{k\in H^-}\left[\frac{L_k^+}{\beta_k - \gamma_k}\right] = \min_{k\in H^-}\left[\frac{L_k^+}{\beta_k - \gamma_k}\right] = \frac{L_1^+}{\beta_1 - \gamma_1}.$$

Theorem 2(b) implies that $S^* > 0$ on H^-, and Theorem 2(c), (d) provide necessary and sufficient conditions (except in the case of equality) for determining whether $S^* \equiv 0$ or $S^* \not\equiv 0$ on H^+. For example, suppose that there are $n = 3$ patches with $H^- = \{1\}$ and $H^+ = \{2, 3\}$. If

$$\frac{1}{d_I} > \max\left\{\frac{L_{21}}{\beta_2 - \gamma_2}, \frac{L_{31}}{\beta_3 - \gamma_3}\right\} + \frac{L_{12} + L_{13}}{\beta_1 - \gamma_1},$$

then $S^* \equiv 0$ on H^+, but if

$$\frac{1}{d_I} < \max\left\{\frac{L_{21}}{\beta_2 - \gamma_2}, \frac{L_{31}}{\beta_3 - \gamma_3}\right\} + \frac{L_{12} + L_{13}}{\beta_1 - \gamma_1},$$

then either $S_2^* = 0$ and $S_3^* > 0$ or $S_2^* > 0$ and $S_3^* = 0$.

Example 3. Suppose there are $n = 9$ patches arranged and connected as in Figure 1. We assume that $L_{ij} \in \{0, 1\}$ with $L_{ij} = 1$ whenever patches i and j are connected by an arrow. In addition, $\gamma_j = 1$ and $\bar{S}_j(0) + \bar{I}_j(0) = 100$ for $j \in \Omega$, so that $\mathcal{R}_0^{[j]} = \beta_j$ and $N = 900$. Four numerical examples (see Figure 2) illustrate the values of S_j^* for $j \in \Omega$. Low-risk patches ($\mathcal{R}_0^{[j]} < 1$) are gray, and high-risk patches ($\mathcal{R}_0^{[j]} > 1$) are white. For $\mathcal{R}_0 > 1$ ($d_I < d_I^*$), the value of S_j^* was approximated by $\tilde{S}_j$, which was calculated using the iterative method (3.11) with $d_S \le 10^{-5}$, $d_I = 1$, and $\sum_{j\in\Omega} \tilde{I}_j < 0.005$.

For the limiting DFE in Figure 2(a), susceptibles can persist only on low-risk patches. In this case, condition (1.6) of Theorem 2(c) is satisfied:

$$1 = \frac{1}{d_I} > \max_{k\in H^+}\left\{\frac{L_k^-}{\beta_k - \gamma_k}\right\} + \max_{k\in H^-}\left\{\frac{L_k^+}{\beta_k - \gamma_k}\right\} = \frac{2}{0.5} - \frac{2}{0.5} = 0.$$

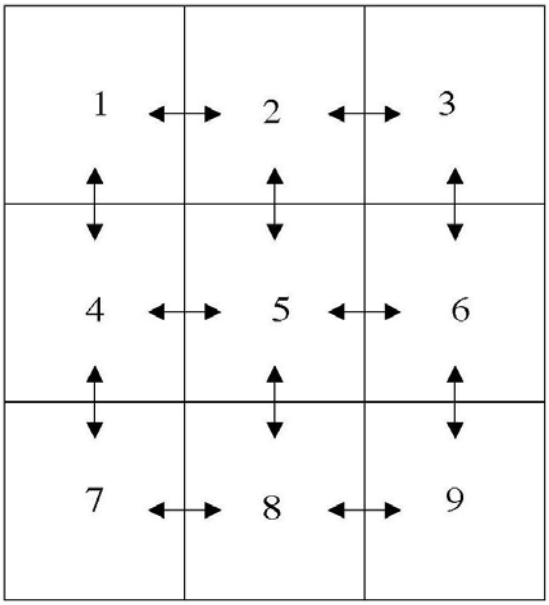


FIG. 1. *Nine patches connected at their boundaries.*

For the limiting DFE in Figure 2(b), susceptibles can persist on several high-risk patches. In this case, condition (1.7) of Theorem 2(d) is satisfied for $j = 2, 5, 6$,

$$1 = \frac{1}{d_I} < \frac{L_j^-}{\beta_j - \gamma_j} + \min_{k \in H^-} \left\{ \frac{L_k^+}{\beta_k - \gamma_k} \right\} = \frac{\{2 \text{ or } 1\}}{0.25} - \frac{1}{0.5} = \{6 \text{ or } 2\},$$

but not for $j = 3$:

$$1 = \frac{1}{d_I} > \frac{L_3^-}{\beta_3 - \gamma_3} + \min_{k \in H^-} \left\{ \frac{L_k^+}{\beta_k - \gamma_k} \right\} = \frac{0}{0.25} - \frac{1}{0.5} = -2.$$

We consider Figure 2(c), (d) in the Discussion.

2. The disease-free equilibrium. Throughout this section, we assume that (A1)–(A3) hold and that N is fixed.

2.1. Existence and uniqueness of the DFE. Equation (1.4) has a unique disease-free solution, and it is spatially homogeneous.

LEMMA 2.1. *A DFE* $(\hat{S}, 0)$ *exists, it is unique, and it is given by* $\hat{S}_j = N/n$ *for* $j \in \Omega$.

Proof. It is clear from (1.4) that $(\hat{S}, 0)$, with $\hat{S}_j = N/n$ for $j \in \Omega$, is a DFE. Now, let $(\tilde{S}, 0)$ be any DFE. Choose $m \in \Omega$ such that $\tilde{S}_m = \min\{\tilde{S}_j : j \in \Omega\}$. Setting $\tilde{I} = 0$ in (1.4a) with $j = m$ leads to $\sum_{k \in \Omega} L_{mk}(\tilde{S}_k - \tilde{S}_m) = 0$. The minimality of $\tilde{S}_m$ implies that $\tilde{S}_k = \tilde{S}_m$ whenever $L_{mk} > 0$. Let $j \in \Omega$ with $j \neq m$. The irreducibility of L implies that there exists a chain from j to m, i.e., a sequence $j_1, j_2, \ldots, j_s \in \Omega$ with $j_1 = j$ and $j_s = m$ such that $L_{j_p j_{p+1}} > 0$ for $1 \leq p \leq s-1$. Thus $\tilde{S}_{j_p} = \tilde{S}_{j_{p+1}}$ for $1 \leq p \leq s-1$. We conclude that $\tilde{S}_j = \tilde{S}_m$. Since j is arbitrary, we must have $\tilde{S}_j = \tilde{S}_m$ for all $j \in \Omega$. In view of (1.4c) with $\tilde{I} = 0$, we obtain $\tilde{S}_j = N/n$ for $j \in \Omega$. ∎

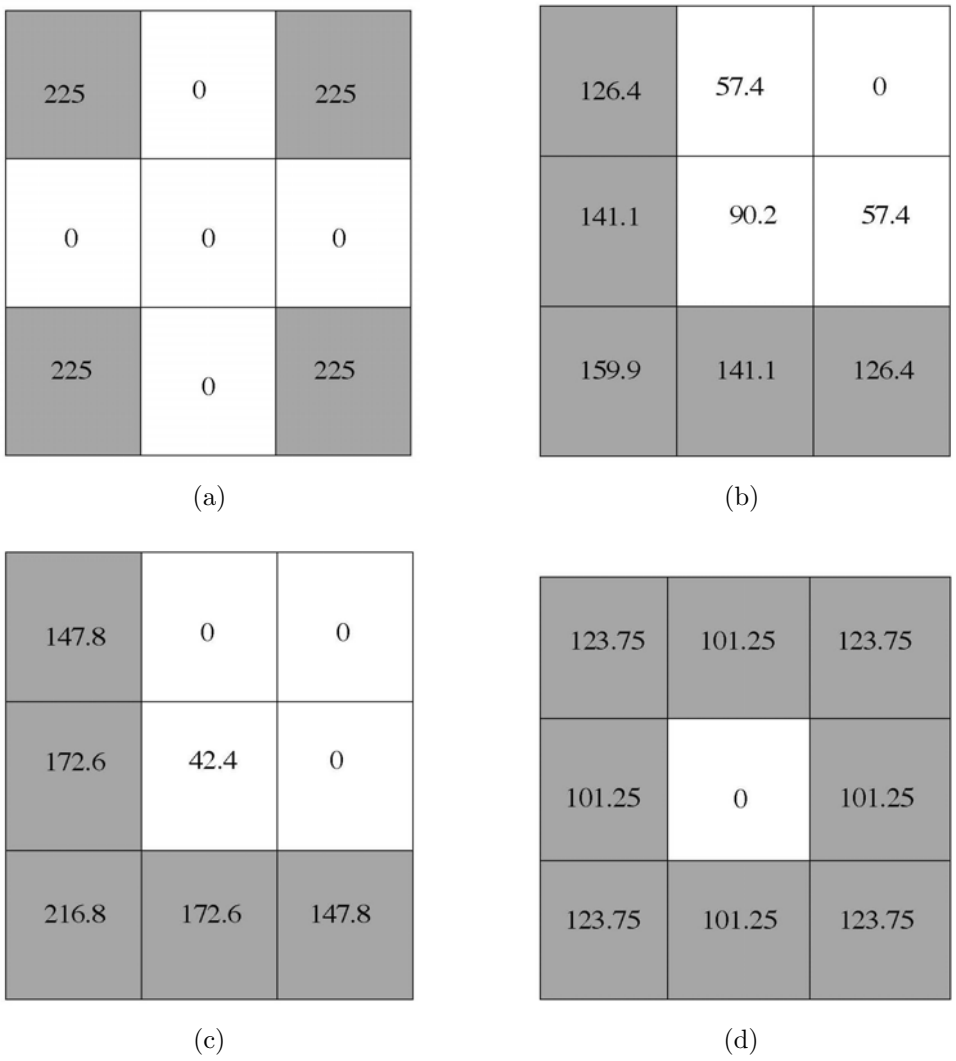


FIG. 2. *The limiting DFE under four parameter conditions.* (a) $\mathcal{R}_0 = 1.20$, $\mathcal{R}_0^{[5]} = 2$, $\mathcal{R}_0^{[j]} = 1.5$, *and* $\mathcal{R}_0^{[k]} = 0.5$ *for* $j = 2, 4, 6, 8$ *and* $k = 1, 3, 7, 9$; (b) $\mathcal{R}_0 = 1.51$, $\mathcal{R}_0^{[3]} = 3$, $\mathcal{R}_0^{[j]} = 1.25$, *and* $\mathcal{R}_0^{[k]} = 0.5$ *for* $j = 2, 5, 6$ *and* $k = 1, 4, 7, 8, 9$; (c) $\mathcal{R}_0 = 1.58$, $\mathcal{R}_0^{[3]} = 3$, $\mathcal{R}_0^{[j]} = 1.5$, *and* $\mathcal{R}_0^{[k]} = 0.5$ *for* $j = 2, 5, 6$ *and* $k = 1, 4, 7, 8, 9$; (d) $\mathcal{R}_0 = 1.03$, $\mathcal{R}_0^{[5]} = 3$, *and* $\mathcal{R}_0^{[j]} = 0.5$ *for* $j \neq 5$.

2.2. Stability of the DFE. Applying Lemma 2.1, we can calculate the basic reproduction number $\mathcal{R}_0$ for (1.1) using the next generation approach (Diekmann, Heesterbeek, and Metz (1990); Diekmann and Heesterbeek (2000); van den Driessche and Watmough (2002)). Since there are n patches, the basic reproduction number will be the spectral radius of an $n \times n$ nonnegative matrix. It is known that if $\mathcal{R}_0 < 1$, then the DFE is locally asymptotically stable, and if $\mathcal{R}_0 > 1$, then the DFE is unstable (van den Driessche and Watmough (2002)).

LEMMA 2.2. *The basic reproduction number for* (1.1) *is the spectral radius of the next generation matrix,*

$$\mathcal{R}_0 = \rho(FV^{-1}),$$

where $F = \operatorname{diag}(\beta_j)$ *and* $V = \operatorname{diag}(\gamma_j + d_I L_j) - d_I L$.

Proof. We can write (1.1b) as

$$\frac{d\bar{I}}{dt} = \mathcal{F} - \mathcal{V},$$

where $\mathcal{F}$ is the vector of new infections and $\mathcal{V}$ is the vector of transitions in the n infected states. Linearization of this system about the DFE yields

$$\frac{dx}{dt} = (F - V)x,$$

where F and V are the Jacobian matrices of $\mathcal{F}$ and $\mathcal{V}$, respectively, evaluated at the DFE. The eigenvalues of $(F - V)$ have negative real part if and only if $\mathcal{R}_0 = \rho(FV^{-1}) < 1$ (van den Driessche and Watmough (2002)). ∎

We now show that if $\mathcal{R}_0 < 1$, then the disease always becomes extinct; i.e., the DFE is globally asymptotically stable.

LEMMA 2.3. *If* $\mathcal{R}_0 < 1$, *then* $(\bar{S}, \bar{I}) \to (\hat{S}, 0)$ *as* $t \to \infty$.

Proof. Suppose that $\mathcal{R}_0 < 1$. We will use the comparison principle to show that $\bar{I}(t) \to 0$ as $t \to \infty$. To begin, observe from (1.1b) that

$$\frac{d\bar{I}_j}{dt} \le d_I \sum_{k\in\Omega} L_{jk}\bar{I}_k + (\beta_j - \gamma_j - d_I L_j)\bar{I}_j, \quad j \in \Omega,$$

or equivalently

$$\frac{d\bar{I}}{dt} \le (F - V)\bar{I},$$

where F and V are as in Lemma 2.2. The linear comparison system

$$\frac{dx}{dt} = (F - V)x, \quad x(0) = \bar{I}(0),$$

which is monotone, has eigenvalues with negative real part because $\mathcal{R}_0 < 1$ (van den Driessche and Watmough (2002)). Consequently, $x(t) \to 0$ as $t \to \infty$. By comparison, $\bar{I}(t) \to 0$ as $t \to \infty$. ∎

The global asymptotic stability of the DFE when $\mathcal{R}_0 < 1$ implies that there can be no EE in this case. In section 3, we consider what happens when $\mathcal{R}_0 > 1$.

3. The endemic equilibrium. Throughout this section, we again assume that (A1)–(A3) hold and that N is fixed.

3.1. Equivalent problems. It will be useful to consider several alternative statements of the equilibrium problem. We present here the first such equivalent problem.

LEMMA 3.1. *The pair* $(\tilde{S}, \tilde{I})$ *is a solution of* (1.4) *if and only if* $(\tilde{S}, \tilde{I})$ *is a solution of*

$$\kappa = d_S\tilde{S}_j + d_I\tilde{I}_j, \qquad j \in \Omega, \tag{3.1a}$$

$$0 = d_I \sum_{k\in\Omega} L_{jk}(\tilde{I}_k - \tilde{I}_j) + \tilde{I}_j\left(\beta_j - \gamma_j - \frac{\beta_j\tilde{I}_j}{\tilde{S}_j + \tilde{I}_j}\right), \quad j \in \Omega, \tag{3.1b}$$

$$N = \sum_{j\in\Omega}\left(\tilde{S}_j + \tilde{I}_j\right), \tag{3.1c}$$

where κ is some positive constant that is independent of $j \in \Omega$.

Proof. Suppose first that $(\tilde{S}, \tilde{I})$ is a solution of (1.4). We will show that there exists some $\kappa > 0$ such that $(\tilde{S}, \tilde{I})$ satisfies (3.1a). Summing (1.4a) and (1.4b) produces the relation

$$d_S \sum_{k\in\Omega} L_{jk}(\tilde{S}_k - \tilde{S}_j) + d_I \sum_{k\in\Omega} L_{jk}(\tilde{I}_k - \tilde{I}_j) = 0, \quad j \in \Omega.$$

We rearrange to get

$$\sum_{k\in\Omega} (L_{jk}/L_j) \left(d_S \tilde{S}_k + d_I \tilde{I}_k \right) = d_S \tilde{S}_j + d_I \tilde{I}_j, \quad j \in \Omega.$$

We can express this system of equations in matrix-vector form as

$$A \left(d_S \tilde{S} + d_I \tilde{I} \right) = d_S \tilde{S} + d_I \tilde{I},$$

where $A = (L_{jk}/L_j)$. Clearly, $A \geq 0$ because $L \geq 0$ and $L_j > 0$ for $j \in \Omega$. Moreover, since A and L are associated with the same adjacency matrix, it follows that A is irreducible. According to the Frobenius theorem (Gantmacher (1960), Theorem 2, p. 53), A has a largest (or *principal*) eigenvalue μ which is real, and μ has a one-dimensional eigenspace $\langle \psi \rangle$ for some positive eigenvector ψ. No other eigenvalue of A has a positive corresponding eigenvector. Since A is a stochastic matrix, the positive vector $x = (1, 1, \ldots, 1)^t$ is an eigenvector for A belonging to the eigenvalue 1. It follows from the remarks above that $\mu = 1$ and we may take $\psi = x$. As the vector $d_S \tilde{S} + d_I \tilde{I}$ is also an eigenvector for A belonging to the eigenvalue 1, we conclude that $d_S \tilde{S} + d_I \tilde{I} = \kappa \psi$ for some $\kappa \in \mathbb{R}$. Since $d_S \tilde{S}_j + d_I \tilde{I}_j > 0$ for at least one $j \in \Omega$ (because $N > 0$) it must be that $\kappa > 0$. Therefore $(\tilde{S}, \tilde{I})$ satisfies (3.1a) for some $\kappa > 0$. The fact that $(\tilde{S}, \tilde{I})$ satisfies (3.1b) and (3.1c) is clear by inspection. If $(\tilde{S}, \tilde{I})$ is a solution of (3.1) for some $\kappa > 0$, then it follows from a direct calculation that $(\tilde{S}, \tilde{I})$ satisfies (1.4). ∎

For our second equivalent formulation of the equilibrium problem, let

$$S_j = \frac{\tilde{S}_j}{\kappa} \quad \text{and} \quad I_j = \frac{d_I \tilde{I}_j}{\kappa}, \tag{3.2}$$

where κ is as in Lemma 3.1. Let $S = (S_j)$, $I = (I_j)$, and

$$f_j(u) = \beta_j \left(1 - \frac{d_S u}{d_I + (d_S - d_I)u} \right) - \gamma_j, \quad u \in [0, 1] \text{ and } j \in \Omega. \tag{3.3}$$

Observe that if $\beta_j > 0$, then f_j decreases from $\beta_j - \gamma_j$ to $-\gamma_j$ as u increases from 0 to 1. The next result follows from a direct calculation.

LEMMA 3.2. *The pair $(\tilde{S}, \tilde{I})$ is a solution of (3.1) if and only if (S, I) is a solution of*

$$1 = d_S S_j + I_j, \qquad j \in \Omega, \tag{3.4a}$$

$$0 = d_I \sum_{k\in\Omega} L_{jk}(I_k - I_j) + I_j f_j(I_j), \quad j \in \Omega, \tag{3.4b}$$

$$\kappa = \frac{d_I N}{\sum_{j\in\Omega} (d_I S_j + I_j)}. \tag{3.4c}$$

The benefit of this second formulation is that (3.4b) depends on I but not S. Thus, once I is determined, it is then a simple matter to determine S from (3.4a) and κ from (3.4c). Observe that κ is in a one-to-one correspondence with N.

3.2. An eigenvalue problem. The linear eigenvalue problem associated with (3.1b) at the DFE is

$$d_I \sum_{k\in\Omega} L_{jk}(\psi_k - \psi_j) + (\beta_j - \gamma_j)\psi_j + \lambda\psi_j = 0, \quad j \in \Omega. \tag{3.5}$$

Observe that (3.5) can be written as

$$d_I \sum_{k\in\Omega} L_{jk}\psi_k + (\beta_j + \theta - \gamma_j - d_I L_j)\psi_j = (\theta - \lambda)\psi_j, \quad j \in \Omega,$$

where $\theta = \max\{\gamma_j + d_I L_j : j \in \Omega\}$, and this equation can be written in the equivalent matrix-vector form $(d_I L + D)\psi = (\theta - \lambda)\psi$, where $D = \mathrm{diag}(\beta_j + \theta - \gamma_j - d_I L_j)$ and $\psi = (\psi_j)$. Thus, (λ, ψ) is a solution of (3.5) if and only if $(\mu, \psi) = (\theta - \lambda, \psi)$ is a solution of

$$Q\psi = \mu\psi, \tag{3.6}$$

where $Q = d_I L + D$.

LEMMA 3.3. *The matrix Q has all real eigenvalues, and it has a largest eigenvalue $\mu^* = \mu^*(d_I)$, which is positive. This eigenvalue μ^* has a one-dimensional eigenspace $\langle\phi\rangle$, where $\phi > 0$. Furthermore, no other eigenvalue of Q has a positive corresponding eigenvector.*

Proof. By construction, $Q_{jk} = d_I L_{jk} \geq 0$ for $j, k \in \Omega$ with $j \neq k$, and $Q_{jj} \geq d_I L_{jj} + \beta_j \geq 0$ for $j \in \Omega$. Therefore, Q is nonnegative. Moreover, Q is irreducible because Q and L are associated with adjacency matrices whose off-diagonal entries are the same. The stated properties of Q now follow from the symmetry of L (which implies that the eigenvalues of Q are real) and the Frobenius theorem (Gantmacher (1960)). ∎

LEMMA 3.4. *Define $\lambda^* = \lambda^*(d_I) = \theta - \mu^*(d_I)$ and let $\phi > 0$ be as in Lemma 3.3. Then*

(a) *λ^* is real and (λ^*, ϕ) satisfies (3.5), i.e.,*

$$d_I \sum_{k\in\Omega} L_{jk}(\phi_k - \phi_j) + (\beta_j - \gamma_j)\phi_j + \lambda^*\phi_j = 0, \quad j \in \Omega. \tag{3.7}$$

Moreover, (λ^*, ψ) satisfies (3.5) if and only if $\psi \in \langle\phi\rangle$. Finally, if (λ, ψ) satisfies (3.5) with $\lambda \neq \lambda^*$, then $\lambda > \lambda^*$ and $\psi_j \leq 0$ for some $j \in \Omega$.

(b) *λ^* is a strictly monotone increasing function of $d_I > 0$.*

(c) *$\lambda^* \to \min\{\gamma_j - \beta_j : j \in \Omega\}$ as $d_I \to 0$.*

(d) *$\lambda^* \to \frac{\Sigma_\gamma - \Sigma_\beta}{n}$ as $d_I \to \infty$.*

(e) *If $\Sigma_\beta \geq \Sigma_\gamma$, then $\lambda^* < 0$ for all $d_I > 0$.*

(f) *If $\Sigma_\beta < \Sigma_\gamma$, then the equation $\lambda^*(d_I) = 0$ has a unique positive root denoted by d_I^*. Furthermore, if $d_I < d_I^*$, then $\lambda^* < 0$, and if $d_I > d_I^*$, then $\lambda^* > 0$.*

The proof of Lemma 3.4 appears in Appendix B. In view of Lemma 3.4(e), (f), let us define $d_I^* = \infty$ when $\Sigma_\beta \geq \Sigma_\gamma$. We now connect λ^* to the basic reproduction number $\mathcal{R}_0$.

LEMMA 3.5. *Let $\mathcal{R}_0$ and λ^* be as in Lemmas 2.2 and 3.4, respectively. Then*

(a) *$\mathcal{R}_0 < 1$ if and only if $\lambda^* > 0$;*

(b) *$\mathcal{R}_0 > 1$ if and only if $\lambda^* < 0$.*

Proof. Observe from (3.7) that

$$(F - V)\phi + \lambda^* \phi = 0, \tag{3.8}$$

where F and V are defined as in Lemma 2.2. Also, since $F - V$ is symmetric, its eigenvalues are all real. Finally, recall from van den Driessche and Watmough (2002) that (i) $\mathcal{R}_0 < 1$ if and only if $F - V$ has all negative eigenvalues, and (ii) $\mathcal{R}_0 > 1$ if and only if $F - V$ has a positive eigenvalue.

(a) Suppose first that $\mathcal{R}_0 < 1$. We see from (3.8) that $(-\lambda^*)$ is an eigenvalue of $F - V$. Since $F - V$ has all negative eigenvalues, we obtain $\lambda^* > 0$. Now suppose that $\lambda^* > 0$. Equation (3.8) and Lemma 3.4(a) imply that $(-\lambda^*)$ is the largest eigenvalue of $F - V$. Thus, all the eigenvalues of $F - V$ are negative, and consequently $\mathcal{R}_0 < 1$.

(b) Suppose first that $\mathcal{R}_0 > 1$. Then $F - V$ has a positive eigenvalue μ. Equation (3.8) and Lemma 3.4(a) imply that $\lambda^* \le -\mu < 0$, i.e., $\lambda^* < 0$. Now suppose that $\lambda^* < 0$. We see from (3.8) that $\mu = -\lambda^*$ is a positive eigenvalue of $F - V$, and hence that $\mathcal{R}_0 > 1$. □

In the next section, we use λ^* and ϕ, rather than $\mathcal{R}_0$, to obtain the existence of an EE when the DFE is unstable.

3.3. Existence of an EE.

LEMMA 3.6. *Suppose that $\mathcal{R}_0 > 1$. Then* (3.4) *has a nonnegative solution (S, I), which can be chosen to satisfy $I \not\equiv 0$. Furthermore, this solution with $I \not\equiv 0$ is unique, $S > 0$, and $0 < I_j < 1$ for every $j \in \Omega$.*

Here we prove the existence of such an (S, I), and in the next section we will demonstrate that it is unique. Suppose that $\mathcal{R}_0 > 1$. In view of (3.4b), consider the related system of differential equations

$$\frac{dI_j}{dt} = G_j(I) \stackrel{\text{def}}{=} d_I \sum_{k \in \Omega} L_{jk}(I_k - I_j) + I_j f_j(I_j), \quad j \in \Omega. \tag{3.9}$$

First, I is a solution of (3.4b) if and only if $G(I) = 0$, where $G = (G_j)$. Second, (3.9) defines a monotone dynamical system because L_{jk} is nonnegative when $j \neq k$. It follows that if $\underline{I}$ and $\overline{I}$ are ordered (i.e., $\underline{I} \le \overline{I}$), and they are sub- and supersolutions of (3.9), respectively, i.e., $G(\underline{I}) \ge 0 \ge G(\overline{I})$, then there must exist some $I \in [\underline{I}, \overline{I}]$ such that $G(I) = 0$, where $[\underline{I}, \overline{I}] = \{I \in \mathbb{R}^n : \underline{I} \le I \le \overline{I}\}$ (Smith (1995)).

With $\phi > 0$ defined as in Lemma 3.3, we now show that $\underline{I} = \epsilon\phi$ and $\overline{I} = (1, 1, \ldots, 1)^t$ can be chosen as sub- and supersolutions for (3.9) if ϵ is chosen to be positive and sufficiently small. We may assume that ϕ is chosen so that $\sum_{j \in \Omega} \phi_j^2 = 1$. Lemma 3.5(b) implies that $\lambda^* < 0$. In view of (3.3), define

$$g(u) = \frac{d_S u}{d_I + (d_S - d_I)u}, \quad u \in [0, 1].$$

We remark that g increases from 0 to 1 as u increases from 0 to 1. Observe from (3.3) and (3.7) that

$$\begin{aligned} G_j(\underline{I}) &= d_I \sum_{k \in \Omega} L_{jk}(\epsilon\phi_k - \epsilon\phi_j) + \epsilon\phi_j f_j(\epsilon\phi_j) \\ &= \epsilon \left[d_I \sum_{k \in \Omega} L_{jk}(\phi_k - \phi_j) + (\beta_j - \gamma_j)\phi_j - \beta_j \phi_j g(\epsilon\phi_j) \right] \\ &= \epsilon\phi_j \left[-\lambda^* - \beta_j g(\epsilon\phi_j) \right] \end{aligned}$$

is positive for $j \in \Omega$ when $0 < \epsilon \ll 1$. Therefore, $\underline{I}$ is a subsolution of (3.9) for ϵ positive and sufficiently small. Next, since $G_j(\overline{I}) = f_j(1) = -\gamma_j$ is nonpositive for $j \in \Omega$, it follows that $\overline{I}$ is a supersolution of (3.9). Also, it is obvious that $\underline{I} \le \overline{I}$ if ϵ is chosen sufficiently small. We conclude from the remarks above that there must be an $I \in [\underline{I}, \overline{I}]$ with $G(I) = 0$. That is, there exists some I satisfying (3.4b) with $0 < I_j \le 1$ for $j \in \Omega$. We argue by contradiction to show that I_j cannot be equal to 1 for any $j \in \Omega$. If $I_j = 1$ for all $j \in \Omega$, then $G_j(I) = -\gamma_j < 0$ for $j \in H^-$, a contradiction. If $I_j = 1$ and $I_m < 1$ for some $j, m \in \Omega$, then there exists a chain from j to m, i.e., a sequence $j_1, j_2, \ldots, j_s \in \Omega$ with $j_1 = j$ and $j_s = m$ such that $L_{j_p j_{p+1}} > 0$ for $1 \le p \le s-1$. Thus, there exists some $k \in \Omega$ for which $I_{j_k} = 1$, $I_{j_{k+1}} < 1$, and $L_{j_k j_{k+1}} > 0$. But then $G_{j_k}(I) \le d_I L_{j_k j_{k+1}}(I_{j_{k+1}} - I_{j_k}) - \gamma_{j_k} < 0$, again a contradiction. We conclude that $0 < I_j < 1$ for $j \in \Omega$. In view of (3.4a), let us define S by $1 = d_S S + I$. Then $S > 0$. Consequently, (S, I) is a positive solution of (3.4) with $I_j < 1$ for $j \in \Omega$.

For sake of completeness, and also for the purpose of proving uniqueness in the next section, we now proceed to construct an iteration algorithm to find I. This algorithm is also used to generate the numerical plots appearing in Figure 2. Equation (3.4b) can be written equivalently as

$$-d_I \sum_{k \in \Omega} L_{jk}(I_k - I_j) = F_j(I_j), \quad j \in \Omega, \tag{3.10}$$

where $F_j(u) = u f_j(u)$. Let $j \in \Omega$. By inspection, the function f_j in (3.3) and its derivative f_j' are bounded for $u \in [0,1]$. We conclude that there exists some $M > 0$ (which can be chosen to be independent of j) such that $|F_j'(u)| < M$ for $u \in [0,1]$. It follows that $F_j'(u) + M > 0$ for $u \in [0,1]$. That is, $F_j(u) + Mu$ is a monotone increasing function of $u \in [0,1]$. Since $F_j'(u) = f_j(u) + u f_j'(u) \le f_j(u)$ for $u \in [0,1]$, it follows that $f_j(u) + M > 0$ for $u \in [0,1]$.

We now add MI_j to both sides of (3.10) to get

$$-d_I \sum_{k \in \Omega} L_{jk}(I_k - I_j) + MI_j = F_j(I_j) + MI_j, \quad j \in \Omega.$$

This equation inspires the vector iteration

$$-d_I \sum_{k \in \Omega} L_{jk}\big(I_k^{(l+1)} - I_j^{(l+1)}\big) + MI_j^{(l+1)} = F_j\big(I_j^{(l)}\big) + MI_j^{(l)}, \quad j \in \Omega, \tag{3.11}$$

with the index $l \ge 0$. This implicit scheme can be made explicit because the left-hand operator, which takes the form $A + M\mathcal{I}_n = M[(1/M)A + \mathcal{I}_n]$, is invertible for M sufficiently large. Let $I^{(l)} = \big(I_j^{(l)}\big)$. For our purposes, we will set $I^{(0)} = \underline{I} = \epsilon\phi$ with ϵ taken to be sufficiently small so that $\underline{I}$ is a subsolution of (3.9) satisfying $\underline{I} < \overline{I} = (1, 1, \ldots, 1)^t$. Similarly, we define the iteration

$$-d_I \sum_{k \in \Omega} L_{jk}\big(I_k^{[l+1]} - I_j^{[l+1]}\big) + MI_j^{[l+1]} = F_j\big(I_j^{[l]}\big) + MI_j^{[l]}, \quad j \in \Omega, \tag{3.12}$$

for $l \ge 0$ with $I^{[0]} = \overline{I}$, a supersolution of (3.9). We want to show that

$$\underline{I} = I^{(0)} \le I^{(1)} \le \cdots \le I^{(n)} \le \cdots \le I^{[n]} \le \cdots \le I^{[1]} \le I^{[0]} = \overline{I},$$

where the symbols surrounding the iteration index indicate the initial condition for the sequence. For convenience, we will refer to the sequences $I^{(l)}$ and $I^{[l]}$ as the *lower* and *upper sequences*, respectively.

It can be shown that all components of both sequences remain within the interval $[0,1]$, that the lower sequence is nondecreasing, that the upper sequence is nonincreasing, and that an iterate of the lower sequence is always less than or equal to the corresponding iterate of the upper sequence. Let $\Delta I^{(l)} = I^{(l+1)} - I^{(l)}$, $\Delta I^{[l]} = I^{[l+1]} - I^{[l]}$, and $\Delta I^{\{l\}} = I^{[l]} - I^{(l)}$.

LEMMA 3.7. *The following statements hold:*

(a) $I_j^{(l)}, I_j^{[l]} \in [0,1]$ *for* $l \geq 0$ *and* $j \in \Omega$*;*

(b) $\Delta I^{(l)} \geq 0$, $\Delta I^{[l]} \leq 0$, *and* $\Delta I^{\{l\}} \geq 0$ *for* $l \geq 0$.

The proof of this result appears in Appendix C. According to Lemma 3.7, the lower and upper sequences are both monotone and bounded. They also satisfy $\underline{I} \leq I^{(l)} \leq I^{[l]} \leq \overline{I}$ for $l \geq 0$. Therefore, there exist $I^{\min}$ and $I^{\max}$ such that $I^{(l)} \to I^{\min}$ and $I^{[l]} \to I^{\max}$ as $l \to \infty$. Clearly, $\underline{I} \leq I^{\min} \leq I^{\max} \leq \overline{I}$. Furthermore, since $I^{\min}$ is a fixed point for (3.11), and $I^{\max}$ is a fixed point for (3.12), each is a solution to (3.4b) with the property that $0 < I_j^{\min} \leq I_j^{\max} \leq 1$ for $j \in \Omega$. By an argument similar to the one given above, we obtain the stronger result that $0 < I_j^{\min} \leq I_j^{\max} < 1$ for $j \in \Omega$. In the next section, we show that $I_j^{\min} = I_j^{\max}$.

3.4. Uniqueness of the EE. Because we are interested only in those (S, I) that satisfy (3.4) with $S \geq 0$ and $I \geq 0$, we will assume throughout this section that if I is a solution to (3.4b), then $0 \leq I_j \leq 1$ for $j \in \Omega$.

LEMMA 3.8. *If* I *is a solution to* (3.4b)*, then either* $I \equiv 0$ *or* $I > 0$.

Proof. We argue by contradiction. Suppose that I is a solution of (3.4b) with $I \not\equiv 0$ and $I \not> 0$. Then there exist nonempty subsets K^- and K^+ of Ω with $I_j = 0$ for $j \in K^-$, $I_j > 0$ for $j \in K^+$, and $K^- \cup K^+ = \Omega$. Equation (3.4b) implies that $\sum_{k\in\Omega} L_{jk} I_k = 0$ for $j \in K^-$. The nonnegativity of L and I implies that $L_{jk} I_k = 0$ for $j \in K^-$ and $k \in \Omega$. Since $I_k > 0$ when $k \in K^+$, we must have $L_{jk} = 0$ when $j \in K^-$ and $k \in K^+$. But this contradicts the irreducibility of L. We conclude that either $I \equiv 0$ or $I > 0$. ∎

The following lemma justifies our referring to $I^{\min}$ and $I^{\max}$ as *minimal* and *maximal* solutions, respectively.

LEMMA 3.9. *If* I *is a positive solution to* (3.4b)*, then* $I \in [I^{\min}, I^{\max}]$.

Proof. Choose ϵ small enough so that $\underline{I} = I^{(0)} \leq I \leq I^{[0]} = \overline{I}$. Arguments similar to the one used in the proof of Lemma 3.7(b) show that $I^{(l)} \leq I \leq I^{[l]}$ for $l \geq 0$. The conclusion follows by letting $l \to \infty$. ∎

We now show that if two positive solutions of (3.4b) are ordered, then they are either strictly ordered or they are equal.

LEMMA 3.10. *If* I^- *and* I^+ *are positive solutions to* (3.4b) *with* $I^- \leq I^+$*, then either* $I^- < I^+$ *or* $I^- \equiv I^+$.

Proof. We argue by contradiction. Suppose that $I^- = (I_j^-)$ and $I^+ = (I_j^+)$ are positive solutions to (3.4b) with $I^- \leq I^+$, and that neither $I^- < I^+$ nor $I^- \equiv I^+$. Then there exist nonempty and disjoint subsets K^- and K^+ of Ω, whose union forms all of Ω, and with the property that $I_j^- < I_j^+$ for $j \in K^-$ and $I_j^- = I_j^+$ for $j \in K^+$. We subtract (3.4b) with $I = I^-$ from (3.4b) with $I = I^+$, and use the fact that $I_j^- = I_j^+$ for $j \in K^+$, to get $\sum_{k\in K^-} L_{jk}(I_k^+ - I_k^-) = 0$ for $j \in K^+$. We sum only over $k \in K^-$ because $I_k^+ = I_k^-$ for $k \in K^+$. By definition, the expression $I_k^+ - I_k^-$ is positive for $k \in K^-$. Consequently, $L_{jk} = 0$ for $j \in K^+$ and $k \in K^-$. But this contradicts the irreducibility of L. We conclude that either $I^- < I^+$ or $I^- \equiv I^+$. ∎

LEMMA 3.11. *If* I^* *and* I^{**} *are positive solutions to* (3.4b)*, then* $I^* \equiv I^{**}$.

Proof. We argue by contradiction. Suppose that I^* and I^{**} are positive solutions to (3.4b) with $I^* \not\equiv I^{**}$. Then $I^*, I^{**} \in [I^{\min}, I^{\max}]$, by Lemma 3.9. Since $I^* \not\equiv I^{**}$, it follows that $I^{\min} \not\equiv I^{\max}$. We conclude from the relation $I^{\min} \le I^{\max}$ and Lemma 3.10 that $I^{\min} < I^{\max}$. So, without loss of generality, we may assume that $I^* < I^{**}$, for otherwise we may replace I^* with $I^{\min}$ and I^{**} with $I^{\max}$. We substitute $I^* = (I_j^*)$ and $I^{**} = (I_j^{**})$ individually into (3.4b) to get

$$d_I \sum_{k\in\Omega} L_{jk}(I_k^* - I_j^*) + I_j^* f_j(I_j^*) = 0, \quad j \in \Omega,$$
$$d_I \sum_{k\in\Omega} L_{jk}(I_k^{**} - I_j^{**}) + I_j^{**} f_j(I_j^{**}) = 0, \quad j \in \Omega.$$

We multiply both sides of the first equation by I_j^{**} and both sides of the second equation by I_j^*, subtract the resulting equations, and then sum over all $j \in \Omega$ to get

$$d_I \sum_{j,k\in\Omega} L_{jk}\left[I_j^{**} I_k^* - I_j^* I_k^{**}\right] + \sum_{j\in\Omega} I_j^* I_j^{**}\left[f_j(I_j^*) - f_j(I_j^{**})\right] = 0.$$

The symmetry of L implies that the first sum vanishes, and the second sum is nonnegative because $I_j^* I_j^{**} > 0$ and $f_j(I_j^*) \ge f_j(I_j^{**})$ for $j \in \Omega$. The fact that $\beta_k > 0$ for $k \in H^+$ implies that $f_k(I_k^*) > f_k(I_k^{**})$, and thus the second sum is in fact positive, a contradiction. We conclude that $I^* \equiv I^{**}$. ∎

Lemmas 3.9 and 3.11 imply that (3.4b) has a unique positive solution given by $I \stackrel{\text{def}}{=} I^{\min} = I^{\max}$. We conclude from Lemma 3.8 that I is the only nonnegative solution of (3.4b) satisfying $I \not\equiv 0$. We have completed the proof of Lemma 3.6. The next result follows from Lemmas 3.2 and 3.6 and (3.2).

LEMMA 3.12. *Suppose that $\mathcal{R}_0 > 1$. Then* (1.4) *has a nonnegative solution $(\tilde{S}, \tilde{I})$ which satisfies $\tilde{I} \not\equiv 0$. Furthermore, this solution is unique; it is given by $(\tilde{S}, \tilde{I}) = (\kappa S, \kappa I/d_I)$, where κ is as in* (3.4c)*; and $\tilde{I} > 0$.*

We have shown that a unique EE exists when $\mathcal{R}_0 > 1$ and that it satisfies $\tilde{I} > 0$. In the next section, we consider the asymptotic behavior of the EE as $d_S \to 0$.

4. Asymptotic behavior of the endemic equilibrium. Throughout this section, we still assume that (A1)–(A3) hold and that N is fixed. We also assume that $\mathcal{R}_0 > 1$, so that Lemma 3.6 for (S, I) and Lemma 3.12 for $(\tilde{S}, \tilde{I})$ always apply.

4.1. The limiting DFE. Observe that $\tilde{S}$, $\tilde{I}$, and κ are all functions of d_S in (1.4) and (3.1). First, we determine the asymptotic behavior of $\tilde{I}$ and κ.

LEMMA 4.1. *As $d_S \to 0$, $\kappa \to 0$ and $\tilde{I} \to 0$.*

Proof. We first show that $\kappa \to 0$ as $d_S \to 0$. Let $j \in H^-$ and $\hat{I}_j$ be a limit point of $\tilde{I}_j$ as $d_S \to 0$. Equation (1.4a) and the nonnegativity of β_j, $\tilde{S}_j$, and $\tilde{I}_j$ imply that

$$d_S \sum_{k\in\Omega} L_{jk}(\tilde{S}_k - \tilde{S}_j) \le \tilde{I}_j\,(\beta_j - \gamma_j)\,.$$

Since $\tilde{S}_k \in [0, N]$ for $k \in \Omega$, it follows that the left-hand side vanishes as $d_S \to 0$. Since $\beta_j < \gamma_j$, it must be that $\hat{I}_j \le 0$. But $\hat{I}_j \ge 0$ because $\tilde{I}_j > 0$ for $d_S > 0$. We conclude that $\hat{I}_j = 0$. Thus, $\tilde{I}_j \to 0$ as $d_S \to 0$ for all $j \in H^-$. Let $k \in H^-$ be fixed. Equation (3.1a) implies that $\kappa = d_S \tilde{S}_k + d_I \tilde{I}_k$. The product $d_S \tilde{S}_k \to 0$ as $d_S \to 0$ because $\tilde{S}_k \in [0, N]$, and $d_I \tilde{I}_k \to 0$ as $d_S \to 0$ by the argument above. Therefore, $\kappa \to 0$ as $d_S \to 0$.

We now show that $\tilde{I} \to 0$. Let $j \in \Omega$. Again, (3.1a) specifies that $\kappa = d_S\tilde{S}_j + d_I\tilde{I}_j$. The left-hand side vanishes as $d_S \to 0$ by part (a). The product $d_S\tilde{S}_j \to 0$ as $d_S \to 0$ because $\tilde{S}_j \in [0, N]$. We conclude that $\tilde{I}_j \to 0$ as $d_S \to 0$. ∎

So that we may determine the asymptotic behavior of $\tilde{S}$, we first consider I in (3.2) as a function of d_S.

LEMMA 4.2. *I_j is a monotone decreasing function of d_S for each $j \in \Omega$.*

Proof. Suppose that $0 < d_{S_1} < d_{S_2}$, and let I^1 and I^2 be corresponding solutions to (3.4b) with $0 < I^1_j, I^2_j < 1$ for $j \in \Omega$. Then

$$d_I \sum_{k\in\Omega} L_{jk}(I^1_k - I^1_j) + I^1_j f_j(I^1_j, d_{S_1}) = 0, \quad j \in \Omega, \tag{4.1a}$$

$$d_I \sum_{k\in\Omega} L_{jk}(I^2_k - I^2_j) + I^2_j f_j(I^2_j, d_{S_2}) = 0, \quad j \in \Omega, \tag{4.1b}$$

where

$$f_j(u, d_S) = \beta_j \left(1 - \frac{d_S u}{d_S u + d_I(1-u)}\right) - \gamma_j, \quad u \in [0,1] \text{ and } j \in \Omega.$$

It is easy to see that $\partial f_j/\partial d_S \le 0$. It follows from this fact and (4.1b), with d_{S_1} in place of d_{S_2}, that

$$d_I \sum_{k\in\Omega} L_{jk}(I^2_k - I^2_j) + I^2_j f_j(I^2_j, d_{S_1}) \ge 0, \quad j \in \Omega.$$

Thus, I^2 is a subsolution of (4.1a). Again, $\overline{I} = (1, 1, \ldots, 1)^t$ is a supersolution of (4.1a). Also, $I^2 < \overline{I}$. By the iteration method presented in sections 3.3 and 3.4, (4.1a) has a unique solution $I^1 \in [I^2, \overline{I}]$. We conclude that $I^1 \ge I^2$. ∎

Recall that $0 < I_j < 1$ for each $j \in \Omega$. The above lemma implies that for every $j \in \Omega$ there exists some I^*_j such that as $d_S \to 0$,

$$I_j \to I^*_j \quad \text{and} \quad 0 < I^*_j \le 1. \tag{4.2}$$

Let $I^* = (I^*_j)$. It remains to establish conditions under which $0 < I^*_j < 1$ or $I^*_j = 1$. Let

$$J^- = \{j \in \Omega : 0 < I^*_j < 1\} \quad \text{and} \quad J^+ = \{j \in \Omega : I^*_j = 1\}.$$

Observe that $J^- \cup J^+ = \Omega$. We will need to know that J^- is nonempty.

LEMMA 4.3. *$H^- \subseteq J^-$.*

Proof. We argue by contradiction. Suppose that there exists some $j \in H^-$ with $j \in J^+$. Then $\beta_j < \gamma_j$ and $I^*_j = 1$. In view of (3.2), we multiply both sides of (3.1b) by d_I/κ and drop the nonnegative term $\beta_j\tilde{I}_j/(\tilde{S}_j + \tilde{I}_j)$ to get

$$d_I \sum_{k\in\Omega} L_{jk}(I_k - I_j) + I_j(\beta_j - \gamma_j) \ge 0.$$

Letting $d_S \to 0$ on both sides yields

$$d_I \sum_{k\in\Omega} L_{jk}(I^*_k - 1) + \beta_j - \gamma_j \ge 0.$$

The negativity of $\beta_j - \gamma_j$ implies that

$$\sum_{k\in\Omega} L_{jk}(I_k^* - 1) > 0.$$

But this inequality contradicts (4.2). We conclude that if $j \in H^-$, then $j \in J^-$. ∎

We are now in a position to determine the asymptotic behavior of $\tilde{S}_j$.

LEMMA 4.4. *The following statements hold:*

(a) $\kappa/d_S \to N^* \stackrel{\text{def}}{=} N/\sum_{j\in\Omega}(1 - I_j^*)$ *as* $d_S \to 0$;

(b) $\tilde{S} \to S^*$ *as* $d_S \to 0$, *where* $S_j^* \stackrel{\text{def}}{=} (1 - I_j^*)N^*$;

(c) $S^* \geq 0$ *and* $\sum_{j\in\Omega} S_j^* = N$.

Proof.

(a) Equations (3.1a), (3.1c), and (3.2) imply that

$$N = \sum_{j\in\Omega}\left(\frac{\kappa - d_I\tilde{I}_j}{d_S}\right) + \sum_{j\in\Omega}\tilde{I}_j = \frac{\kappa}{d_S}\sum_{j\in\Omega}(1 - I_j) + \sum_{j\in\Omega}\tilde{I}_j.$$

Lemma 4.1 and (4.2) imply that

$$\frac{\kappa}{d_S} \to \frac{N}{\sum_{j\in\Omega}(1 - I_j^*)} \quad \text{as} \quad d_S \to 0.$$

This limit is well defined because J^- is nonempty.

(b) Again, (3.1a) and (3.2) imply that

$$\tilde{S}_j = \frac{\kappa - d_I\tilde{I}_j}{d_S} = (1 - I_j)\frac{\kappa}{d_S}.$$

Equation (4.2) and part (a) imply that $\tilde{S}_j \to (1 - I_j^*)N^*$ as $d_S \to 0$.

(c) This part follows immediately from parts (a) and (b), the positivity of N, and (4.2). ∎

4.2. The limiting DFE on high-risk patches. Observe from Lemma 4.4(b) that $S^* > 0$ on J^- and $S^* \equiv 0$ on J^+. We know from Lemma 4.3 that J^- is nonempty because it contains H^-. Next we show that J^+, which is a subset of H^+, is also nonempty.

LEMMA 4.5. J^+ *is nonempty.*

Proof. We argue by contradiction. Suppose that J^+ is empty, i.e., $J^- = \Omega$. Multiply both sides of (3.1b) by d_I/κ to get

$$d_I\sum_{k\in\Omega} L_{jk}(I_k - I_j) + I_j\left(\beta_j - \gamma_j - \frac{\beta_j\tilde{I}_j}{\tilde{S}_j + \tilde{I}_j}\right) = 0, \quad j \in \Omega. \tag{4.3}$$

Since $I_j^* \in (0, 1)$ for $j \in \Omega$, we have from (3.4a) and (4.2) that $S_j = (1 - I_j)/d_S \to \infty$ as $d_S \to 0$ for $j \in \Omega$. It follows from this fact and (3.2) that

$$\frac{\beta_j\tilde{I}_j}{\tilde{S}_j + \tilde{I}_j} = \frac{\beta_j I_j}{d_I S_j + I_j} \to 0$$

as $d_S \to 0$ for $j \in \Omega$. Letting $d_S \to 0$ in (4.3), we get

$$d_I\sum_{k\in\Omega} L_{jk}(I_k^* - I_j^*) + I_j^*(\beta_j - \gamma_j) = 0, \quad j \in \Omega.$$

Thus $(\lambda, \psi) = (0, I^*)$ satisfies (3.5). Since $I^* > 0$, we obtain from Lemma 3.4(a) that $\lambda^* = 0$. But this contradicts Lemma 3.5(a). We conclude that J^+ is nonempty. ∎

Next, we determine a condition under which J^+ is as large as it can be.

LEMMA 4.6. *If condition* (1.6) *holds, then* $J^+ = H^+$.

Proof. Recall that $J^+ \subseteq H^+$. We argue by contradiction to show that if condition (1.6) holds, then $H^+ \subseteq J^+$. Suppose that condition (1.6) holds and that there exists some $j \in H^+$ with the property that $j \in J^-$. Without loss of generality, we may assume that $I_j^* = \min\{I_k^* : k \in H^+\}$. Choose $m \in H^-$ so that $I_m^* = \min\{I_k^* : k \in H^-\}$. Letting $d_S \to 0$ in (3.4b) implies that

$$d_I \sum_{k \in \Omega} L_{jk}(I_k^* - I_j^*) + I_j^*(\beta_j - \gamma_j) = 0.$$

Here we used the fact that $0 < I_j^* < 1$. Since $\Omega = H^- \cup H^+$, we obtain

$$d_I \sum_{k \in H^+} L_{jk}(I_k^* - I_j^*) + d_I \sum_{k \in H^-} L_{jk} I_k^* + I_j^*(\beta_j - \gamma_j - d_I L_j^-) = 0.$$

The minimality of I_j^* over H^+ and I_m^* over H^- implies that

$$(d_I L_j^-) I_m^* \le I_j^*(\gamma_j - \beta_j + d_I L_j^-). \tag{4.4}$$

A similar argument shows that

$$(d_I L_m^+) I_j^* \le I_m^*(\gamma_m - \beta_m + d_I L_m^+). \tag{4.5}$$

We multiply corresponding sides of (4.4) and (4.5) together and simplify to get

$$(\gamma_j - \beta_j)(\gamma_m - \beta_m) + (\gamma_j - \beta_j) d_I L_m^+ + (\gamma_m - \beta_m) d_I L_j^- \ge 0.$$

We divide both sides by $d_I(\gamma_j - \beta_j)(\gamma_m - \beta_m)$, which is negative, and rearrange to get

$$\frac{1}{d_I} \le \frac{L_j^-}{\beta_j - \gamma_j} + \frac{L_m^+}{\beta_m - \gamma_m} \le \max_{j \in H^+} \left[\frac{L_j^-}{\beta_j - \gamma_j} \right] + \max_{k \in H^-} \left[\frac{L_k^+}{\beta_k - \gamma_k} \right].$$

This contradicts (1.6). We conclude that $H^+ \subseteq J^+$, and therefore that $H^+ = J^+$. ∎

Finally, we determine a condition under which J^+ is a proper subset of H^+.

LEMMA 4.7. *If condition* (1.7) *holds for some* $j \in H^+$, *then* $j \in J^-$.

Proof. We argue by contradiction. Suppose that condition (1.7) holds for some $p \in H^+$, and that $p \in J^+$. Choose $m \in H^-$ such that $I_m^* = \max\{I_k^* : k \in H^-\} < 1$. We let $d_S \to 0$ in (3.4b) with $j = m$ to get

$$0 = d_I \sum_{k \in \Omega} L_{mk}(I_k^* - I_m^*) + I_m^*(\beta_m - \gamma_m).$$

We rearrange to get

$$0 = d_I \sum_{k \in H^+} L_{mk} I_k^* + d_I \sum_{k \in H^-} L_{mk} I_k^* + I_m^*(\beta_m - \gamma_m - d_I L_m).$$

The upper bound of 1 on I_k^* and the maximality of I_m^* imply that

$$0 \le d_I L_m^+ + I_m^*(\beta_m - \gamma_m + d_I L_m^- - d_I L_m).$$

The relation $L_m^+ + L_m^- = L_m$ implies that

$$I_m^*(\gamma_m - \beta_m + d_I L_m^+) \le d_I L_m^+.$$

The positivity of $\gamma_m - \beta_m$ implies that

$$I_m^* \le \frac{d_I L_m^+}{\gamma_m - \beta_m + d_I L_m^+}. \tag{4.6}$$

Letting $d_S \to 0$ in (3.4b), we get

$$0 \le d_I \sum_{k\in\Omega} L_{jk}(I_k^* - I_j^*) + I_j^*(\beta_j - \gamma_j), \quad j \in \Omega.$$

Again, we rearrange to get

$$0 \le d_I \sum_{k\in H^+} L_{jk} I_k^* + d_I \sum_{k\in H^-} L_{jk} I_k^* + I_j^*(\beta_j - \gamma_j - d_I L_j), \quad j \in \Omega.$$

In particular, if $j = p$, then

$$0 \le d_I L_p^+ + d_I \sum_{k\in H^-} L_{pk} I_k^* + (\beta_p - \gamma_p - d_I L_p).$$

The relation $L_p^+ + L_p^- = L_p$ and the maximality of I_m^* imply that

$$0 \le d_I(I_m^* - 1)L_p^- + \beta_p - \gamma_p.$$

Equation (4.6) implies that

$$0 \le d_I \left(\frac{d_I L_m^+}{\gamma_m - \beta_m + d_I L_m^+} - 1 \right) L_p^- + \beta_p - \gamma_p.$$

We rearrange to get

$$\frac{\beta_p - \gamma_p}{d_I} \ge \left(\frac{\gamma_m - \beta_m}{\gamma_m - \beta_m + d_I L_m^+} \right) L_p^-.$$

The positivity of $\gamma_m - \beta_m$ and $\beta_p - \gamma_p$ implies that

$$\frac{1}{d_I} \ge \frac{L_p^-}{\beta_p - \gamma_p} + \frac{L_m^+}{\beta_m - \gamma_m} \ge \frac{L_p^-}{\beta_p - \gamma_p} + \min_{k\in H^-} \left[\frac{L_k^+}{\beta_k - \gamma_k} \right].$$

But this contradicts (1.7) with $j = p$. We conclude that if (1.7) holds for some $p \in H^+$, then $p \in J^-$. □

5. Discussion. We first mention some limiting cases for which we can simplify the expression for the basic reproduction number $\mathcal{R}_0$. We next state some open problems that relate to our work, and then finish with some concluding remarks.

5.1. Limiting cases. In the general case, we will not be able to obtain a simple expression for $\mathcal{R}_0$. However, we can compute an explicit expression for $\mathcal{R}_0$ in special cases.

In two limiting cases for n patches, the expression for $\mathcal{R}_0$ can be simplified. The basic reproduction number tends to the maximum ratio of the transmission rate to the recovery rate as infected movement becomes arbitrarily small ($\mathcal{R}_0 \to \max\{\beta_j/\gamma_j : j \in \Omega\}$ as $d_I \to 0$), and it tends to the average transmission rate divided by the average recovery rate as infected movement becomes arbitrarily large ($\mathcal{R}_0 \to \Sigma_\beta/\Sigma_\gamma$ as $d_I \to \infty$). The latter limit can be verified for a given value of n by calculating the limit V_∞^{-1} of V^{-1} as $d_I \to \infty$ (McCormack (2006)). For example, if $n = 2$ patches, then $F = \left(\begin{smallmatrix} \beta_1 & 0 \\ 0 & \beta_2 \end{smallmatrix}\right)$ and $V_\infty^{-1} = \frac{1}{\gamma_1+\gamma_2}\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, so that the limiting value of $\mathcal{R}_0$ is $\rho(FV_\infty^{-1}) = \frac{\beta_1+\beta_2}{\gamma_1+\gamma_2}$.

In the general case that $n = 2$, then $\mathcal{R}_0$ in Lemma 2.2 becomes

$$\mathcal{R}_0 = \frac{\beta_2\gamma_1 + \beta_1\gamma_2 + d_I\ell(\beta_1+\beta_2) + \sqrt{[\beta_2\gamma_1 - \beta_1\gamma_2 + d_I\ell(\beta_2-\beta_1)]^2 + (2d_I\ell)^2\beta_1\beta_2}}{2[\gamma_1\gamma_2 + d_I\ell(\gamma_1+\gamma_2)]},$$

where $\ell = L_{12} = L_{21}$. In this case, the condition $\mathcal{R}_0 < 1$ is equivalent to the following conditions:

$$(\beta_1 - \gamma_1 - d_I\ell) + (\beta_2 - \gamma_2 - d_I\ell) < 0,$$
$$(\beta_1 - \gamma_1 - d_I\ell)(\beta_2 - \gamma_2 - d_I\ell) - (d_I\ell)^2 > 0.$$

That is, the eigenvalues of $F - V$ are negative if and only if the conditions above are satisfied.

5.2. Open problems. Some open mathematical questions remain in connection with this research.

- We conjecture that the basic reproductive number $\mathcal{R}_0$ is a monotone decreasing function of d_I. The difficulty in showing this directly is that, although V^{-1} is a function of d_I, a general expression for V^{-1} is not simple. If this conjecture is true, then $\max\{\beta_j/\gamma_j : j \in \Omega\}$ and $\Sigma_\beta/\Sigma_\gamma$ are upper and lower bounds, respectively, on $\mathcal{R}_0$.
- We did not prove stability of the EE in Theorem 1. We conjecture that the EE globally attracts all solutions of (1.1) satisfying (1.2), and numerical simulations suggest that this is indeed the case. The uniform persistence of $\bar{I} = (\bar{I}_j)$ in (1.1) when the DFE is unstable has recently been established by Dhirasakdanon, Thieme, and van den Driessche (2007).
- We do not yet fully understand the asymptotic behavior of the EE as $d_S \to 0$. For example, condition (1.6) is not necessary for S_j^* to be zero on a patch. In Figure 2(c) susceptibles can persist on only one of four high-risk patches. In this case, condition (1.7) is satisfied for $j = 5$,

$$1 = \frac{1}{d_I} < \frac{L_5^-}{\beta_5 - \gamma_5} + \min_{k\in H^-}\left\{\frac{L_k^+}{\beta_k - \gamma_k}\right\} = \frac{2}{0.5} - \frac{1}{0.5} = 2,$$

but neither condition (1.6) nor (1.7) is satisfied for $j = 2, 3, 6$. Similarly, condition (1.7) is probably not necessary for S_j^* to be positive on a patch. It is easy to see that if some high-risk patch ($j \in H^+$) is not directly connected to any low-risk patches ($L_j^- = 0$), then condition (1.7) cannot be satisfied because $L_k^+/(\beta_k - \gamma_k)$ is nonpositive for $k \in H^-$. However, this does not rule

out the possibility that $S_j^* > 0$. In general, an open problem is to determine the distribution of high-risk patches for which S^* is either positive or zero when neither condition (1.6) nor (1.7) is satisfied. In Figure 2(d), susceptibles cannot persist on the single high-risk patch, and neither condition (1.6) nor (1.7) is satisfied:

$$\max_{j\in H^+}\left\{\frac{L_j^-}{\beta_j-\gamma_j}\right\} = \frac{L_5^-}{\beta_5-\gamma_5} = \frac{4}{2} = 2,$$

$$\min_{j\in H^-}\left\{\frac{L_j^+}{\beta_j-\gamma_j}\right\} = -\frac{1}{0.5} = -2, \quad \text{and} \quad \max_{j\in H^-}\left\{\frac{L_j^+}{\beta_j-\gamma_j}\right\} = 0.$$

As always, there are biological realities that we did not take into account, but which ecologists have suggested are important determinants of community dynamics.

- We neglect *population dynamics* (births or deaths) within the patches. At a very crude level, we can either ignore these dynamics on the grounds that epidemic dynamics often occur on a faster time scale than host demography, or we can say heuristically that death of an infected individual and subsequent replacement by a susceptible (in the absence of vertical transmission) is equivalent to a recovery event. Of course, either of these claims is an approximation, and it remains to be seen whether the results would be sensitive to such details.
- *Death during movement* may occur, especially if patches are separated by hostile "matrix" habitat. Adding this phenomenon might simply mean that some component of mortality (corresponding to a loss of infective potential, as argued in the previous point) scaled with d_I.
- *Density-dependent movement*—typically increasing rates of movement at higher population density, by organisms seeking to avoid competition—is generally an important factor in determining the behavior of spatial population dynamics models (Amarasekare (2004)).

These factors suggest possible directions for future exploration.

5.3. Mathematical and biological conclusions. Some of the relationships and techniques applied here have been applied by others. The relationship between the high rate of movement for infectives ($d_I > d_I^*$) and the basic reproduction number ($\mathcal{R}_0 < 1$) was noted by Salmani and van den Driessche (2006) in a two-patch SIS epidemic model. In addition, global stability of the DFE using comparison or monotone techniques has been applied by others (Arino et al. (2005); Arino and van den Driessche (2006); Wang and Mulone (2003); Wang and Zhao (2004)).

Our new results relate spatial heterogeneity, habitat connectivity, and rates of movement to disease persistence and extinction. We showed for populations with low mobility of susceptibles ($d_S \approx 0$) and moderate mobility of infectives ($0 \ll d_I < d_I^*$) that disease prevalence is very low ($\tilde{I} \approx 0$) in a spatial environment that includes both low-risk and high-risk patches. These results may have implications for disease control. If the environment is low-risk, but infectives move a lot, the disease may die out; conversely, restricting movement of infectives among patches (e.g., by habitat fragmentation) may allow the disease to persist and/or reemerge. In contrast, if a high-risk spatial environment can be modified to include low-risk patches (i.e., low transmission rates or high recovery rates) and if the mobility of susceptible individuals can be restricted, then it may be possible to eliminate the disease. In epidemiology,

quarantine attempts to prevent infected individuals from moving into a patch with a susceptible population; a *cordon sanitaire* attempts to restrict the movement of infected individuals out of a restricted area. The control strategy suggested by these results most closely resembles the movement restrictions imposed on *all* individuals (susceptible as well as infective) during the 2001 foot-and-mouth disease virus epidemic in Britain.

In a broader sense, these results fall under the ecological rubric of *source-sink dynamics*—population dynamics in heterogeneous environments with both "good" and "bad" patches (in our terminology, high- or low-risk patches, depending on whether we mean "good for the host" or "good for the disease"). The initially counterintuitive result that movement of infectives leads to disease extinction in a high-risk environment, which seems at odds with the idea of preventing disease from spreading between high-risk core groups and the general population (Jacquez, Simon, and Koopman (1995)), or between patches in a metapopulation (Hess (1996)), makes sense when we consider that (unlike in the core-group example), high infection rates are a property of the environment rather than of the individual. Ecologists usually want to prevent the extinction of threatened species; in contrast, epidemiologists want to promote the extinction of disease. However, ecologists have explored a broad range of questions, including evolutionary dynamics (Gomulkiewicz, Holt, and Barfield (1995)) and community structure (Namba and Hashimoto (2004)), in the context of source-sink dynamics. In the long run, linking the mathematical analyses of theoretical epidemiological and ecological models in heterogeneous landscapes can lead to broader mathematical and biological insights.

Appendix A. The irreducibility of L implies that, given any $j, k \in \Omega$ with $j \neq k$, there exists a distinct sequence $j_1, j_2, \ldots, j_s \in \Omega$, with $j_1 = j$ and $j_s = k$, such that $L_{j_p j_{p+1}} > 0$ for $1 \leq p \leq s-1$ (Seneta (1973), Exercise 1.3). We call such a sequence a *chain* from j to k. Second, the irreducibility of L implies that there exists no nonempty proper subset K of Ω with the property that $L_{jk} = 0$ for $j \in K$ and $k \notin K$ (Bapat and Raghavan (1997), Lemma 1.1.1). Finally, L is associated with an adjacency matrix $B = (B_{jk})$ for which $B_{jk} = 1$ if $L_{jk} > 0$ and $B_{jk} = 0$ if $L_{jk} = 0$. If the corresponding adjacency matrix for another nonnegative matrix A has the same off-diagonal entries as B, then A is also irreducible (Ortega (1987)).

Appendix B.

Proof of Lemma 3.4.

(a) As θ and μ^* are both real, so is λ^*. The fact that (μ^*, ϕ) is a solution of (3.6) implies that (λ^*, ϕ) is a solution of (3.5). Since (λ^*, ψ) satisfies (3.5) if and only if (μ^*, ψ) satisfies (3.6), and (μ^*, ψ) is a solution of (3.6) if and only if $\psi \in \langle \phi \rangle$, it follows that (λ^*, ψ) is a solution of (3.5) if and only if $\psi \in \langle \phi \rangle$. Suppose that (λ, ψ) satisfies (3.5) with $\lambda \neq \lambda^*$. Then (μ, ψ) satisfies (3.6) with $\mu = \theta - \lambda \neq \theta - \lambda^* = \mu^*$. Lemma 3.3 implies that $\mu < \mu^*$ and $\psi_j \leq 0$ for some $j \in \Omega$. We conclude that $\lambda = \theta - \mu > \theta - \mu^* = \lambda^*$.

(b) Observe from (3.7) that both ϕ and λ^* are functions of d_I. Both ϕ and λ^* are, in fact, differentiable functions of d_I by the implicit function theorem. We differentiate both sides of (3.7) by d_I to obtain

$$\sum_{k \in \Omega} L_{jk}(\phi_k - \phi_j) + d_I \sum_{k \in \Omega} L_{jk}\left(\phi_k' - \phi_j'\right) + (\beta_j - \gamma_j + \lambda^*)\phi_j' + (\lambda^*)'\phi_j = 0, \qquad j \in \Omega.$$

It suffices to show that $(\lambda^*)' > 0$. We multiply both sides of the equation above by ϕ_j and sum over all $j \in \Omega$ to get

$$\sum_{j,k\in\Omega} L_{jk}(\phi_k - \phi_j)\phi_j + d_I \sum_{j,k\in\Omega} L_{jk}\left(\phi_k' - \phi_j'\right)\phi_j + \sum_{j\in\Omega}(\beta_j - \gamma_j + \lambda^*)\phi_j\phi_j' + (\lambda^*)' \sum_{j\in\Omega}\phi_j^2 = 0.$$

Equation (3.7) and the symmetry of L imply that the second and third sums on the left-hand side cancel:

$$\begin{aligned}\sum_{j\in\Omega}(\beta_j - \gamma_j + \lambda^*)\phi_j\phi_j' &= d_I \sum_{j,k\in\Omega} L_{jk}(\phi_j - \phi_k)\phi_j' \\ &= d_I \sum_{j,k\in\Omega} L_{jk}(\phi_j' - \phi_k')\phi_j.\end{aligned}$$

Therefore,

$$\sum_{j,k\in\Omega} L_{jk}(\phi_k - \phi_j)\phi_j + (\lambda^*)' \sum_{j\in\Omega}\phi_j^2 = 0.$$

The symmetry of L implies that

$$(\lambda^*)' \sum_{j\in\Omega}\phi_j^2 = \frac{1}{2}\sum_{j,k\in\Omega} L_{jk}(\phi_j - \phi_k)^2. \tag{B.1}$$

Clearly, the right-hand side is nonnegative. We now show that it is in fact positive. We argue by contradiction. Suppose that

$$\sum_{j,k\in\Omega} L_{jk}(\phi_j - \phi_k)^2 = 0. \tag{B.2}$$

If $\phi_j = \phi_1$ for all $j \in \Omega$, then (3.7) and the positivity of ϕ imply that $\beta_j - \gamma_j + \lambda^* = 0$ for all $j \in \Omega$. But this is impossible, because H^- and H^+ are both nonempty. Therefore, it must be that $\phi_m \neq \phi_1$ for some $m \in \Omega$. The irreducibility of L implies that there exists a chain from 1 to m, i.e., a sequence $j_1, j_2, \ldots, j_s \in \Omega$ with $j_1 = 1$ and $j_s = m$ such that $L_{j_p j_{p+1}} > 0$ for $1 \le p \le s-1$. Equation (B.2) implies that $\phi_{j_p} = \phi_{j_{p+1}}$ for $1 \le p \le s-1$. Hence, $\phi_1 = \phi_m$, another contradiction. We conclude that the right-hand side of (B.1) is positive, and thus that $(\lambda^*)'$ is also positive.

(c) A variational characterization of λ^* is given by

$$\lambda^* = \inf_{\sum_{j\in\Omega}\varphi_j^2 = 1}\left\{\frac{d_I}{2}\sum_{j,k\in\Omega} L_{jk}(\varphi_j - \varphi_k)^2 + \sum_{j\in\Omega}(\gamma_j - \beta_j)\varphi_j^2\right\}, \tag{B.3}$$

and (3.7) is its corresponding Euler–Lagrange equation. Thus,

$$\lim_{d_I\to 0}\lambda^* = \inf_{\sum_{j\in\Omega}\varphi_j^2 = 1}\left\{\sum_{j\in\Omega}(\gamma_j - \beta_j)\varphi_j^2\right\}.$$

The right-hand side is minimized by setting $\varphi_j = 1$ for a single $j \in \Omega$ with the property that $\gamma_j - \beta_j = \min\{\gamma_k - \beta_k : k \in \Omega\}$ and letting $\varphi_j = 0$ otherwise.

(d) Parts (b) and (c) show that λ^* is a strictly monotone increasing function of $d_I > 0$ that is bounded from below, and substituting $\phi_j = 1/\sqrt{n}$ for $j \in \Omega$ into (B.3) shows that λ^* is bounded from above. Therefore, λ^* has a limit $\lambda^*_\infty \in (\min\{\gamma_k - \beta_k : k \in \Omega\}, \infty)$ as $d_I \to \infty$. We divide both sides of (3.7) by d_I to get

$$\sum_{k\in\Omega} L_{jk}(\phi_k - \phi_j) + \frac{(\beta_j - \gamma_j)\phi_j}{d_I} + \frac{\lambda^*\phi_j}{d_I} = 0, \quad j \in \Omega. \tag{B.4}$$

Without loss of generality, we may assume that $\phi = (\phi_j)$ is a unit vector, i.e., $\sum_{j\in\Omega} \phi_j^2 = 1$. It follows from the positivity of ϕ and compactness that $\phi \to \bar{\phi}$, where $\bar{\phi}_j \geq 0$ for $j \in \Omega$ and $\sum_{j\in\Omega} \bar{\phi}_j^2 = 1$, for some positive sequence of values $d_I \to \infty$. Let this sequence be denoted by $d_I^{(l)}$. Taking such a limit in (B.4) produces

$$\sum_{k\in\Omega} L_{jk}(\bar{\phi}_k - \bar{\phi}_j) = 0, \quad j \in \Omega.$$

We can write this equation in matrix-vector form as $A\bar{\phi} = \bar{\phi}$, where $A = (L_{jk}/L_j)$. The nonnegativity and irreducibility of A implies that $\bar{\phi}$ is proportional to $(1, 1, \ldots, 1)^t$, as both vectors belong to the principal eigenvalue $\mu = 1$. The fact that $\bar{\phi}$ is a nonnegative unit vector implies that $\bar{\phi}_j = 1/\sqrt{n}$ for $j \in \Omega$. Observe from the symmetry of L that $\sum_{j,k\in\Omega} L_{jk}(\phi_k - \phi_j) = 0$. Therefore, summing (3.7) with $d_I = d_I^{(l)}$ over all $j \in \Omega$ yields

$$\sum_{j\in\Omega} (\beta_j - \gamma_j)\phi_j + \lambda^*\big(d_I^{(l)}\big) \sum_{j\in\Omega} \phi_j = 0. \tag{B.5}$$

We let $d_I^{(l)} \to \infty$ to get

$$\sum_{j\in\Omega} (\beta_j - \gamma_j)\bar{\phi}_j + \lambda^*_\infty \sum_{j\in\Omega} \bar{\phi}_j = 0.$$

Since $\bar{\phi}_j = 1/\sqrt{n}$ for $j \in \Omega$, we obtain

$$\lambda^*_\infty = \frac{1}{n} \sum_{j\in\Omega} (\gamma_j - \beta_j) = \frac{\Sigma_\gamma - \Sigma_\beta}{n}.$$

Finally, parts (e) and (f) follow directly from parts (b), (c), and (d) together with the fact that H^+ is nonempty. ∎

Appendix C.

Proof of Lemma 3.7.

(a) We argue by induction on l for $I^{(l)}$. If $l = 0$, then the result is immediate because $I_j^{(0)} = \underline{I}_j = \epsilon\phi_j \in [0, 1]$ for $j \in \Omega$. Suppose now that $I_j^{(l)} \in [0, 1]$ for $0 \leq l \leq s$ and $j \in \Omega$, where $s \geq 0$, but that $I_m^{(s+1)} \notin [0, 1]$ for some $m \in \Omega$. Suppose first that $I_m^{(s+1)} < 0$. Without loss of generality, we may assume that $I_m^{(s+1)} = \min\{I_j^{(s+1)} : j \in \Omega\}$. Equation (3.11) with $l = s$ and $j = m$ implies that

$$-d_I \sum_{k\in\Omega} L_{mk}\big(I_k^{(s+1)} - I_m^{(s+1)}\big) + MI_m^{(s+1)} = \Big[f_m\big(I_m^{(s)}\big) + M\Big] I_m^{(s)}.$$

Since $I_m^{(s)} \in [0,1]$, we have, by the properties of M,

$$-d_I \sum_{k\in\Omega} L_{mk}\big(I_k^{(s+1)} - I_m^{(s+1)}\big) + MI_m^{(s+1)} \geq 0.$$

Since $MI_m^{(s+1)} < 0$, it follows that

$$d_I \sum_{k\in\Omega} L_{mk}\big(I_k^{(s+1)} - I_m^{(s+1)}\big) < 0.$$

However, this result contradicts the minimality of $I_m^{(s+1)}$. Suppose now that $I_m^{(s+1)} > 1$, and without loss of generality that $I_m^{(s+1)} = \max\{I_j^{(s+1)} : j \in \Omega\}$. Again, (3.11) with $l = s$ and $j = m$ implies that

$$-d_I \sum_{k\in\Omega} L_{mk}\big(I_k^{(s+1)} - I_m^{(s+1)}\big) + MI_m^{(s+1)} = F_m\big(I_m^{(s)}\big) + MI_m^{(s)}.$$

Since $I_m^{(s)} \in [0,1]$, and $F_m(u) + Mu$ is a monotone increasing function of $u \in [0,1]$, we have

$$F_m\big(I_m^{(s)}\big) + MI_m^{(s)} \leq F_m(1) + M \leq M,$$

where the last inequality follows from $F_m(1) = f_m(1) = -\gamma_m \leq 0$. Hence

$$-d_I \sum_{k\in\Omega} L_{mk}\big(I_k^{(s+1)} - I_m^{(s+1)}\big) + MI_m^{(s+1)} \leq M.$$

Since $I_m^{(s+1)} > 1$, we have

$$-d_I \sum_{k\in\Omega} L_{mk}\big(I_k^{(s+1)} - I_m^{(s+1)}\big) < 0.$$

However, this result contradicts the maximality of $I_m^{(s+1)}$. We conclude that $I_j^{(s+1)} \in [0,1]$ for all $j \in \Omega$, and, by induction, $I_j^{(l)} \in [0,1]$ for $l \geq 0$ and $j \in \Omega$. The argument for $I^{[l]}$ is similar.

(b) We argue by induction on l for $\Delta I^{(l)}$. To show that $\Delta I^{(0)} \geq 0$, we suppose otherwise and obtain a contradiction. If $\Delta I^{(0)} \not\geq 0$, then there exists some $m \in \Omega$ such that $\Delta I_m^{(0)} < 0$. We may assume that $\Delta I_m^{(0)} = \min\{\Delta I_j^{(0)} : j \in \Omega\}$. Recall that ϵ was chosen so that $G(\underline{I}) = G(\epsilon\phi) \geq 0$. Equation (3.9) with $j = m$ and $I = I^{(0)} = \underline{I}$ implies that

$$G_m\big(I^{(0)}\big) = d_I \sum_{k\in\Omega} L_{mk}\big(I_k^{(0)} - I_m^{(0)}\big) + F_m\big(I_m^{(0)}\big) \geq 0.$$

It follows from this inequality and (3.11) with $l = 0$ and $j = m$ that

$$-d_I \sum_{k\in\Omega} L_{mk}\big(I_k^{(1)} - I_m^{(1)}\big) + MI_m^{(1)} \geq d_I \sum_{k\in\Omega} L_{mk}\big(I_m^{(0)} - I_k^{(0)}\big) + MI_m^{(0)}.$$

Therefore,

$$d_I \sum_{k\in\Omega} L_{mk}\big(\Delta I_k^{(0)} - \Delta I_m^{(0)}\big) \leq M\Delta I_m^{(0)} < 0.$$

However, this result contradicts the minimality of $\Delta I_m^{(0)}$. We conclude that $\Delta I^{(0)} \geq 0$.

Suppose now that $\Delta I^{(l)} \geq 0$ for $0 \leq l \leq s$, where $s \geq 0$, but that $\Delta I^{(s+1)} \not\geq 0$. Then there exists some $m \in \Omega$ such that $\Delta I_m^{(s+1)} < 0$. We may assume that $\Delta I_m^{(s+1)} = \min\{\Delta I_j^{(s+1)} : j \in \Omega\}$. We subtract (3.11) with $l = s$ and $j = m$ from (3.11) with $l = s+1$ and $j = m$ to get

$$\begin{aligned} -d_I \sum_{k\in\Omega} L_{mk}\big(\Delta I_k^{(s+1)} - \Delta I_m^{(s+1)}\big) + M\Delta I_m^{(s+1)} \\ = F_m\big(I_m^{(s+1)}\big) - F_m\big(I_m^{(s)}\big) + M\Delta I_m^{(s)}. \end{aligned}$$

Recall from part (a) that $I_m^{(s)}$ and $I_m^{(s+1)}$ both lie within the interval $[0,1]$. There exists ζ between $I_m^{(s)}$ and $I_m^{(s+1)}$ such that

$$-d_I \sum_{k\in\Omega} L_{mk}\big(\Delta I_k^{(s+1)} - \Delta I_m^{(s+1)}\big) + M\Delta I_m^{(s+1)} = \Big[F_m'(\zeta) + M\Big]\Delta I_m^{(s)}.$$

The right-hand side is nonnegative because $F_m'(\zeta) + M > 0$ and $\Delta I_m^{(s)} \geq 0$. Therefore,

$$d_I \sum_{k\in\Omega} L_{mk}\big(\Delta I_k^{(s+1)} - \Delta I_m^{(s+1)}\big) \leq M\Delta I_m^{(s+1)} < 0.$$

However, this result contradicts the minimality of $\Delta I_m^{(s+1)}$. We conclude that $\Delta I^{(s+1)} \geq 0$, and, by induction, $\Delta I^{(l)} \geq 0$ for $l \geq 0$. The arguments for $\Delta I^{[l]}$ and $\Delta I^{\{l\}}$ are similar, except that $\Delta I^{[1]} = I^{[1]} - I^{[0]} \leq 0$ by part (a) and $\Delta I^{\{0\}} = I^{[0]} - I^{(0)} \geq 0$ is immediately clear. ∎

Acknowledgment. We thank two anonymous referees for their helpful comments.

REFERENCES

L. J. S. Allen, B. M. Bolker, Y. Lou, and A. L. Nevai (2007), *Asymptotic profile of the steady states for an SIS epidemic disease reaction-diffusion model*, Discrete Contin. Dynam. Systems, to appear.

L. J. S. Allen, N. Kirupaharan, and S. M. Wilson (2004), *SIS epidemic models with multiple pathogen strains*, J. Difference Equations Appl., 10, pp. 53–75.

P. Amarasekare (2004), *The role of density-dependent dispersal in source-sink dynamics*, J. Theoret. Biol., 226, pp. 159–168.

J. Arino, J. R. Davis, D. Hartley, R. Jordan, J. M. Miller, and P. van den Driessche (2005), *A multi-species epidemic model with spatial dynamics*, Math. Med. Biol., 22, pp. 129–142.

J. Arino and P. van den Driessche (2003a), *The basic reproduction number in a multi-city compartmental epidemic model*, in Proceedings of Posta, Lecture Notes in Control and Inform. Sci. 294, L. Benvenuti, A. De Santis, and L. Farina, eds., Springer, New York, pp. 135–142.

J. Arino and P. van den Driessche (2003b), *A multi-city epidemic model*, Math. Popul. Stud., 10, pp. 175–193.

J. Arino and P. van den Driessche (2006), *Disease spread in metapopulations*, in Nonlinear Dynamics and Evolution Equations, Fields Inst. Commun. 48, H. Brunner, X.-O. Zhao, and X. Zou, eds., AMS, Providence, RI, pp. 1–13.

R. B. Bapat and T. E. S. Raghavan (1997), *Nonnegative Matrices and Applications*, Cambridge University Press, Cambrige, UK.

B. M. Bolker and B. T. Grenfell (1995), *Space, persistence, and dynamics of measles epidemics*, Phil. Trans. R. Soc. Lond. B, 348, pp. 309–320.

F. Brauer and J. A. Nohel (1989), *The Qualitative Theory of Ordinary Differential Equations*, Dover, New York.

C. Castillo-Chavez and A.-A. Yakubu (2001), *Dispersal, disease and life-history evolution*, Math. Biosci., 173, pp. 35–53.

C. Castillo-Chavez and A.-A. Yakubu (2002), *Intraspecific competition, dispersal, and disease dynamics in discrete-time patchy environments*, in Mathematical Approaches for Emerging and Reemerging Infectious Diseases: An Introduction, C. Castillo-Chavez, S. Blower, P. van den Driessche, D. Kirschner, and A.-A. Yakubu, eds., Springer-Verlag, New York, pp. 165–181.

T. Dhirasakdanon, H. Thieme, and P. van den Driessche (2007), *A Sharp Threshold for Disease Persistence in Patchy Host Populations*, preprint.

O. Diekmann and J. A. P. Heesterbeek (2000), *Mathematical Epidemiology of Infectious Diseases*, John Wiley and Sons, Chichester, New York.

O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz (1990), *On the definition and the computation of the basic reproduction ratio $\mathcal{R}_0$ in models for infectious diseases in heterogeneous populations*, J. Math. Biol., 28, pp. 365–382.

F. R. Gantmacher (1960), *The Theory of Matrices, Vol.* II, Chelsea, New York.

R. Gomulkiewicz, R. D. Holt, and M. Barfield (1995), *The effects of density dependence and immigration on local adaptation and niche evolution in a black-hole sink environment*, Theoret. Pop. Biol., 55, pp. 283–296.

G. Hess (1996), *Disease in metapopulation models: Implications for conservation*, Ecology, 77, pp. 1617–1632.

Y. Jin and W. Wang (2005), *The effect of population dispersal on the spread of a disease*, J. Math. Anal. Appl., 308, pp. 343–364.

J. A. Jacquez, C. P. Simon, and J. S. Koopman (1995), *Core groups and the $\mathcal{R}_0$'s for subgroups in heterogeneous SIS and SI models*, in Epidemic Models: Their Structure and Relation to Data, D. Mollison, ed., Cambridge University Press, Cambridge, UK.

A. L. Lloyd and V. A. A. Jansen (2004), *Spatiotemporal dynamics of epidemics: Synchrony in metapopulation models*, Math. Biosci., 188, pp. 1–16.

A. Lloyd and R. M. May (1996), *Spatial heterogeneity in epidemic models*, J. Theoret. Biol., 179, pp. 1–11.

R. K. McCormack (2006), *Multi-Host and Multi-Patch Mathematical Epidemic Models for Disease Emergence with Applications to Hantavirus in Wild Rodent Populations*, Ph.D. dissertation, Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX.

T. Namba and C. Hashimoto (2004), *Dispersal-mediated coexistence of competing predators*, Theoret. Pop. Biol., 66, pp. 53–70.

J. M. Ortega (1987), *Matrix Theory. A Second Course*, Plenum Press, New York, London.

S. Ruan (2006), *Spatial-temporal dynamics in nonlocal epidemiological models*, in Mathematics for Life Science and Medicine, Vol. 2, Y. Takeuchi, K. Sato, and Y. Iwasa, eds., Springer-Verlag, New York, pp. 99–122.

S. Ruan, W. Wang, and S. A. Levin (2006), *The effect of global travel on the spread of SARS*, Math. Biosci., 3, pp. 205–218.

L. A. Rvachev and I. M. Longini (1985), *A mathematical model for the global spread of influenza*, Math. Biosci., 75, pp. 3–22.

M. Salmani and P. van den Driessche (2006), *A model for disease transmission in a patchy environment*, Discrete Contin. Dynam. Systems Ser. B, 6, pp. 185–202.

L. Sattenspiel and K. Dietz (1995), *A structured epidemic model incorporating geographic mobility among regions*, Math. Biosci., 128, pp. 71–91.

E. Seneta (1973), *Non-Negative Matrices: An Introduction to Theory and Applications*, Wiley, New York.

H. Smith (1995), *Monotone Dynamical Systems, An Introduction to the Theory of Competitive and Cooperative Systems*, Mathematical Surveys and Monographs, AMS, Providence, RI.

P. van den Driessche and J. Watmough (2002), *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci., 180, pp. 29–48.

W. Wang and G. Mulone (2003), *Threshold of disease transmission in a patch environment*, J. Math. Anal. Appl., 285, pp. 321–335.

W. Wang and X.-Q. Zhao (2004), *An epidemic model in a patchy environment*, Math. Biosci., 190, pp. 97–112.

SIAM J. Appl. Math.
Vol. 67, No. 5, pp. 1310–1329

NONLINEAR DYNAMICS OF ELECTRIFIED THIN LIQUID FILMS*

DMITRI TSELUIKO† AND DEMETRIOS T. PAPAGEORGIOU†

Abstract. We study a nonlinear nonlocal evolution equation describing the hydrodynamics of thin films in the presence of normal electric fields. The liquid film is assumed to be perfectly conducting and to completely wet the upper or lower surface of a horizontal flat plate. The flat plate is held at constant voltage, and a vertical electric field is generated by a second parallel electrode kept at a different constant voltage and placed at a large vertical distance from the bottom plate. The fluid is viscous, and gravity and surface tension act. The equation is derived using lubrication theory and contains an additional nonlinear nonlocal term representing the electric field. The electric field is linearly destabilizing and is particularly important in producing nontrivial dynamics in the case when the film rests on the upper side of the plate. We give rigorous results on the global boundedness of positive periodic smooth solutions, using an appropriate energy functional. We also implement a fully implicit numerical scheme and perform extensive numerical experiments. Through a combination of analysis and numerical experiments we present evidence for the global existence of positive smooth solutions. This means, in turn, that the film does not touch the wall in finite time but asymptotically at infinite time. Numerical solutions are presented to support such phenomena, which are also observed in hanging films when electric fields are absent.

Key words. thin film, electrohydrodynamics, nonlocal evolution equation

AMS subject classifications. 76D03, 76D08, 76D27, 76E17

DOI. 10.1137/060663532

1. Introduction.

A viscous liquid film wetting the upper surface of a flat horizontal substrate is expected to be stable, under normal conditions, and eventually returns to its uniform undisturbed value of the film thickness. If the film is hanging (i.e., it wets the underside of the substrate), then gravity is destabilizing. This paper is concerned with the addition of electric fields normal to the substrate. Using experiments and linear theory, Taylor and McEwan [26] observed that a sufficiently strong field can overcome viscous forces in overlying films and induce wavy perturbations. We aim to model and analyze the nonlinear stages of this phenomenon for overlying and hanging films.

In the absence of electric fields the problem was studied by Ehrhard and Davis [9], who considered spreading of viscous drops on smooth horizontal surfaces which are uniformly heated or cooled. Their isothermal evolution equation for the interface coincides with ours when there is no electric field. Yiantsios and Higgins [30] considered the behavior of a viscous film bounded below by a wall and above by an unbounded second heavier immiscible fluid. For the case when the Bond number $B \ll 1$ (it measures the ratio between gravitational and interfacial forces) and the viscosity ratio $m = \mu_1/\mu_2$ is $O(1)$, where μ_2 is the film viscosity, they obtained the same evolution equation for the interface as did Ehrhard and Davis [9]. Ehrhard [8] used the model derived by Erhard and Davis [9] to describe the quasi-steady evolution of a viscous drop hanging on the earth-facing side of a smooth horizontal plate, which is either uni-

*Received by the editors June 22, 2006; accepted for publication (in revised form) March 15, 2007; published electronically June 21, 2007.
http://www.siam.org/journals/siap/67-5/66353.html
†Department of Mathematical Sciences and Center for Applied Mathematics and Statistics, New Jersey Institute of Technology, Newark, NJ 07102 (dt8@njit.edu, depapa@oak.njit.edu). The first author acknowledges support from NJIT through its Presidential Research Initiative. The work of the second author was supported by National Science Foundation grant number DMS-0072228.

formly heated or cooled. Also, other similar equations arising in the modeling of thin liquid films have been derived and studied by Bertozzi [1], Dussan [7], Greenspan [13], Haley and Miksis [14], Hocking [16], Myers [18], and Oron, Davis, and Bankoff [19].

Here we consider the problem of a perfectly conducting liquid film on a horizontal plane with the upper electrode placed far from the grounded substrate. There has been considerable interest in electrically induced instabilities and their use in pattern formation and transfer in photolithographic applications. Demonstrations of such instabilities in this context have been made by Schaffer et al. [23], [24] and Lin et al. [17], for example. Theoretical works have focused on linear theory, as in Pease and Russel [21] and references therein, as well as long wave theories in the thin film and small gap approximation (i.e., the second electrode is placed close to the grounded bottom plate) in Shankar and Sharma [25], and the more recent leaky dielectric study of Craster and Matar [5]. The present work is related to but different from those cited above, but we expect that the mathematical tools developed here can be used in those problems also.

A long wave theory leads to a nonlocal nonlinear evolution equation at the leading order, which by a change of the sign of the gravitational parameter also describes hanging films. (Nonlocal equations have also been derived in related problems by Papageorgiou and Vanden-Broeck [20], Savettaseranee et al. [22], and Tilley, Petropoulos, and Papageorgiou [27].) If $H(x, t)$ denotes the scaled interfacial position, then the equation takes the form

$$H_t + \frac{1}{3}\left[H^3\left(\frac{1}{C}H_{xxx} - GH_x + 2W_e\mathcal{H}[H_{xx}]\right)\right]_x = 0, \quad (x, t) \in \mathbb{R} \times \mathbb{R}_+, \tag{1.1}$$

$$H(x, t) = H(x + 2L, t),$$

where $C > 0$, $W_e > 0$, and G can be positive or negative for overlying or hanging films, respectively. In the absence of an electric field the film is linearly stable or unstable, depending on whether $G > 0$ or $G < 0$. The addition of an electric field can always make the film unstable irrespective of the sign of G.

We prove that positive smooth solutions of (1.1) do not blow up and are uniformly bounded for all time in the Sobolev H^1-norm. This is done by constructing an appropriate energy functional $\mathcal{E}[H]$ having the steady-state solutions as extrema. Our analysis extends that of Bertozzi and Pugh [3] to the nonlocal equation (1.1). We also note that Hocherman and Rosenau [15] considered a class of thin film equations with the coefficients in front of the spatial derivatives being polynomials of higher or lower degree of the unknown function (it is unclear whether such equations arise in physical applications). They were interested in identifying equations whose solutions blow up in finite time, and they made a conjecture regarding this. This possibility was also recently studied by Bertozzi and Pugh [3], [4], and Witelski, Bernoff, and Bertozzi [29] based on both rigorous analysis and numerical computations.

For (1.1) we also establish analytically that for positive solutions the integral $\int_{-L}^{L}(1/H)dx$ is bounded on each time interval. Extensive numerical experiments indicate that $\max|H_{xx}|$ is bounded on each time interval; if we assume this, then we can use the observation on the evolution of $\int_{-L}^{L}(1/H)dx$ to prove global existence of positive smooth solutions (i.e., that the film does not touch down in finite time). A rigorous proof of boundedness of $\max|H_{xx}|$ is under investigation. Results of extensive computations and their relation to the analytical results are also reported.

The outline of the paper is as follows. In section 2 we describe the physical problem and give the governing equations. Section 3 develops the asymptotic long

wave theory that leads to the scaled evolution equation (1.1). Section 4 is devoted to rigorous analytical results, and section 5 describes the numerical method used to solve (1.1). Section 6 contains the numerical results, and finally, in section 7 we give our conclusions.

2. Physical model and governing equations. Consider a viscous liquid film completely wetting a solid horizontal substrate. Two related configurations are of interest: *overlying* two-dimensional films with the liquid layer resting on the substrate, and *overhanging* two-dimensional films with the liquid layer wetting the underside of the horizontal substrate. A schematic is provided in Figure 1 for the overlying film with a normal electric field present.

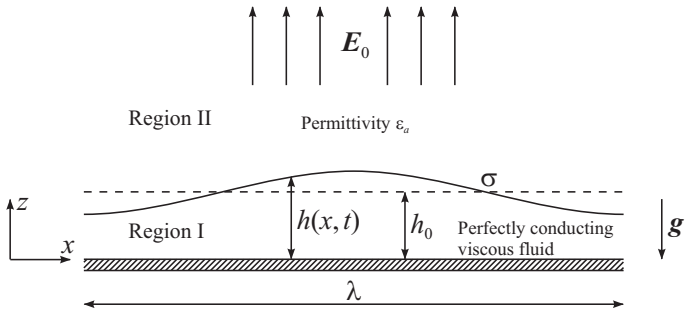


FIG. 1. *Schematic of the problem.*

The fluid is Newtonian of a constant density ρ and dynamic viscosity μ and is assumed to be a perfect conductor. The surface tension coefficient between the liquid and the surrounding medium is σ. We denote by $h(x, t)$ the local film thickness, which is a function of space and time, and the unperturbed thickness of the liquid layer is h_0. The gravitational acceleration $\boldsymbol{g}$ acts in the vertical direction. The plate is a grounded infinite electrode held at zero voltage. Another flat parallel electrode is placed infinitely far from the wetted substrate, so that a uniform vertical electric field is set up at infinity; i.e., at infinity the electric field $\boldsymbol{E}$ approaches a constant value $\boldsymbol{E}_0$ which is normal to the plate. The surrounding medium is assumed to be a perfect dielectric with permittivity ε_a, and the corresponding voltage potential in it is denoted by V. Since the liquid is a perfect electric conductor, the potential is zero on the interface, and there is no electric field inside the liquid layer. We use a rectangular coordinate system (x, z) with the x-axis pointing along the plate and the z-axis pointing up and being perpendicular to the plate. The associated velocity field is denoted by $\boldsymbol{u} = (u, v)$, the liquid layer is denoted by Region I, and the surrounding medium by Region II.

The hydrodynamics in Region I is governed by the Navier–Stokes equations. In Region II the electric field can be written as $\boldsymbol{E} = -\nabla V$, with V satisfying the Laplace equation. The equations are made dimensionless by scaling lengths with h_0, velocities with a typical velocity U_0, time with h_0/U_0, pressure with $\mu U_0/h_0$, and voltages by $E_0 h_0$. This leads to the following equations in Region I:

$$u_t + uu_x + vu_z = \frac{1}{R}(-p_x + u_{xx} + u_{zz}), \tag{2.1}$$

$$v_t + uv_x + vv_z = \frac{1}{R}(-p_z + v_{xx} + v_{zz} - G), \tag{2.2}$$

$$u_x + v_z = 0, \tag{2.3}$$

and in Region II we obtain

$$V_{xx} + V_{zz} = 0. \tag{2.4}$$

The dimensionless boundary conditions of no slip at the wall and the far field condition in Region II for the voltage become

$$u|_{z=0} = 0, \quad v|_{z=0} = 0, \tag{2.5}$$

$$V_x \to 0,\ V_z \to -1 \quad \text{as } z \to \infty. \tag{2.6}$$

At the interface $z = h(x, t)$ we have

$$V = 0, \tag{2.7}$$

$$v = h_t + uh_x, \tag{2.8}$$

$$(1 - h_x^2)(u_z + v_x) + 4h_x v_z = 0, \tag{2.9}$$

$$-\frac{W_e}{2}(1 + h_x^2)V_z^2 + \frac{1 + h_x^2}{1 - h_x^2}v_z + \frac{1}{2}(\bar{p}_{\text{atm}} - p) = \frac{h_{xx}}{2C(1 + h_x^2)^{3/2}}, \tag{2.10}$$

where $\bar{p}_{\text{atm}} = p_{\text{atm}}h_0/\mu U_0$, with p_{atm} a constant, is the nondimensional constant pressure in Region II. The boundary condition (2.7) reflects the fact that $z = h(x, t)$ is an equipotential surface, (2.8) is the kinematic condition, and (2.9) and (2.10) follow from the balance of tangential and normal stresses at the interface. The parameters

$$R = \frac{U_0 h_0}{\nu}, \qquad C = \frac{U_0 \mu}{\sigma}, \qquad W_e = \frac{\varepsilon_a E_0^2 h_0}{2\mu U_0}, \qquad G = \frac{\rho g h_0^2}{\mu U_0} \tag{2.11}$$

are a Reynolds number, a Capillary number measuring the ratio of inertial to capillary forces, an electric Weber number measuring the ratio of electrical to fluid pressures, and a gravity number G measuring the ratio of gravitational to viscous forces.

It is useful to use the following exact solution to (2.1)–(2.10),

$$\bar{u} = 0, \qquad \bar{v} = 0, \qquad \bar{p} = \bar{p}_{\text{atm}} - W_e - G(z - 1), \qquad \overline{V} = 1 - z, \tag{2.12}$$

and introduce new unknown functions $\tilde{u}$, $\tilde{v}$, $\tilde{p}$, $\widetilde{V}$ by $u = \bar{u} + \tilde{u}$, etc., and drop tildes from the transformed equations and boundary conditions. The resulting fully nonlinear dimensionless system is exact and presents a formidable computational and analytical task. In what follows we make analytical progress by studying the physically relevant case of thin films using a long wave nonlinear theory.

3. Long wave asymptotics. In the asymptotic analysis presented next we assume that the typical length λ of the interface deformation is long compared to the undisturbed thickness; that is, $\delta = h_0/\lambda \ll 1$ is a small parameter. In Region I, we introduce the lubrication scalings

$$x = \frac{1}{\delta}\xi, \quad t = \frac{1}{\delta}\tau, \quad v = \delta w, \quad p = \frac{1}{\delta}P. \tag{3.1}$$

The conditions at the interface $z = h(\xi, t)$ become

$$w = h_\tau + uh_\xi, \tag{3.2}$$

$$(1 - \delta^2 h_\xi^2)(u_z + \delta^2 w_\xi) + 4\delta^2 h_\xi w_z = 0, \tag{3.3}$$

$$-\frac{W_e}{2}\left[(V_z-1)^2(1+\delta^2 h_\xi^2)-1\right]+\delta\frac{1+\delta^2 h_x^2}{1-\delta^2 h_x^2}w_z$$
$$-\frac{G}{2}(1-h)-\frac{P}{2\delta}=\frac{\delta^2 h_{\xi\xi}}{2C(1+\delta^2 h_\xi^2)^{3/2}}. \tag{3.4}$$

The last boundary condition contains a nonlocal contribution since V satisfies the Laplace equation in the potential region above the fluid layer. This is obtained by considering the problem in Region II. The analysis is the same as in Tseluiko and Papageorgiou [28] and results in the following expression for V_z in terms of the interfacial position (keeping the $O(\delta)$ term and dropping the higher order terms):

$$V_z(\xi,0)=-\delta\mathcal{H}[h_\xi], \tag{3.5}$$

where $\mathcal{H}$ is the Hilbert transform operator defined by

$$\mathcal{H}[g](\xi)=\frac{1}{\pi}PV\int_{-\infty}^{\infty}\frac{g(\xi')}{\xi-\xi'}d\xi', \tag{3.6}$$

where PV denotes the principal value. Using (3.5) in (3.4) yields

$$-\delta W_e\mathcal{H}[h_\xi]+\delta w_z-\frac{G}{2}(1-h)-\frac{P}{2\delta}=\frac{\delta^2 h_{\xi\xi}}{2C}+O(\delta^2), \tag{3.7}$$

from which we deduce the following canonical scalings that retain the effects of the electric field, gravity, and surface tension (bar quantities are of order one):

$$C=\delta^3\overline{C},\quad W_e=\frac{\overline{W}_e}{\delta^2},\quad G=\frac{1}{\delta}\overline{G}. \tag{3.8}$$

The Reynolds number R is assumed to be $o(\delta^{-1})$ throughout. Expanding in powers of δ, e.g., $u=u_0+\delta u_1+\cdots$, etc., gives the following leading order solutions:

$$u_0=\frac{P_{0\xi}}{2}z^2-P_{0\xi}H_0 z, \tag{3.9}$$

$$w_0=-\frac{P_{0\xi\xi}z^3}{6}+\frac{P_{0\xi\xi}H_0 z^2}{2}+\frac{P_{0\xi}H_{0\xi}z^2}{2}, \tag{3.10}$$

$$P_0=-2\overline{W}_e\mathcal{H}[H_{0\xi}]-\overline{G}(1-H_0)-\frac{1}{\overline{C}}H_{0\xi\xi}. \tag{3.11}$$

Using the velocities (3.9) and (3.10) in the kinematic condition (3.2) gives $H_{0\tau}=\frac{1}{3}[H_0^3P_{0\xi}]_\xi$; using (3.11) for P_0, the evolution equation becomes

$$H_t+\frac{1}{3}\left[H^3\left(\frac{1}{C}H_{xxx}-GH_x+2W_e\mathcal{H}[H_{xx}]\right)\right]_x=0, \tag{3.12}$$

where for simplicity we write t and x for τ and ξ, H for H_0, and drop the bars from $\overline{C}$, $\overline{G}$, $\overline{W}_e$. There are several noteworthy features of (3.12). The electric field enters through a nonlocal term and is destabilizing as in falling films; see Tseluiko and Papageorgiou [28]. Gravity is present, and if we allow G to be negative, we obtain the long wave thin film dynamics of hanging films. In the absence of an electric field ($W_e=0$) and if $G>0$, the flow is linearly stable; instability is possible if $G<0$, as is intuitive for hanging films (this case has been considered by Bertozzi and Pugh [3],

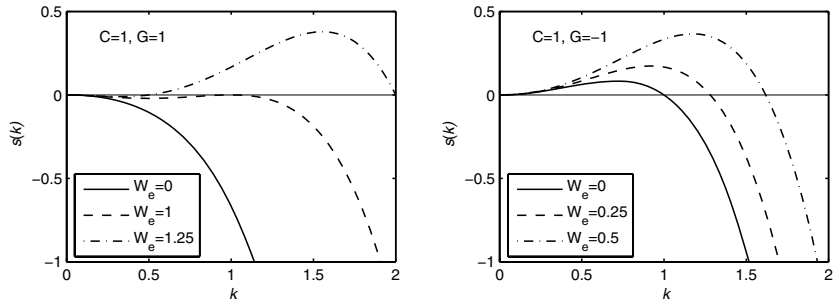


FIG. 2. *The effect of the electric field on linear growth rates $s(k)$. The left panel has $G = 1$ (overlying film) and the right panel $G = -1$ (under-hanging film). In both cases $C = 1$, and the values of W_e are shown in the figures.*

Ehrhard [8], Ehrhard and Davis [9], Yiantsios and Higgins [30], [31]). The electric field, however, can be utilized to destabilize liquid films lying on top of a substrate electrode ($G > 0$), and the novel equation (3.12) enables a quantitative study of such phenomena. To quantify some of these observations we perform a linear stability analysis to identify stable and unstable regimes when $W_e \neq 0$. Writing $H = 1 + \epsilon\eta$, linearizing with respect to ϵ, and seeking solutions proportional to $\eta = \hat{\eta}\exp(st + \mathrm{i}kx)$, where $\hat{\eta}$ is a complex constant, leads to the following linear dispersion relation:

$$s(k) = -\frac{1}{3C}k^4 + \frac{2W_e}{3}k^2|k| - \frac{G}{3}k^2. \tag{3.13}$$

(We have used the Fourier transform property $\mathcal{F}\big[\mathcal{H}[u]\big](k) = -\mathrm{i}\,\mathrm{sign}(\mathrm{Re}k)\hat{u}(k)$.)

When $G > 0$ (the film is resting on a substrate), we see that for $W_e < (G/C)^{1/2}$ all modes are stable; i.e., $s(k) < 0$ for all k. For $W_e > (G/C)^{1/2}$, however, there is a band of unstable waves with wavenumbers extending from $k_L = CW_e - \sqrt{(CW_e)^2 - CG}$ to $k_R = CW_e + \sqrt{(CW_e)^2 - CG}$, and the electric field is destabilizing. The most unstable mode has wavenumber $k = (3CW_e + \sqrt{9(CW_e)^2 - 8CG})/4$. Also, $k_L > 0$ for all W_e, and hence all waves in the immediate vicinity of $k = 0$ are stable—for large W_e, for example, all waves longer that $4\pi W_e/G$ are stable. Typical results are depicted in Figure 2 (the left panel), for the particular values $C = 1$, $G = 1$ for which $W_e > 1$ yields instability, as is clear from the figure.

When $G < 0$, there is always a band of unstable waves extending from $k_L = 0$ to $k_R = CW_e + \sqrt{(CW_e)^2 - CG}$; see Figure 2 (the right panel) for typical results for the case $C = 1$, $G = 1$. For large W_e we have $k_R \sim 2CW_e$, and so increasingly shorter wavelengths become linearly unstable as W_e increases. Damping of sufficiently high wavenumbers (and, hence, well-posedness) is provided by the presence of surface tension which is extremely important in this case.

The linear results set the stage for nonlinear computations and analysis, and we can expect nontrivial behavior in parameter regimes that support unstable linear waves. We concentrate on such calculations next.

4. Analytical results. In this section we prove the global boundedness of positive classical solutions of the evolution equation (3.12). This is achieved by constructing an energy functional whose extrema are steady state solutions of (3.12) and using it to estimate the H^1-norm of the solution and show that it is uniformly bounded.

The main result is Proposition 4.2 in section 4.2. The results extend those of Bertozzi and Pugh [3] to a class of physically meaningful nonlocal equations.

4.1. The energy functional. We consider the generalized equation

$$H_t + \big[f(H)\mathcal{A}[H]_x\big]_x = 0, \tag{4.1}$$

where f is a function which takes positive values for positive arguments and is zero only at zero, and $\mathcal{A}[H]$ is some integro-differential operator, which involves the function H, its first and second spatial derivatives, and the Hilbert transform operator. The additional condition that is satisfied by this operator will be given below. We consider (4.1) on a periodic interval $[-L, L]$ with positive initial data $H(x, 0) = H_0(x)$.

Steady state solutions of (4.1) are found by integrating once to obtain

$$f(H)\mathcal{A}[H]_x = C_1, \tag{4.2}$$

where C_1 is some constant. If H vanishes at some point, then $C_1 = 0$. Otherwise $\mathcal{A}[H]_x = C_1/f(H)$; integration gives $C_1 \int_{-L}^{L}(1/f(H))dx = 0$, which in turn implies that $C_1 = 0$. So, for steady state solutions, $\mathcal{A}[H]_x = 0$, i.e.,

$$\mathcal{A}[H] = C_2, \tag{4.3}$$

where C_2 is some constant.

Let $\mathcal{E}[H]$ be the energy functional having the form

$$\mathcal{E}[H] = \int_{-L}^{L} \mathcal{L}\big(H, H_x, \mathcal{H}[H]\big)dx, \tag{4.4}$$

and whose extrema are the steady state solutions of (4.1). More precisely, we assume that the following generalized Euler–Lagrange equation,

$$\frac{\partial \mathcal{L}}{\partial H} - \frac{d}{dx}\left[\frac{\partial \mathcal{L}}{\partial H_x}\right] - \mathcal{H}\left[\frac{\partial \mathcal{L}}{\partial \mathcal{H}[H]}\right] = 0, \tag{4.5}$$

coincides with the equation $C_2 - \mathcal{A}[H] = 0$. Then

$$\frac{d\mathcal{E}[H]}{dt} = \int_{-L}^{L}\left(\frac{\partial \mathcal{L}}{\partial H}H_t + \frac{\partial \mathcal{L}}{\partial H_x}H_{xt} + \frac{\partial \mathcal{L}}{\partial \mathcal{H}[H]}\mathcal{H}[H_t]\right)dx. \tag{4.6}$$

Integrating the second term by parts and applying the property $\int_I u(x)\mathcal{H}[v](x)dx = -\int_I v(x)\mathcal{H}[u](x)dx$ to the third term gives

$$\frac{d\mathcal{E}[H]}{dt} = \int_{-L}^{L}\left(\frac{\partial \mathcal{L}}{\partial H} - \frac{d}{dx}\left[\frac{\partial \mathcal{L}}{\partial H_x}\right] - \mathcal{H}\left[\frac{\partial \mathcal{L}}{\partial \mathcal{H}[H]}\right]\right)H_t dx \tag{4.7}$$

$$= -\int_{-L}^{L}(C_2 - \mathcal{A}[H])\big[f(H)\mathcal{A}[H]_x\big]_x dx = -\int_{-L}^{L} f(H)\mathcal{A}[H]_x^2 dx. \tag{4.8}$$

Therefore, $d\mathcal{E}[H]/dt \le 0$ for nonnegative H i.e., $\mathcal{E}[H]$ is bounded above.

For (3.12) we have $f(H) = H^3/3$ and $\mathcal{A}[H] = (1/C)H_{xx} - GH + 2W_e\mathcal{H}[H_x]$. The steady state solutions are determined by the equation

$$\frac{1}{C}H_{xx} - GH + 2W_e\mathcal{H}[H_x] = C_2, \tag{4.9}$$

which on integration yields

$$C_2 = -\frac{G}{2L}\int_{-L}^{L} H dx. \tag{4.10}$$

It follows that the functional $\mathcal{L}\big(H,\, H_x,\, \mathcal{H}[H]\big)$ can be chosen in the following form:

$$\mathcal{L}\big(H,\, H_x,\, \mathcal{H}[H]\big) = \frac{1}{2C}H_x^2 + \frac{G}{2}H^2 + W_e H_x \mathcal{H}[H] + C_2 H. \tag{4.11}$$

Thus, the energy functional

$$\mathcal{E}[H] = \int_{-L}^{L}\left(\frac{1}{2C}H_x^2 + \frac{G}{2}H^2 + W_e H_x \mathcal{H}[H] + C_2 H\right)dx \tag{4.12}$$

is a nonincreasing function of time for a nonnegative solution H.

4.2. Uniform boundedness of positive smooth solutions. In the previous section we have shown that the energy functional $\mathcal{E}[H]$ is bounded above by its initial value for a nonnegative solution H of (3.12). In this section we will show uniform boundedness of solutions. We will restrict our consideration to positive solutions, since, given upper and lower bounds for positive solutions, the equation is uniformly parabolic, which implies small time smoothness of the solutions (see Eidelman [10], Friedman [11]).

We begin with the following lemma.

LEMMA 4.1. *Let $\mathcal{E}[H]$ be the functional defined on $H^1(-L,\, L)$ by* (4.12). *Then there exist constants $\alpha > 0$, $\beta > 0$, and γ such that (s.t.)*

$$\|H\|_{H^1}^2 \le \alpha \mathcal{E}[H] + \beta \|H\|_1^2 + \gamma \|H\|_1 \tag{4.13}$$

for all nonnegative $H \in H^1_{\mathrm{per}}$. (Here and everywhere else we denote by L^2_{per}, H^k_{per}, $k = 1, 2, \ldots$, the subspaces of the Sobolev spaces $L^2(-L,\, L)$, $H^k(-L,\, L)$ consisting of periodic functions with period $2L$.)

Proof. First, using the Cauchy–Schwartz and Young's inequalities and the property $\|\mathcal{H}[u]\| = \|u\|$ of the Hilbert transform gives

$$\begin{aligned}\int_{-L}^{L} H_x \mathcal{H}[H] dx &\ge -\|H_x\|_2 \|\mathcal{H}[H]\|_2 = -\|H_x\|_2 \|H\|_2 \\ &\ge -\frac{\varepsilon_1}{2}\|H_x\|_2^2 - \frac{1}{2\varepsilon_1}\|H\|_2^2, \end{aligned} \tag{4.14}$$

where ε_1 is some positive number. Hence,

$$\begin{aligned}\mathcal{E}[H] &\ge \int_{-L}^{L}\left(\frac{1}{2C}H_x^2 + \frac{G}{2}H^2 + C_2 H\right)dx - \frac{\varepsilon_1 W_e}{2}\|H_x\|_2^2 - \frac{W_e}{2\varepsilon_1}\|H\|_2^2 \\ &= \left(\frac{1}{2C} - \frac{\varepsilon_1 W_e}{2}\right)\|H_x\|_2^2 + \left(\frac{G}{2} - \frac{W_e}{2\varepsilon_1}\right)\|H\|_2^2 + C_2\int_{-L}^{L} H dx \\ &= \left(\frac{1}{2C} - \frac{\varepsilon_1 W_e}{2}\right)\|H\|_{H^1}^2 + \left(\frac{G}{2} - \frac{W_e}{2\varepsilon_1} - \frac{1}{2C} + \frac{\varepsilon_1 W_e}{2}\right)\|H\|_2^2 + C_2\int_{-L}^{L} H dx. \end{aligned} \tag{4.15}$$

Note that $\int_{-L}^{L} H dx = \|H\|_1$ for nonnegative H. We define $A = 1/2C - \varepsilon_1 W_e/2$ and $B = -G/2 + W_e/2\varepsilon_1 + 1/2C - \varepsilon_1 W_e/2$. Choosing ε_1 sufficiently small gives $A > 0$, $B > 0$. Also, using the interpolation inequality

$$\|H\|_2 \le C_3 \|H\|_{H^1}^{1/3} \|H\|_1^{2/3} \tag{4.16}$$

and applying Young's inequality for the right-hand side of the expression above gives

$$\|H\|_2 \le C_3 \left(\frac{\varepsilon_2}{3} \|H\|_{H^1} + \frac{2}{3\varepsilon_2^{1/2}} \|H\|_1 \right), \tag{4.17}$$

where ε_2 is some positive number. Therefore,

$$\|H\|_2^2 \le \frac{2\varepsilon_2^2 C_3^2}{9} \|H\|_{H^1}^2 + \frac{8C_3^2}{9\varepsilon_2} \|H\|_1^2, \tag{4.18}$$

and we get

$$\mathcal{E}[H] \ge A\|H\|_{H^1}^2 - B\|H\|_2^2 + C_2\|H\|_1 \tag{4.19}$$

$$\ge A\|H\|_{H^1}^2 - B\left(\frac{2\varepsilon_2^2 C_3^2}{9} \|H\|_{H^1}^2 + \frac{8C_3^2}{9\varepsilon_2} \|H\|_1^2 \right) + C_2\|H\|_1 \tag{4.20}$$

$$= \left(A - \frac{2\varepsilon_2^2 B C_3^2}{9} \right) \|H\|_{H^1}^2 - \frac{8BC_3^2}{9\varepsilon_2} \|H\|_1^2 + C_2\|H\|_1. \tag{4.21}$$

We define $\tilde{A} = A - 2\varepsilon_2^2 B C_3^2/9$ and $\tilde{B} = 8BC_3^2/9\varepsilon_2$. Note that $\tilde{B}$ is positive and choosing ε_2 small enough makes $\tilde{A}$ also positive. Thus,

$$\mathcal{E}[H] \ge \tilde{A}\|H\|_{H^1}^2 - \tilde{B}\|H\|_1^2 + C_2\|H\|_1, \tag{4.22}$$

i.e.,

$$\|H\|_{H^1}^2 \le \alpha \mathcal{E}[H] + \beta\|H\|_1^2 + \gamma\|H\|_1, \tag{4.23}$$

where $\alpha = 1/\tilde{A}$, $\beta = \tilde{B}/\tilde{A}$, $\gamma = -C_2/\tilde{A}$, as required. □

We can now prove uniform boundedness of positive smooth solutions to (3.12).

PROPOSITION 4.2. *Let $H(x, t)$ be a positive smooth solution of* (3.12) *with periodic boundary conditions on some time interval* $[0, T]$. *If $H(x, 0) = H_0(x) \in H^1_{\mathrm{per}}$, then $\|H\|_{H^1}$ is uniformly bounded.*

Proof. First, note that (3.12) is a conservation law. The spatial integral of the solution is conserved. Indeed, integrating over $[-L, L]$ gives $(d/dt)\int_{-L}^{L} H dx = 0$ and hence $\|H\|_1 = \|H_0\|_1$. Also, since $\mathcal{E}[H]$ is a nonincreasing function of time, (4.13) implies

$$\|H\|_{H^1}^2 \le \alpha \mathcal{E}[H] + \beta\|H\|_1^2 + \gamma\|H\|_1 \tag{4.24}$$

$$= \alpha \mathcal{E}[H] + \beta\|H_0\|_1^2 + \gamma\|H_0\|_1 \tag{4.25}$$

$$\le \alpha \mathcal{E}[H_0] + \beta\|H_0\|_1^2 + \gamma\|H_0\|_1, \tag{4.26}$$

which was to be proved. □

Remark. Proposition 4.2 is essentially a no-blow-up theorem for the solution H. The result does not prevent a touchdown (the numerics predicts a touchdown in infinite time—see Figures 3–6).

4.3. Evolution of $\int_{-L}^{L} H^{-1}dx$. In this section we show that the spatial integral of H^{-1} is bounded on each finite time interval. Indeed,

$$\frac{d}{dt}\int_{-L}^{L}\frac{dx}{H} = -\int_{-L}^{L}\frac{H_t}{H^2}dx. \tag{4.27}$$

Substituting the expression for H_t from (3.12) into (4.27) gives

$$\begin{aligned}\frac{d}{dt}\int_{-L}^{L}\frac{dx}{H} &= \frac{1}{3}\int_{-L}^{L}\frac{1}{H^2}\left[H^3\left(\frac{1}{C}H_{xxx} - GH_x + 2W_e\mathcal{H}[H_{xx}]\right)\right]_x dx\\ &= \frac{1}{3}\int_{-L}^{L}\frac{1}{H^2}(3H^2H_x)\left(\frac{1}{C}H_{xxx} - GH_x + 2W_e\mathcal{H}[H_{xx}]\right)dx\\ &\quad + \frac{1}{3}\int_{-L}^{L}\frac{1}{H^2}H^3\left(\frac{1}{C}H_{xxx} - GH_x + 2W_e\mathcal{H}[H_{xx}]\right)_x dx\\ &= \frac{1}{C}\int_{-L}^{L}H_xH_{xxx}dx - G\int_{-L}^{L}H_x^2dx + 2W_e\int_{-L}^{L}H_x\mathcal{H}[H_{xx}]dx\\ &\quad + \frac{1}{3C}\int_{-L}^{L}HH_{xxxx}dx - \frac{G}{3}\int_{-L}^{L}HH_{xx}dx + \frac{2W_e}{3}\int_{-L}^{L}H\mathcal{H}[H_{xxx}]dx.\end{aligned} \tag{4.28}$$

Integration by parts then gives

$$\begin{aligned}\frac{d}{dt}\int_{-L}^{L}\frac{dx}{H} &= -\frac{1}{C}\int_{-L}^{L}H_{xx}^2dx - G\int_{-L}^{L}H_x^2dx + 2W_e\int_{-L}^{L}H_x\mathcal{H}[H_{xx}]dx\\ &\quad + \frac{1}{3C}\int_{-L}^{L}H_{xx}^2dx + \frac{G}{3}\int_{-L}^{L}H_x^2dx - \frac{2W_e}{3}\int_{-L}^{L}H_x\mathcal{H}[H_{xx}]dx.\end{aligned} \tag{4.29}$$

Therefore,

$$\frac{d}{dt}\int_{-L}^{L}\frac{dx}{H} = -\frac{2}{3C}\int_{-L}^{L}H_{xx}^2dx - \frac{2G}{3}\int_{-L}^{L}H_x^2dx + \frac{4W_e}{3}\int_{-L}^{L}H_x\mathcal{H}[H_{xx}]dx. \tag{4.30}$$

Using the Cauchy–Schwartz and Young's inequalities and the property $\|\mathcal{H}[u]\| = \|u\|$ gives

$$\begin{aligned}\int_{-L}^{L}H_x\mathcal{H}[H_{xx}]dx &\le \|H_x\|_2\|\mathcal{H}[H_{xx}]\|_2 = \|H_x\|_2\|H_{xx}]\|_2\\ &\le \frac{1}{2\varepsilon}\|H_x\|_2^2 + \frac{\varepsilon}{2}\|H_{xx}\|_2^2,\end{aligned} \tag{4.31}$$

where ε is some positive number. Hence,

$$\begin{aligned}\frac{d}{dt}\int_{-L}^{L}\frac{dx}{H} &\le -\frac{2}{3C}\|H_{xx}\|_2^2 - \frac{2G}{3}\|H_x\|_2^2 + \frac{4W_e}{3}\left(\frac{1}{2\varepsilon}\|H_x\|_2^2 + \frac{\varepsilon}{2}\|H_{xx}\|_2^2\right)\\ &= A\|H_{xx}\|_2^2 + B\|H_x\|_2^2,\end{aligned} \tag{4.32}$$

where $A = -2/3C + 2W_e\varepsilon/3$, $B = -2G/3 + 2W_e/3\varepsilon$. Choosing ε small enough (s.t. $A < 0$) implies

$$\frac{d}{dt}\int_{-L}^{L}\frac{dx}{H} \le B\|H_x\|_2^2. \tag{4.33}$$

Since the H^1-norm of the positive solution H is uniformly bounded (as was shown in the previous section), we obtain the desired boundedness result since there exists a constant D s.t. $(d/dt)\int_{-L}^{L}(1/H)dx \le D$.

It was proved in section 4.2 that the H^1-norm of a positive solution is bounded for all time. Due to Agmon's inequality this also implies boundedness of the maximum of the solution. Hence, if the solution is positive for all time, then it is also uniformly bounded above; i.e., it does not blow up in finite or infinite time. In this section we have shown that the spatial integral of $1/H$ is bounded on each finite time interval; i.e., it does not blow up in finite time (though this can happen in infinite time). This result can be used to show that as long as $\max|H_{xx}|$ remains bounded on each finite time interval, an initially positive solution will remain positive for all time; that is, the interface does not touch down in finite time. We prove this below. (The boundedness of $\max|H_{xx}|$ comes from extensive computations (see section 6) and is used as an assumption in what follows. A rigorous proof has not yet been found.)

PROPOSITION 4.3. *Let $H(x, t)$ be a positive smooth solution of* (3.12) *with periodic boundary conditions on some time interval $[0, T)$. In addition, assume that $\max|H_{xx}|$ is bounded above on each finite time interval. Then the solution H remains positive for all time, i.e., $T = \infty$.*

Proof. Suppose that T is finite and the solution becomes zero at some point $x = x_0$ in finite time $t = T$, and seek a contradiction. This means that at $t = T$ the solution obtains a minimum at $x = x_0$. Denote by $\xi(t)$ the point at which the solution has a (local) minimum at a given time t s.t. $\xi(t)$ is a continuous function of t and $\xi(T) = x_0$. Also, let $\max|H_{xx}| = A$, where A is a function of time which is bounded on $[0, T)$. At each time $t < T$ we expand the function H into a Taylor series about $x = \xi(t)$ and use the fact that $H_x(\xi(t), t) = 0$,

$$H(x, t) = H(\xi(t), t) + \frac{1}{2}(x - \xi(t))^2 H_{xx}(\zeta, t), \tag{4.34}$$

where $\zeta = \zeta(x, t)$ is some point between x and $\xi(t)$. We get

$$\begin{aligned}\int_{-L}^{L}\frac{dx}{H(x, t)} &= \int_{-L}^{L}\frac{dx}{H(\xi(t), t) + \frac{1}{2}(x - \xi(t))^2 H_{xx}(\zeta, t)}\\ &\ge \int_{-L}^{L}\frac{dx}{H(\xi(t), t) + \frac{A}{2}(x - \xi(t))^2}\\ &= -\sqrt{\frac{2}{AH(\xi(t), t)}}\tan^{-1}\left(\sqrt{\frac{A}{2H(\xi(t), t)}}(\xi(t) - x)\right)\Bigg|_{-L}^{L}. \qquad (4.35)\end{aligned}$$

It follows that the right-hand side of (4.35) blows up when $t \to T$, since then $H(\xi(t), t) \to 0$ by the assumption. This contradicts the boundedness of $\int_{-L}^{L}(1/H)dx$ on each finite time interval. Thus, if $\max|H_{xx}|$ is bounded on each finite time interval, then the solution cannot become zero in finite time; i.e., it remains positive for all time. ∎

5. Numerical method. We use a fully implicit two-level scheme with Newton iterations for (3.12) on a finite periodic interval $[-L, L]$. The method was developed for the following more general equation (this also includes the falling film equation derived in Tseluiko and Papageorgiou [28]):

$$H_t + \left[f_1(H)H_{xxx}\right]_x + \left[f_2(H)\right]_{xx} + \left[f_3(H)\mathcal{H}[H_{xx}]\right]_x + \left[f_4(H)\right]_x = 0, \tag{5.1}$$

with $f_1, \dots, f_4$ polynomials in H. We incorporate the ideas of Bertozzi and Pugh [2], Diez, Kondic, and Bertozzi [6] into nonlocal problems. The equation is solved on a uniform spatial grid $x_m = (m - M)\Delta x$, $m = 1, 2, \dots, 2M$, where $\Delta x = L/M$, and spatial derivatives are discretized using central differences; H_m denote the values of a $2L$-periodic function H at the mesh points. We also set $H_0 = H_{2M}$, $H_{-1} = H_{2M-1}$, etc., and $H_{2M+1} = H_1$, $H_{2M+2} = H_2$, etc., which follow by the periodicity of H. For $m = 1, 2, \dots, 2M - 1$, we introduce the midpoints $x_{m+1/2} = (x_m + x_{m+1})/2$ with $x_{1/2} = (-L + x_1)/2$ and define $H_{m+1/2} = (H_m + H_{m+1})/2$. Second order accurate central differences are used to approximate odd derivatives at $x_{m+1/2}$ and even derivatives at x_m. To approximate the Hilbert transform of H_{xx} at $x = x_{m+1/2}$ we use trapezoidal quadrature in the periodic representation $\mathcal{H}[g](x) = (1/2L)PV \int_{-L}^{L} g(\xi) \cot\big(\pi(x - \xi)/2L\big) d\xi$,

$$\mathcal{H}[H_{xx}](x_{m+1/2}) \approx \widetilde{\mathcal{H}}[\partial_2(H)]_{m+1/2} \equiv \frac{\Delta x}{2L} \sum_{k=1}^{2M} \partial_2(H)_k \cot\left(\frac{\pi(x_{m+1/2} - x_k)}{2L}\right). \tag{5.2}$$

This leads to the system of ordinary differential equations for $H_1, H_2, \dots, H_{2M}$:

$$\begin{aligned} \frac{dH_m}{dt} = &- \frac{f_1(H_{m+1/2})\partial_3(H)_{m+1/2} - f_1(H_{m-1/2})\partial_3(H)_{m-1/2}}{\Delta x} - \partial_2(f_2(H))_m \\ &- \frac{f_3(H_{m+1/2})\widetilde{\mathcal{H}}[\partial_2(H)]_{m+1/2} - f_3(H_{m-1/2})\widetilde{\mathcal{H}}[\partial_2(H)]_{m-1/2}}{\Delta x} \\ &- \frac{f_4(H_{m+1/2}) - f_4(H_{m-1/2})}{\Delta x}, \end{aligned} \tag{5.3}$$

where the grid operators ∂_2 and ∂_3 correspond to the second- and third order spatial derivatives, respectively. We write (5.3) in the following compact form:

$$\frac{d\boldsymbol{H}}{dt} = \boldsymbol{F}(\boldsymbol{H}), \tag{5.4}$$

where $\boldsymbol{H} = (H_1, H_2, \dots, H_{2M})^T$, $\boldsymbol{F}(\boldsymbol{H}) = (F_1(\boldsymbol{H}), F_2(\boldsymbol{H}), \dots, F_{2M}(\boldsymbol{H}))^T$ are given by the right-hand side of (5.3). Note that, unlike similar thin film problems that have been studied previously, $F_m(\boldsymbol{H})$ depends on all the components of $\boldsymbol{H}$ due to the presence of the nonlocal Hilbert transform.

This semidiscrete scheme preserves the discrete approximation of the volume. We show this by multiplying (5.3) by Δx and summing over $m = 1, 2, \dots, 2M$ to obtain $\sum_{m=1}^{2M} (dH_m/dt)\Delta x = 0$. A time integration yields $\sum_{m=1}^{2M} H_m(t)\Delta x = \sum_{m=1}^{2M} H_m(0)\Delta x$, as desired.

For the time discretization of (5.4) we use the usual implicit two-level scheme,

$$\frac{\boldsymbol{H}^{n+1} - \boldsymbol{H}^n}{\Delta t_n} = \boldsymbol{F}(\theta \boldsymbol{H}^{n+1} + (1 - \theta)\boldsymbol{H}^n), \tag{5.5}$$

where $\boldsymbol{H}^n = (H_1^n, H_2^n, \dots, H_{2M}^n)^T$ is the numerical solution for $\boldsymbol{H}$ at $t = t_n$, $\Delta t_n = t_{n+1} - t_n$, and θ is some real number between 0 and 1 (the scheme is first order accurate in time). To advance from the time level n to the time level $n + 1$, the algebraic system of nonlinear equations (5.5) for $\boldsymbol{H}^{n+1}$ is solved iteratively using Newton's method. The time step is chosen dynamically for each time level by requiring several constraints to be satisfied as described below (see also Bertozzi and Pugh [2] and Diez,

Kondic, and Bertozzi [6]). If the numerical solution violates one of the constraints, then the time step is reduced and the calculation is repeated until all the constraints are met. On the other hand, if all the constraints are met after the first application of Newton's method, the time step is increased at the next time level in order to prevent using unnecessarily small time steps. The constraints are the following: (a) the minimum of the solution should change by no more than 10%, (b) the local relative error should be small (10^{-3}, say). The local relative error e_m approximates $((\Delta t_{n-1})^2/H_m^n)(d^2H_m^n/dt^2)$ and is computed as follows (see [2], [6]):

$$e_m = \frac{2\Delta t_{n-1}}{\Delta t_{n-2}} \frac{\Delta t_{n-2} H_m^{n+1} + \Delta t_{n-1} H_m^{n-1} - (\Delta t_{n-2} + \Delta t_{n-1}) H_m^n}{(\Delta t_{n-2} + \Delta t_{n-1}) H_m^n}. \tag{5.6}$$

In addition, the spatial grid is refined during the calculation to get better resolution of the solution. This is done by doubling the number of mesh points when the magnitude of more than 2/3 of the Fourier modes is larger than a set tolerance of 10^{-13}. (The fast Fourier transform is used as an accuracy diagnostic.)

The numerical method has been described and implemented for solutions without any assumed symmetry. If $f_4 \equiv 0$, however, as is the case here, and the initial condition is an even function, then H remains even for all time. In this case we can consider $2L$-periodic even solutions and discretize the equation on the interval $[0, L]$ alone, thus halving the number of unknowns. The appropriate boundary conditions are

$$H_x(0, t) = H_{xxx}(0, t) = 0, \qquad H_x(L, t) = H_{xxx}(L, t) = 0, \tag{5.7}$$

with periodicity used as needed in calculating difference formulas.

6. Numerical results. The code was validated by reproducing the results of Yiantsios and Higgins [30], who solved (3.12) with $C = 1$, $G = -1$, $W_e = 0$, on periodic intervals of lengths $2\sqrt{2}\pi$, $4\sqrt{2}\pi$, $6\sqrt{2}\pi$, 5π. We also reproduced the results of Bertozzi and Pugh [3]. Part of their work involved the numerical solution of (3.12) on the interval $[-1, 1]$, with $C = 1$, $G = -80$, $W_e = 0$. Our code has reproduced these results with indistinguishable differences at $t = 100$, which is the largest time that Bertozzi and Pugh [3] report.

For the results presented here we take $C = 1$, $G = -1$ or $G = 1$, and increase W_e to enhance the instability. We take $L = 10$ and an initial condition

$$H(x, 0) = 1 + 0.1\cos(\pi x/L). \tag{6.1}$$

Thus, without loss of generality the mean value of $H(x, 0)$ over a period is taken to be 1. If it were $d > 0$, then the change of time scale, $t \to d^{-3}t$, leaves the evolution equation (3.12) unchanged but normalizes the initial condition to have unit mean.

6.1. Films wetting the underside of the plate, $G < 0$. As explained previously, when $G < 0$ the flat film state is long-wave unstable even when $W_e = 0$. We present results for fixed $G = -1$ and $C = 1$ as the electric field parameter W_e increases. For these parameters with $W_e = 0$ and $L = 10$, the first two harmonics are linearly unstable modes of the flat state, as can be seen from the linear result (6.2) below. We aim to systematically quantify the dynamics in the nonlinear regime as W_e increases.

The first set of results is presented in Figure 3 for $W_e = 0$. There are eight panels in the figure depicting the evolution over 1000 time units. The plots of the

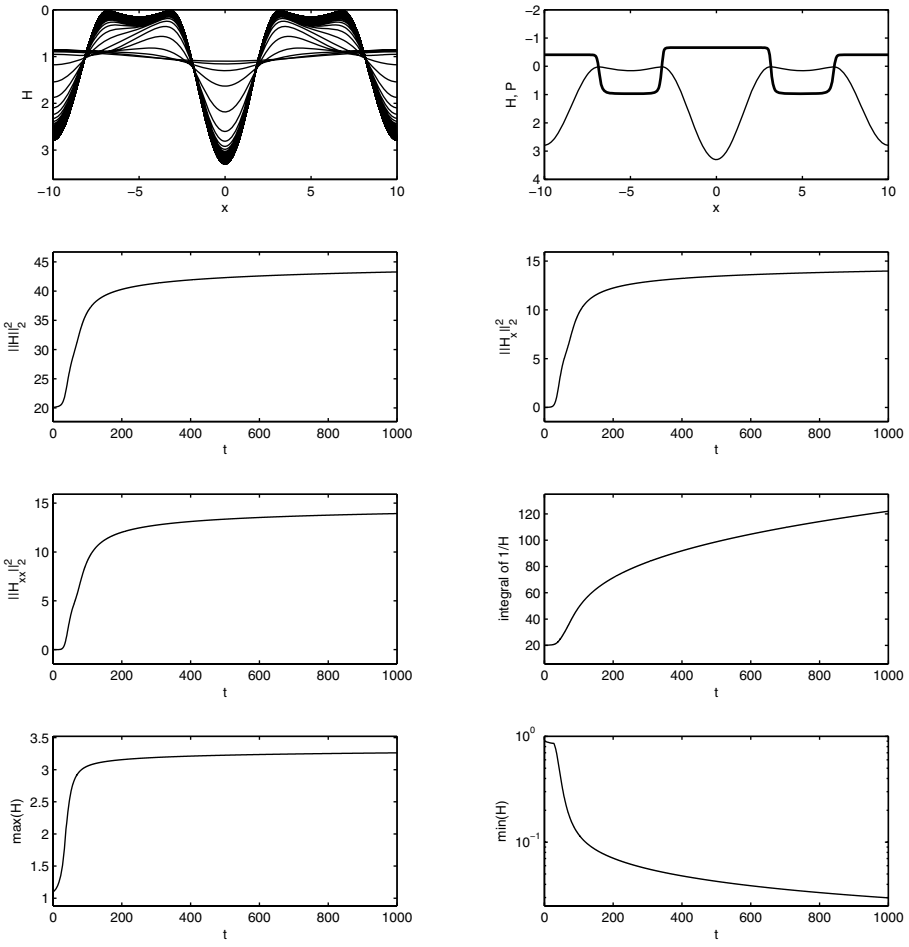


FIG. 3. *Evolution of the spatially periodic interface for* $C = 1$, $G = -1$, $W_e = 0$. *The equation was integrated for* $0 \le t \le 1000$. *The upper-left panel shows the evolution of the profile* H *(the time interval between the plots is* 10*). The upper-right panel shows the profile* H *(thin line) and the pressure* P *given by* (3.11) *(solid line) at* $t = 1000$. *Also, the evolution of* $\|H\|_2^2$, $\|H_x\|_2^2$, $\|H_{xx}\|_2^2$, *as well as the evolution of* $\int_{-10}^{10}(1/H)dx$ *and the maximum and minimum of* H *are shown. (For the minimum we use a log-linear plot.)*

interface are reflected about the x-axis to emphasize that we are dealing with hanging films. The interface evolution is shown in the top-left panel, and the top-right panel shows the solution at the last computed time $t = 1000$; the thin line curve represents the interface $H(x, 1000)$, and the thick line curve the corresponding perturbation pressure distribution $P(x, 1000)$ given by (3.11) (note that the subscript zero has been dropped). It can be seen that the pressure is essentially uniform and negative in the regions where large drops are forming (the pressure has different values in different-sized drops), and uniform and positive in the thinning regions between large drops. The resulting pressure gradient acts to push fluid out of the thinning regions and into the larger drops. This mechanism is at play for all computed results presented here and is responsible for the asymptotic thinning of the regions between the larger

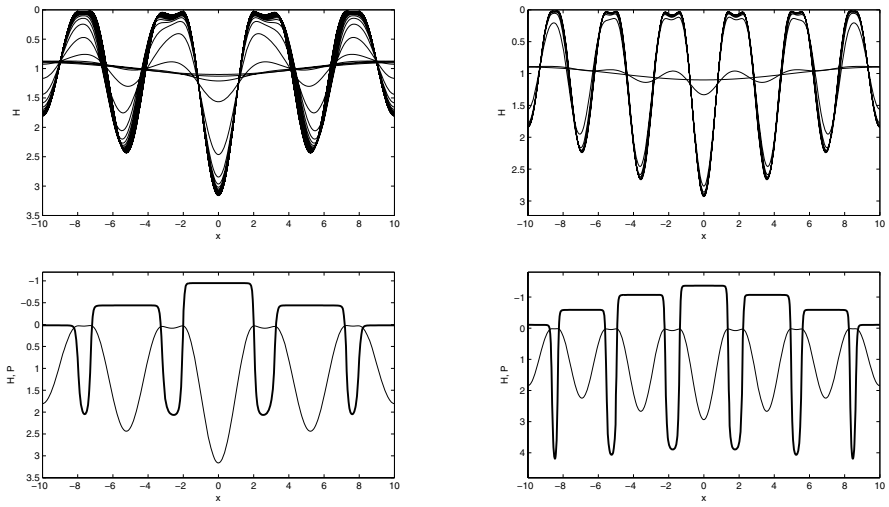


FIG. 4. *Evolution for $C = 1$, $G = -1$. The top row shows the interface evolution for $W_e = 0.5$ (left) and 1 (right); the bottom row shows the corresponding profiles H (thin line) and the pressures P (solid line) at the final times.*

quasi-static drops. The other six panels contain information on the evolution of different norms of H, namely $\|H\|_2^2$, $\|H_x\|_2^2$, and $\|H_{xx}\|_2^2$ (labeled in the figure); the evolution of $\int_{-L}^{L}(1/H)dx$; and the evolution of the maximum and minimum of H over the spatial domain, $\max(H)$ and $\min(H)$, respectively (the latter is plotted using log-linear scales). The boundedness of $\|H\|_2^2$ and $\|H_x\|_2^2$ (equivalently of the L^2- and H^1-norms) is in line with the rigorous results of section 4, as is the at most linear growth of the integral $\int_{-L}^{L}(1/H)dx$. The evolution of $\|H_{xx}\|_2^2$ (equivalently the H^2-norm) indicates that it is bounded (moreover, from numerics it can be seen that $\max|H_{xx}|$ is bounded as well)—this is used as an assumption in producing a proof that the interface cannot touch down in finite time but can do so asymptotically in infinite time (see section 4.3). The log-linear evolution of $\min(H)$ also provides strong evidence of an asymptotic touchdown after infinite time in line with the conjecture in section 4.3. Finally, we note that the profile at large time contains two large drops (see the top-right panel), and this number coincides with the number of linearly unstable modes. Nonlinearity produces small daughter drops between the main mother drops that cannot be predicted by linear theory. This phenomenon is generic and holds when W_e is nonzero also (see Figures 4, 5, 6 also).

Figure 4 contains results for nonzero electric fields with $W_e = 0.5$ and 1. The top row shows the evolution of the interface for $W_e = 0.5$ (left) and $W_e = 1$ (right), and the bottom row shows the corresponding final computed interfacial profiles and pressure distributions. The calculations were carried out to 100 and 30 time units for $W_e = 0.5$ and 1, respectively, and profiles in the top row are depicted at intervals of 4 time units. Once more we see main drops forming at large times with thinning regions between them containing smaller humps. The pressure distribution is uniform in the larger drops with maxima in the thinning regions, producing the draining mechanism discussed earlier. The main difference between the two cases is that for $W_e = 0.5$ we ultimately have four drops forming, while for $W_e = 1$ we have six. This can be explained using linear theory and (3.13). The wavenumber of the maximally unstable

mode on 2π-periodic domains is $k_{max} = (3CW_e + \sqrt{9(CW_e)^2 - 8CG})/4$. Modifying this to $2L$-periodic domains gives

$$k_{max} = \left(\frac{L}{\pi}\right) \frac{3CW_e + \sqrt{9(CW_e)^2 - 8CG}}{4}. \tag{6.2}$$

The value of k_{max} provides a qualitative estimate of the main features of the interface at large times; for example, in the results of Figure 3 we have $k_{max} = 2.25$, which explains the two large drops that form. For the parameters of Figure 4 we have $k_{max} = 3.74$ and $k_{max} = 5.67$, which explain the four and six drops formed. The smaller drops forming in the thinning regions are due to nonlinearity and cannot be explained using a simple linear theory. We have also monitored norms and other diagnostics as in Figure 3 and have found similar behavior. Most notably, $\max |H_{xx}|$ remains bounded in time.

6.2. Films wetting the upper side of the plate, $G > 0$. Here we present results for $G = 1$, $C = 1$, and increasing values of W_e. As noted earlier, if $W_e = 0$, the flat state is stable—the solutions to the initial value problem are damped and produce the uniform trivial state at large times (this has been confirmed numerically also). Support for this is also provided by the linear result (3.13), since $s(k) < 0$ for all $k \neq 0$. If W_e exceeds the critical value $W_{ec} = (G/C)^{1/2}$, instability sets in over a band of wavenumbers $k_L < k < k_R$, using the notation of section 3. For our parameters, $W_{ec} = 1$, and in what follows we present results for increasing $W_e > 1$.

The first set of results has $W_e = 1.02$, which is just above critical. According to (6.2), $k_{max} = 3.37$, and so we may expect three drops to form at large times. The results are shown in Figure 5; the integration was carried out to $t = 5000$. The format of the figure is the same as that of Figure 3, and the results are qualitatively similar. The top-right panel shows an enlargement of the solution and the corresponding pressure distribution in the vicinity of the first thinning region to the left of the origin. Again we see essentially uniform negative pressure in the main drops and a pressure maximum in the thinning region in between, so that the fluid draining mechanism described earlier is seen to operate. The other diagnostics are in agreement with the analytical results.

In Figure 6 we present results for $W_e = 1.1$, 1.5, and 2.0, respectively. The top row shows the time evolution of H as W_e increases from the left to the right, and the bottom row shows the corresponding final computed profiles and pressure distributions (thin and thick solid lines, respectively)—for $W_e = 1.1$ this is enlarged accordingly. The computations were carried out to 1000, 75, and 20 time units for $W_e = 1.1$, 1.5, and 2, respectively, and profiles are plotted every 10, 1, and 1 time units, respectively. The pressure gradient draining mechanism is operational throughout, and uniform but different negative pressures are attained inside the large drops. The number of drops formed at large times is again in excellent agreement with linear theory (linear theory cannot provide the volumes of the drops or the formation of smaller daughter droplets in the thinning regions). For example, the values of k_{max} given by (6.2) are 3.98, 6.37, and 8.99 for $W_e = 1.1$, 1.5, and 2.0, respectively, while the numbers of the main computed drops are 4, 6, and 9, respectively. It can be concluded, therefore, that linear theory can be used in a very simple way to predict how many drops will form at large time. Details, including drop volumes, must be calculated numerically by solving the nonlinear problem.

7. Conclusions. We have derived and studied analytically and numerically a nonlinear nonlocal evolution equation that describes the evolution of thin films wetting

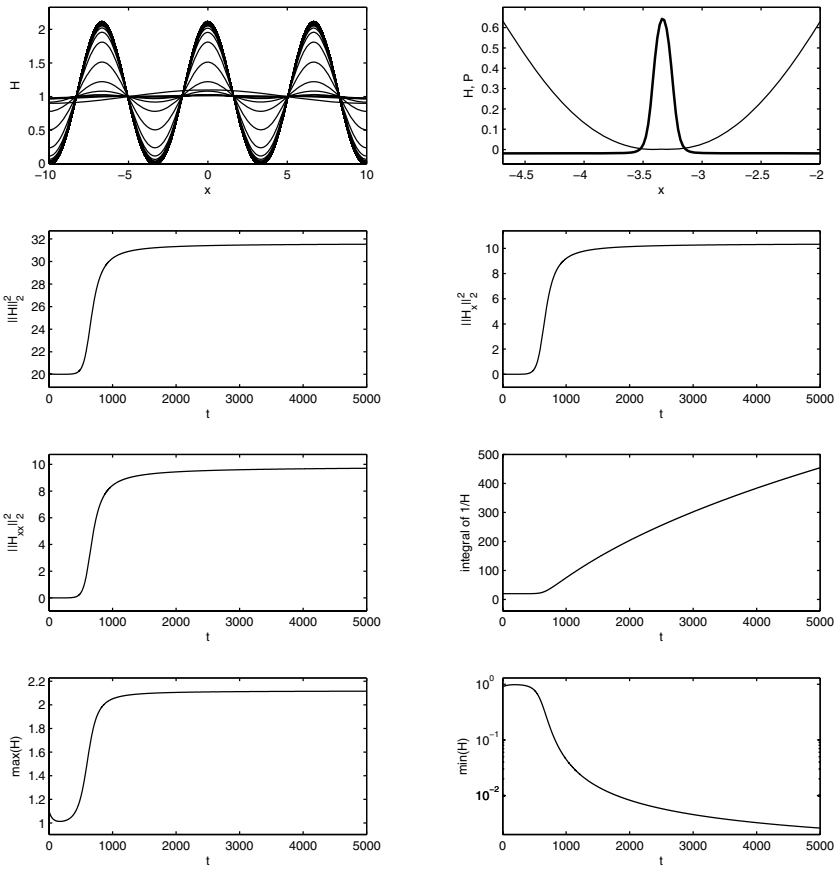


FIG. 5. *Evolution of the spatially periodic interface for* $C = 1$, $G = 1$, $W_e = 1.02$. *The equation was integrated for* $0 \le t \le 5000$. *The upper-left panel shows the evolution of the profile* H *(the time interval between the plots is* 100*). The upper-right panel shows the profile* H *(thin line) and the pressure* P *given by* (3.11) *(solid line) at the final time. Also, the evolution of* $\|H\|_2^2$, $\|H_x\|_2^2$, $\|H_{xx}\|_2^2$, *as well as the evolution of* $\int_{-10}^{10}(1/H)dx$ *and the maximum and minimum of* H, *are shown (for the minimum we use a log-linear plot).*

a horizontal plate in the presence of a vertical electric field. The field introduces the instability when $G > 0$ (films wetting the upper side of the plate) and enhances the instability when $G < 0$ (films wetting the underside side of the plate).

By extending previous analytical studies to incorporate nonlocal terms, we have proved a no-blow-up theorem of positive smooth solutions of the evolution equation. Using an estimate of the integral of the reciprocal of the solution and assuming that $\max|H_{xx}|$ is uniformly bounded (this is suggested by extensive numerical work), we have also presented a conjecture that the film cannot touch down in finite time but can do so only asymptotically in infinite time. This also holds in the absence of the field. All rigorous results are seen in the numerical solutions, thus providing additional accuracy checks for the latter.

Extensive numerical experiments have been carried out to describe the salient features of thin electrified film dynamics. Initially, the evolution follows the predictions of linear theory, and the solution grows exponentially. As time increases, higher

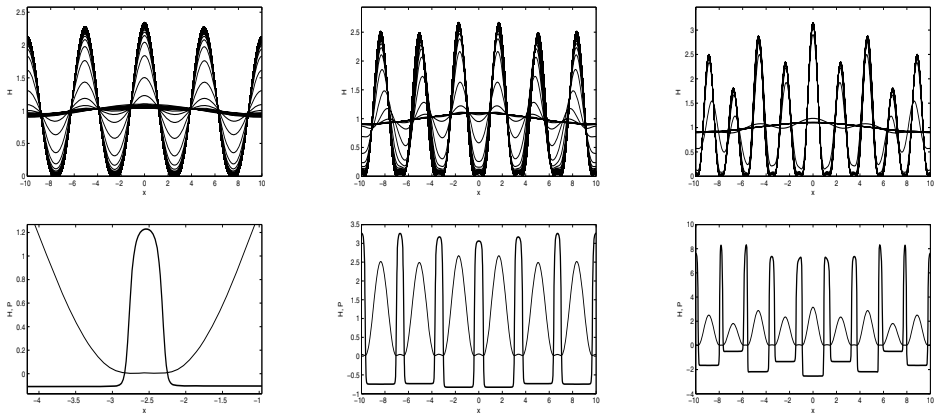


FIG. 6. *Evolution for $C = 1$, $G = 1$. The top row shows the interface evolution for $W_e = 1.1$, 1.5, and 2, from left to right; the bottom row shows the corresponding profiles H (thin line) and the pressures P (solid line) at the final times.*

harmonics are generated due to the nonlinearities, and the most dominant mode appears to be the most unstable mode predicted by linear theory; note that this mode corresponds to the number of the drops which appear at large times. The qualitative features of the solutions for $G < 0$ and W_e zero or nonzero are similar to those with $G > 0$ and $W_e > 0$ (in the latter case a nonzero electric field is required to destabilize the flow and to produce nontrivial dynamics). An increase in W_e for fixed G and C (or equivalently a decrease of a negative G with fixed W_e and C), produces increasingly more drops at large times, whose number is predicted by linear theory. This drop-formation behavior is one of the main features of the dynamics as additional unstable modes enter. In all computed cases, as the time increases the evolution slows down (this can be seen in any of the different computational panels in the figures, but is most clearly evidenced by the evolution of $\min(H)$). The spatial features at large times are also quite intricate: First, as the interface reaches the vicinity of the wall it tends to flatten, and after that the solution tends to bulge near the ends of the flat regions forming a secondary hump in between—see Figure 3, for example. All the results indicate that the solution remains positive—the film does not touch down within finite time. Also, the solution is bounded for all time (this has been proven rigorously), despite the fact that the electric field increases the instability and promotes the process of the formation of the increasingly larger numbers of drops.

Finally, we comment on the possibility of coarsening dynamics, as seen in other thin film studies by Glasner and Witelski [12], for example. Even though it is not clear a priori whether neighboring drops communicate with each other in order to trigger merging and coarsening, the numerical solutions with electric fields present suggest that such dynamics is not seen. Our results suggest that the thinning of the interdrop regions feeds fluid into main drops and that the large time dynamics of the latter remain independent from each other, drops remaining fixed and not moving. It would be interesting to add a disjoining pressure in the manner of Glasner and Witelski [12] that prevents asymptotic thinning and allows communication between main drops. It is also interesting to revisit the calculations of Yiantsios and Higgins [30] for $W_e = 0$, where they do not impose symmetry and see that drops move (slowly) with the larger drop moving more. The calculations failed to see merging, however, due to a critical

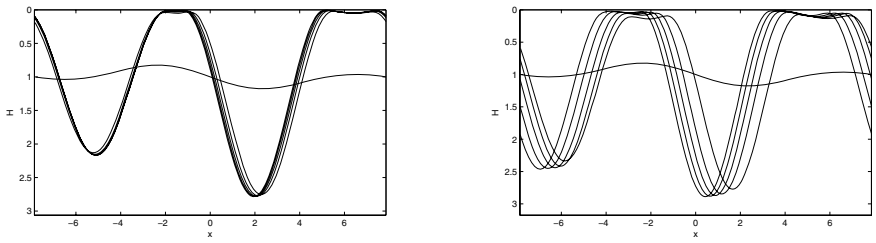


FIG. 7. *Nonsymmetric initial conditions for hanging drops, $C = 1$, $G = -1$. Left panel, $W_e = 0$, and right panel, $W_e = 0.1$. The domain length is 5π, and the initial condition in both runs is $H(x,0) = 1 + 0.1\sin(2x/5) + 0.1\sin(4x/5)$. As time increases, the minimum film thickness decreases.*

slow-down of drop motion as thinning of interdrop regions takes place. Following [30], we take $C = 1$, $G = -1$, and $W_e = 0$ and 0.1 with the length $L = 2.5\pi$ and the initial condition $H(x,0) = 1 + 0.1\sin(2x/5) + 0.1\sin(4x/5)$. The results are shown in Figure 7, with $W_e = 0$ for the left panel and $W_e = 0.1$ for the right panel. For $W_e = 0$ two drops are formed and are approaching each other, with the larger drop being more mobile and moving from the right to the left, while the smaller drop moves from the left to the right. The rate of approach slows down critically (the integration is carried out to 2000 time units, and the profiles are shown every 400 units). For $W_e = 0.1$, two drops are formed but move in the same direction (from the right to the left) with the electric field present. The electric field significantly increases the mobility of the drops. The velocity of the larger drop is slowing down, while the velocity of the smaller drop is increasing with time, and the distance between the drops is decreasing with increasing time. We did not see merging again due to a critical slow-down of drop motion (the integration is carried out to 500 time units, and the profiles are shown every 100 units).

REFERENCES

[1] A. L. BERTOZZI, *The mathematics of moving contact lines in thin liquid films*, Notices Amer. Math. Soc., 45 (1998), pp. 689–697.

[2] A. L. BERTOZZI AND M. C. PUGH, *The lubrication approximation for thin viscous films: The moving contact line with 'porous media' cut off of Van der Waals interactions*, Nonlinearity, 7 (1994), pp. 1535–1564.

[3] A. L. BERTOZZI AND M. C. PUGH, *Long-wave instabilities and saturation in thin film equations*, Comm. Pure Appl. Math., 51 (1998), pp. 625–661.

[4] A. L. BERTOZZI AND M. C. PUGH, *Finite-time blow-up of solutions of some long-wave unstable thin film equations*, Indiana Univ. Math. J., 49 (2000), pp. 1323–1366.

[5] R. V. CRASTER AND O. MATAR, *Electrically induced pattern formation in thin leaky dielectric films*, Phys. Fluids, 17 (2005), paper 032104.

[6] J. A. DIEZ, L. KONDIC, AND A. L. BERTOZZI, *Global models for moving contact lines*, Phys. Rev. E, 63 (2000), paper 011208.

[7] E. B. V. DUSSAN, *On the spreading of liquids on solid surfaces, static and dynamic contact angles*, Ann. Rev. Fluid Mech., 11 (1979), pp. 371–400.

[8] P. EHRHARD, *The spreading of hanging drops*, J. Colloid Interface Sci., 168 (1994), pp. 242–246.

[9] P. EHRHARD AND S. H. DAVIS, *Nonisothermal spreading of liquid drops on horizontal plates*, J. Fluid. Mech., 229 (1991), pp. 365–388.

[10] E. D. EIDELMAN, *Parabolic Systems*, North-Holland, Amsterdam, 1969.

[11] A. FRIEDMAN, *Interior estimates for parabolic systems of partial differential equations*, J. Math. Mech., 7 (1958), pp. 393–418.

[12] K. B. GLASNER AND T. P. WITELSKI, *Coarsening dynamics of dewetting films*, Phys. Rev. E, 67 (2003), paper 016302.

[13] H. P. Greenspan, *On the motion of a small viscous droplet that wets a surface*, J. Fluid Mech., 84 (1978), pp. 125–143.
[14] P. J. Haley and M. J. Miksis, *The effect of the contact line on droplet spreading*, J. Fluid Mech., 223 (1991), pp. 57–81.
[15] T. Hocherman and P. Rosenau, *On KS-type equations describing the evolution and rupture of a liquid interface*, Phys. D, 67 (1993), pp. 113–125.
[16] L. M. Hocking, *Rival contact-angle models and the spreading of drops*, J. Fluid Mech., 239 (1992), pp. 671–681.
[17] Z. Lin, T. Kerle, E. Schaffer, U. Steiner, and T. P. Russel, *Structure formation at the interface of liquid/liquid bilayers in electric fields*, Macromolecules, 35 (2002), pp. 3971–3976.
[18] T. G. Myers, *Thin films with high surface tension*, SIAM Rev., 40 (1998), pp. 441–462.
[19] A. Oron, S. H. Davis, and S. G. Bankoff, *Long-scale evolution of thin liquid films*, Rev. Mod. Phys., 69 (1997), pp. 931–980.
[20] D. T. Papageorgiou and J.-M. Vanden-Broeck, *Large amplitude capillary waves in electrified fluid sheets*, J. Fluid. Mech., 508 (2004), pp. 71–88.
[21] L. F. Pease and W. B. Russel, *Linear stability analysis of thin leaky dielectric films subjected to electric fields*, J. Non-Newtonian Fluid Mech., 102 (2002), pp. 233–250.
[22] K. Savettaseranee, P. G. Petropoulos, D. T. Papageorgiou, and B. S. Tilley, *The effect of electric fields on the rupture of thin viscous films by Van der Waals forces*, Phys. Fluids, 15 (2003), pp. 641–652.
[23] E. Schaffer, T. Thurn-Albrecht, T. P. Russel, and U. Steiner, *Electrically induced structure formation and pattern transfer*, Nature, 403 (2000), pp. 874–877.
[24] E. Schaffer, T. Thurn-Albrecht, T. P. Russel, and U. Steiner, *Electrohydrodynamic instabilities in polymer films*, Europhys. Lett., 53 (2001), pp. 518–524.
[25] V. Shankar and A. Sharma, *Instability of the interface between thin fluid films subjected to electric fields*, J. Colloid Interface Sci., 274 (2004), pp. 294–308.
[26] G. I. Taylor and A. D. McEwan, *The stability of a horizontal fluid interface in a vertical electric field*, J. Fluid Mech., 22 (1965), pp. 1–16.
[27] B. S. Tilley, P. G. Petropoulos, and D. T. Papageorgiou, *Dynamics and rupture of planar electrified liquid sheets*, Phys. Fluids, 13 (2001), pp. 3547–3563.
[28] D. Tseluiko and D. T. Papageorgiou, *Wave evolution on electrified falling films*, J. Fluid Mech., 556 (2006), pp. 361–386.
[29] T. P. Witelski, A. J. Bernoff, and A. L. Bertozzi, *Blowup and dissipation in a critical-case unstable thin film equation*, European J. Appl. Math., 15 (2004), pp. 223–256.
[30] S. G. Yiantsios and B. G. Higgins, *Rayleigh–Taylor instability in thin viscous films*, Phys. Fluids A, 1 (1989), pp. 1484–1501.
[31] S. G. Yiantsios and B. G. Higgins, *Rupture of thin films: Nonlinear stability analysis*, J. Colloid Interface Sci., 147 (1991), pp. 341–350.

SIAM J. APPL. MATH.
Vol. 67, No. 5, pp. 1330–1352

ELASTIC SCATTERER RECONSTRUCTION VIA THE ADJOINT SAMPLING METHOD*

S. NINTCHEU FATA† AND B. B. GUZINA‡

Abstract. An inverse problem dealing with the reconstruction of voids in a uniform semi-infinite solid from near-field elastodynamic waveforms is investigated via the linear sampling method. To cater to active imaging applications that are characterized by a limited density of illuminating sources, existing formulation of the linear sampling method is advanced in terms of its adjoint statement that features integration over the receiver surface rather than its source counterpart. To deal with an ill-posedness of the integral equation that is used to reconstruct the obstacle, the problem is solved by alternative means of Tikhonov regularization and a preconditioned conjugate gradient method. Through a set of numerical examples, it is shown (i) that the adjoint statement elevates the performance of the linear sampling method when dealing with scarce illuminating sources, and (ii) that a combined use of the existing formulation together with its adjoint counterpart represents an effective tool for exposing an undersampling of the experimental input, e.g., in terms of the density of source points used to illuminate the obstacle.

Key words. inverse scattering, elastic waves, near-field waveforms, linear sampling

AMS subject classifications. 35R30, 65R30, 74J05, 74J20

DOI. 10.1137/060653123

1. Introduction.

Remote sensing of internal defects or obstacles using elastic waves with "long" wavelengths, i.e., those inside the so-called resonance region [8], is an inverse scattering problem relevant to a variety of applications such as nondestructive material testing, hydrocarbon prospecting, and medical diagnosis. In the context of seismic inversion, such low-frequency waveforms are often interpreted by means of the full waveform tomography [36], which typically couples gradient-based nonlinear minimization with a finite-difference (forward) simulation of elastic wave propagation [4, 35]. For simple (e.g., homogeneous) background media, the waveform tomography approach to seismic imaging can be alternatively established within the framework of elastodynamic boundary integral equation methods, especially when equipped by the analytical sensitivity estimates [2, 19] (see also [6] for acoustic problems). Irrespective of the type of forward model, however, the high resolution of full waveform inversion is commonly balanced by its lack of robustness, manifest in the "trapping" of iterative solution within local minima [31, 36].

Over the past decade, Colton and coworkers [7, 9, 11] introduced an alternative, point-probing technique for solving inverse scattering problems in acoustics and electromagnetism, the so-called linear sampling method (LSM), that circumvents many of the foregoing impediments. This minimization-free approach to waveform tomography makes use of an ill-posed integral equation, written with reference to the obstacle-free domain, whose kernel is constructed from the observed waveforms and whose solution norm (used as an obstacle indicator) remains bounded *only* for sampling points strik-

*Received by the editors February 27, 2006; accepted for publication (in revised form) March 12, 2007; published electronically July 11, 2007.
http://www.siam.org/journals/siap/67-5/65312.html
†Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831 (nintcheufats@ornl.gov).
‡Corresponding author. Department of Civil Engineering, University of Minnesota, Minneapolis, MN 55455 (guzina@ wave.ce.umn.edu). This author's research was supported by National Science Foundation grant CMS-324348.

ing the support of the scatterer. Owing to its computational efficiency and relative robustness (stemming from the circumvention of nonlinear optimization), the LSM has since been adapted to both far-field [1, 5, 34] and near-field [32] elastic scattering problems. Despite its inherent appeal, however, the existing formulation of the LSM for active imaging configurations in near-field elastodynamics [32], which postulates integration over the source region, may not be applicable to testing arrangements that are characterized by a limited density of "illuminating" sources (e.g., magnetic resonance elastography [23, 12, 39]). Beyond the issue of source *density*, the existing algorithm in [32], which makes use of the singular value decomposition, may face an additional set of difficulties related to significant computational cost and inaccurate singular values when dealing with large *amounts* of experimental data (e.g., large numbers of sources and receivers, regardless of their density) as in three-dimensional (3D) seismic imaging [38].

To transcend the foregoing impediments, the focus of this study is two-fold and includes (i) a reformulation of the LSM for near-field elastodynamics, to cater for active imaging configurations with only a limited density of excitation sources, and (ii) computational treatment of ill-conditioned linear systems, which establishes an alternative to singular value decomposition in situations involving significant amounts of experimental data. To this end, an *adjoint statement* of the so-called direct sampling method in [32] is proposed, wherein the inverse problem is formulated as a linear integral equation of the first kind, involving integration over the *measurement* (as opposed to the source) surface, whose solution becomes unbounded in the exterior of a hidden scatterer. A preconditioned conjugate gradient algorithm for solving ill-posed linear systems, established earlier for X-ray computed tomography [37], is adapted to provide an alternative computational treatment of the LSM in dealing with extensive data sets. A set of numerical examples is included to illustrate the performance of the proposed developments.

2. Problem formulation. This investigation deals with time-harmonic elastic wave imaging of an obstacle Ω_C that is strictly embedded in a uniform, isotropic, semi-infinite solid; see Figure 2.1. With reference to the Cartesian frame $\{O;\xi_1,\xi_2,\xi_3\}$ set at the top of the half-space, the background domain $\Omega = \{(\xi_1,\xi_2,\xi_3)|\,\xi_3>0\}$ is characterized by the Lamé constants λ and μ and the mass density ρ; its free surface $\{(\xi_1,\xi_2,\xi_3)|\,\xi_3=0\}$ is denoted by Σ. Adopting further the hypothesis of an "impenetrable" scatterer, Ω_C is taken to be in the form of a *cavity* with smooth boundary Γ of class $C^{1,\alpha}$, $\alpha\in(0,1]$. For further reference, let

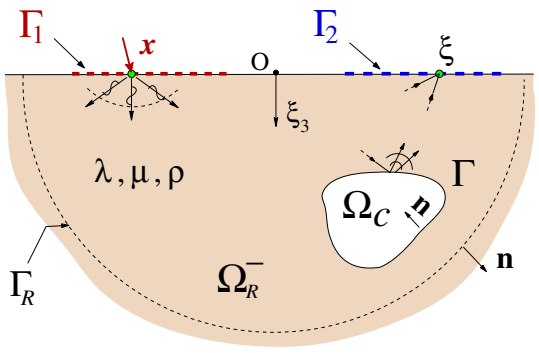


FIG. 2.1. *Illumination of an impenetrable obstacle in the semi-infinite solid by elastic waves.*

$\Omega^- = \Omega \setminus (\Omega_{\rm C} \cup \Gamma)$ denote the unbounded region surrounding the obstacle, and let $\Gamma_{\rm R} \subset \Omega$ be a hemisphere of radius R centered at the origin. The respective subsets of Ω and Ω^- that are bounded by $\Gamma_{\rm R}$ are denoted by $\Omega_{\rm R}$ and $\Omega^-_{\rm R}$, with an implicit assumption that R is sufficiently large such that $\Omega_{\rm C} \subset \Omega^-_{\rm R}$. As implied by the figure, the scatterer $\Omega_{\rm C}$ is exposed by time-harmonic sources acting on the source surface $\Gamma_1 \subset \Sigma$, with the induced solid motion monitored over the measurement area $\Gamma_2 \subset \Sigma$. In what follows, the frequency of excitation is denoted by ω, with the implicit time-harmonic factor $\mathrm{e}^{\mathrm{i}\omega t}$ omitted for brevity.

2.1. Direct scattering problem. In situations where the excitation source used to illuminate the obstacle is a *point force* of unit magnitude acting at $\boldsymbol{x} \in \Gamma_1$ in the kth coordinate direction, the elastodynamic displacement response of the perturbed reference solid Ω^- can be conveniently decomposed as

$$\boldsymbol{u}^k(\boldsymbol{\xi}, \boldsymbol{x}) \,=\, \mathring{\boldsymbol{u}}^k(\boldsymbol{\xi}, \boldsymbol{x}) + \tilde{\boldsymbol{u}}^k(\boldsymbol{\xi}, \boldsymbol{x}), \qquad \boldsymbol{\xi} \in \Omega^-, \quad \boldsymbol{x} \in \Gamma_1, \tag{2.1}$$

where $\tilde{\boldsymbol{u}}^k$ represents the scattered field ($\tilde{\boldsymbol{u}}^k \equiv \boldsymbol{0}$ in the absence of a scatterer) and $\mathring{\boldsymbol{u}}^k$ denotes the free field, i.e., the response of the obstacle-free solid Ω due to prescribed excitation. With such definitions, one has

$$\mathring{\boldsymbol{u}}^k(\boldsymbol{\xi}, \boldsymbol{x}) = \hat{\boldsymbol{u}}^k(\boldsymbol{\xi}, \boldsymbol{x}), \quad \boldsymbol{\xi} \neq \boldsymbol{x}, \quad \boldsymbol{\xi} \in \Omega, \quad \boldsymbol{x} \in \Gamma_1, \tag{2.2}$$

where $\hat{\boldsymbol{u}}^k(\boldsymbol{\xi}, \boldsymbol{x})$ is the elastodynamic (displacement) Green's function for a homogeneous semi-infinite solid [18] at $\boldsymbol{\xi} \in \Omega$ due to a unit time-harmonic point force acting at $\boldsymbol{x} \in \Gamma_1$ in the kth direction.

With reference to any smooth surface $S \subset \Omega$ with unit normal $\boldsymbol{n}$, it is further useful to introduce the traction vector

$$\boldsymbol{t}(\boldsymbol{\xi}; \boldsymbol{u}) = \boldsymbol{n}(\boldsymbol{\xi}) \cdot \mathbf{C} : \boldsymbol{\nabla} \boldsymbol{u}(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \in S, \tag{2.3}$$

associated with the displacement field $\boldsymbol{u}$, where $\mathbf{C} = \lambda\, \mathbf{I}_2 \otimes \mathbf{I}_2 \,+\, 2\,\mu\, \mathbf{I}_4$ denotes the isotropic elasticity tensor and $\mathbf{I}_k$ ($k = 2, 4$) is the symmetric kth order identity tensor.

On the basis of (2.1)–(2.3), the forward problem associated with Figure 2.1 can be specified as a task of resolving the scattered field $\tilde{\boldsymbol{u}}^k$ from the knowledge of the free field $\mathring{\boldsymbol{u}}^k$ and the exact geometry of (impenetrable) scatterer $\Omega_{\rm C}$. More precisely, one is to find a $\tilde{\boldsymbol{u}}^k \in C^2(\Omega^-) \cap C^1(\Omega^- \cup \Gamma \cup \Sigma)$ satisfying the homogeneous Navier equation

$$\boldsymbol{L} \tilde{\boldsymbol{u}}^k(\boldsymbol{\xi}, \boldsymbol{x}) + \rho \omega^2 \tilde{\boldsymbol{u}}^k(\boldsymbol{\xi}, \boldsymbol{x}) = \boldsymbol{0}, \quad \boldsymbol{\xi} \in \Omega^-, \quad \boldsymbol{x} \in \Gamma_1, \tag{2.4}$$

and Neumann boundary conditions

$$\tilde{\boldsymbol{t}}^k(\boldsymbol{\xi}, \boldsymbol{x}) = \begin{cases} \boldsymbol{0}, & \boldsymbol{\xi} \in \Sigma, \\ -\mathring{\boldsymbol{t}}^k(\boldsymbol{\xi}, \boldsymbol{x}), & \boldsymbol{\xi} \in \Gamma, \end{cases} \qquad \boldsymbol{x} \in \Gamma_1, \tag{2.5}$$

where $\mathring{\boldsymbol{t}}^k = \boldsymbol{t}(\boldsymbol{\xi}; \mathring{\boldsymbol{u}}^k)$; $\tilde{\boldsymbol{t}}^k = \boldsymbol{t}(\boldsymbol{\xi}; \tilde{\boldsymbol{u}}^k)$ is understood in the sense of the trace [28, 30], and $\boldsymbol{L} = \mu \boldsymbol{\nabla}^2 + (\lambda + \mu) \boldsymbol{\nabla} \boldsymbol{\nabla} \cdot$ is the Lamé operator (see also [15]). To ensure the well-posedness of the forward scattering problem, the scattered field $\tilde{\boldsymbol{u}}^k$ must also satisfy the generalized radiation condition

$$\mathcal{R}(\tilde{\boldsymbol{u}}^k) := \lim_{R \to \infty} \int_{\Gamma_{\rm R}} \left\{ \hat{\boldsymbol{u}}^j(\boldsymbol{\xi}, \boldsymbol{y}) \cdot \tilde{\boldsymbol{t}}^k(\boldsymbol{\xi}, \boldsymbol{x}) - \hat{\boldsymbol{t}}^j(\boldsymbol{\xi}, \boldsymbol{y}) \cdot \tilde{\boldsymbol{u}}^k(\boldsymbol{\xi}, \boldsymbol{x}) \right\} dS_{\boldsymbol{\xi}} = 0, \tag{2.6}$$

$$\boldsymbol{y} \in \Omega, \quad j = 1, 2, 3,$$

common to all radiating solutions in a semi-infinite elastic solid [27]. In (2.6), $\hat{\boldsymbol{t}}^j(\boldsymbol{\xi},\boldsymbol{y}) = \boldsymbol{t}(\boldsymbol{\xi};\hat{\boldsymbol{u}}^j)$ is the traction vector associated with $\hat{\boldsymbol{u}}^j(\boldsymbol{\xi},\boldsymbol{y})$, i.e., the elastodynamic traction Green's function for the reference half-space Ω. As shown in [19], the generalized radiation condition applies equally when $\hat{\boldsymbol{u}}^j$ and $\hat{\boldsymbol{t}}^j$ in (2.6) are superseded by the respective displacements and tractions corresponding to *any radiating* elastodynamic state in Ω^-. To facilitate the ensuing presentation, this result can be applied to establish the equality

$$\lim_{R\to\infty}\int_{\Gamma_R}\left\{\tilde{\boldsymbol{u}}^j(\boldsymbol{\xi},\boldsymbol{y})\cdot\tilde{\boldsymbol{t}}^k(\boldsymbol{\xi},\boldsymbol{x})-\tilde{\boldsymbol{t}}^j(\boldsymbol{\xi},\boldsymbol{y})\cdot\tilde{\boldsymbol{u}}^k(\boldsymbol{\xi},\boldsymbol{x})\right\}dS_{\boldsymbol{\xi}} \;=\; 0, \qquad \boldsymbol{x},\boldsymbol{y}\in\Omega^-. \tag{2.7}$$

For completeness, it is also noted that the Green's functions featured in (2.6) satisfy

$$\begin{aligned} &\boldsymbol{L}\hat{\boldsymbol{u}}^j(\boldsymbol{\xi},\boldsymbol{y})+\rho\omega^2\hat{\boldsymbol{u}}^j(\boldsymbol{\xi},\boldsymbol{y})+\delta(\boldsymbol{\xi}-\boldsymbol{y})\boldsymbol{e}^j=\boldsymbol{0}, \qquad \boldsymbol{\xi},\boldsymbol{y}\in\Omega,\\ &\hat{\boldsymbol{t}}^j(\boldsymbol{\xi},\boldsymbol{y})=\boldsymbol{0}, \qquad \boldsymbol{\xi}\in\Sigma, \quad \boldsymbol{y}\in\Omega,\\ &\mathcal{R}(\hat{\boldsymbol{u}}^j)=0, \qquad \boldsymbol{\xi},\boldsymbol{y}\in\Omega, \end{aligned} \tag{2.8}$$

where $\boldsymbol{e}^j$ is the unit vector in the jth coordinate direction and $\hat{\boldsymbol{t}}^j$ on Σ is understood in the sense of the trace. In this setting, the Green's functions due to a "surface" point load as in (2.2) are interpreted in the limit as $\Omega\ni\boldsymbol{y}\to\Sigma$. Throughout this investigation, it is assumed that the forward scattering problem for the semi-infinite solid Ω^- given by (2.4), (2.5), and (2.6) admits a unique solution $\tilde{\boldsymbol{u}}^k\in H^1_{loc}(\Omega^-)$.

2.2. Inverse scattering problem. To formulate the reconstruction method, it is next instructive to introduce the *Green's tensor* $\widehat{\boldsymbol{U}}(\boldsymbol{\xi},\boldsymbol{x})$ and the *scattered tensor* $\widetilde{\boldsymbol{U}}(\boldsymbol{\xi},\boldsymbol{x})$, both associated with a unit point source acting at $\boldsymbol{x}\in\Gamma_1$. In the reference Cartesian frame, the components of $\widehat{\boldsymbol{U}}(\boldsymbol{\xi},\boldsymbol{x})$ can be arranged in a 3×3 matrix

$$\widehat{\boldsymbol{U}}(\boldsymbol{\xi},\boldsymbol{x})=\begin{pmatrix}\hat{u}_1^1(\boldsymbol{\xi},\boldsymbol{x}) & \hat{u}_1^2(\boldsymbol{\xi},\boldsymbol{x}) & \hat{u}_1^3(\boldsymbol{\xi},\boldsymbol{x})\\ \hat{u}_2^1(\boldsymbol{\xi},\boldsymbol{x}) & \hat{u}_2^2(\boldsymbol{\xi},\boldsymbol{x}) & \hat{u}_2^3(\boldsymbol{\xi},\boldsymbol{x})\\ \hat{u}_3^1(\boldsymbol{\xi},\boldsymbol{x}) & \hat{u}_3^2(\boldsymbol{\xi},\boldsymbol{x}) & \hat{u}_3^3(\boldsymbol{\xi},\boldsymbol{x})\end{pmatrix}, \quad \boldsymbol{\xi}\in\Omega\setminus\{\boldsymbol{x}\}, \quad \boldsymbol{x}\in\Gamma_1, \tag{2.9}$$

where $\hat{\boldsymbol{u}}^k=(\hat{u}_1^k,\hat{u}_2^k,\hat{u}_3^k)$ is the elastodynamic displacement Green's function for the semi-infinite solid Ω as examined before. Here it is useful to note that $\widehat{\boldsymbol{U}}$ is characterized by the reciprocity property [19], i.e., that $\widehat{\boldsymbol{U}}(\boldsymbol{\xi},\boldsymbol{x})=[\widehat{\boldsymbol{U}}(\boldsymbol{x},\boldsymbol{\xi})]^{\mathrm{T}}$ $(\boldsymbol{x}\neq\boldsymbol{\xi},\boldsymbol{x},\boldsymbol{\xi}\in\Omega)$, where superscript "$_{\mathrm{T}}$" denotes the matrix transpose.

By analogy to (2.9), the perturbation of $\widehat{\boldsymbol{U}}$ due to the presence of an obstacle can be written in the form of the scattered tensor

$$\widetilde{\boldsymbol{U}}(\boldsymbol{\xi},\boldsymbol{x})=\begin{pmatrix}\tilde{u}_1^1(\boldsymbol{\xi},\boldsymbol{x}) & \tilde{u}_1^2(\boldsymbol{\xi},\boldsymbol{x}) & \tilde{u}_1^3(\boldsymbol{\xi},\boldsymbol{x})\\ \tilde{u}_2^1(\boldsymbol{\xi},\boldsymbol{x}) & \tilde{u}_2^2(\boldsymbol{\xi},\boldsymbol{x}) & \tilde{u}_2^3(\boldsymbol{\xi},\boldsymbol{x})\\ \tilde{u}_3^1(\boldsymbol{\xi},\boldsymbol{x}) & \tilde{u}_3^2(\boldsymbol{\xi},\boldsymbol{x}) & \tilde{u}_3^3(\boldsymbol{\xi},\boldsymbol{x})\end{pmatrix}, \quad \boldsymbol{\xi}\in\Omega^-, \quad \boldsymbol{x}\in\Gamma_1, \tag{2.10}$$

where $\tilde{u}_j^k$ is the jth Cartesian component of the scattered field at $\boldsymbol{\xi}\in\Omega^-$ due to a unit point source acting at $\boldsymbol{x}\in\Gamma_1$ in the kth coordinate direction so that $\tilde{\boldsymbol{u}}^k=(\tilde{u}_1^k,\tilde{u}_2^k,\tilde{u}_3^k)$. With reference to (2.3) and (2.9), one may further introduce the *traction Green's tensor*

$$\widehat{\boldsymbol{T}}(\boldsymbol{\xi},\boldsymbol{x})=\boldsymbol{n}(\boldsymbol{\xi})\cdot\mathbf{C}:\boldsymbol{\nabla}_{\boldsymbol{\xi}}\widehat{\boldsymbol{U}}(\boldsymbol{\xi},\boldsymbol{x}), \quad \boldsymbol{\xi}\in S\setminus\{\boldsymbol{x}\}, \quad \boldsymbol{x}\in\Gamma_1, \tag{2.11}$$

for any smooth surface $S\subset\Omega$ with unit normal $\boldsymbol{n}$.

With the above definitions, the inverse problem of interest can be specified as a task of reconstructing an impenetrable obstacle $\Omega_{\rm C}$ from the knowledge of the scattered tensor $\widetilde{\boldsymbol{U}}(\boldsymbol{\xi},\boldsymbol{x})$ for *all* observation points $\boldsymbol{\xi}\in\Gamma_2\subset\Sigma$ and *all* source points $\boldsymbol{x}\in\Gamma_1\subset\Sigma$. In what follows, this problem will be solved by generalizing upon the LSM for near-field elastodynamics proposed in [32], which assumes a continuous representation of $\widetilde{\boldsymbol{U}}(\boldsymbol{\xi},\boldsymbol{x})$ over $\Gamma_1\times\Gamma_2$ as an experimental input. In practice, however, $\widetilde{\boldsymbol{U}}$ is constructed using spatially discrete measurements, which necessitates a sufficient density of the source and observation points. One of the key objectives in this study is to relax the former requirement in terms of the density of excitation sources through a rigorous mathematical reformulation of the existing technique.

3. Preliminaries. Initially developed by Colton and Kirsch [7] in the context of far-field acoustics, the LSM for the *full waveform* (i.e., near-field) obstacle identification in elastodynamics was shown in [32] to revolve around the linear integral equation

$$(\boldsymbol{F}\boldsymbol{g}_{\boldsymbol{z},\boldsymbol{d}})(\boldsymbol{\xi}) = \widehat{\boldsymbol{U}}(\boldsymbol{\xi},\boldsymbol{z})\cdot\boldsymbol{d},\qquad \boldsymbol{\xi}\in\Gamma_2,\quad \boldsymbol{z}\in\Omega,\quad \boldsymbol{d}\in\mathbb{R}^3, \tag{3.1}$$

of the first kind, where the near-field operator $\boldsymbol{F}\colon L_2(\Gamma_1)\to L_2(\Gamma_2)$ is defined as

$$(\boldsymbol{F}\boldsymbol{g}_{\boldsymbol{z},\boldsymbol{d}})(\boldsymbol{\xi}) := \int_{\Gamma_1}\widetilde{\boldsymbol{U}}(\boldsymbol{\xi},\boldsymbol{x})\cdot\boldsymbol{g}_{\boldsymbol{z},\boldsymbol{d}}(\boldsymbol{x})\,dS_{\boldsymbol{x}},\quad \boldsymbol{\xi}\in\Gamma_2; \tag{3.2}$$

$\widetilde{\boldsymbol{U}}$ synthesizes the experimental observations, $\boldsymbol{g}_{\boldsymbol{z},\boldsymbol{d}}(\cdot)\equiv\boldsymbol{g}(\cdot;\boldsymbol{z},\boldsymbol{d})\in L_2(\Gamma_1)$ is the unknown vector density, and $\boldsymbol{d}$ is a unit vector ($\|\boldsymbol{d}\|=1$) signifying the polarization of a "fictitious" point source on the right-hand side of (3.1) acting at *sampling point* $\boldsymbol{z}$. Here $L_2(S)$ denotes the Hilbert space of square-integrable vector fields equipped with the inner product

$$(\boldsymbol{g},\boldsymbol{h})_{L_2(S)} = \int_S \overline{\boldsymbol{g}}(\boldsymbol{x})\cdot\boldsymbol{h}(\boldsymbol{x})\,dS_{\boldsymbol{x}}, \tag{3.3}$$

the overbar implies complex conjugation, and $S\subset\Sigma$ is a generic planar surface of finite extent. In what follows, it is assumed that $\boldsymbol{d}\in\mathbb{R}^3$ and $\|\boldsymbol{d}\|=1$.

For sampling points inside the support of the obstacle, i.e., $\boldsymbol{z}\in\Omega_{\rm C}$, it can be shown under certain restrictions on ω as in [32] that (i) the near-field operator $\boldsymbol{F}$ is injective, (ii) $\boldsymbol{F}$ has a dense range so that (3.1) can be solved at least approximately, and (iii) the solution norm $\|\boldsymbol{g}_{\boldsymbol{z},\boldsymbol{d}}\|_{L_2(\Gamma_1)}$ becomes unbounded as the sampling point $\boldsymbol{z}\in\Omega_{\rm C}$ approaches boundary Γ of the scatterer $\Omega_{\rm C}$ from its interior. Using the concept of topological derivative [20], it is also shown that $\|\boldsymbol{g}_{\boldsymbol{z},\boldsymbol{d}}\|_{L_2(\Gamma_1)}$ can be made arbitrarily large when $\boldsymbol{z}$ lies outside of the support of the scatterer, i.e., $\boldsymbol{z}\in\Omega^-$. This unbounded behavior of $\boldsymbol{g}_{\boldsymbol{z},\boldsymbol{d}}$ has prompted the use of $1/\|\boldsymbol{g}_{\boldsymbol{z},\boldsymbol{d}}\|_{L_2(\Gamma_1)}$, $\boldsymbol{z}\in\Omega$, as a characteristic function of the hidden obstacle $\Omega_{\rm C}$.

Unfortunately, integral representation (3.2) and thus (3.1) do not make much sense if the density of source points on the source surface Γ_1 is insufficient, a situation that is common to many physical testing configurations. To mitigate the problem, it is useful to consider an alternative statement of the LSM wherein the integrals involved are taken over the observation surface Γ_2 rather than the source surface Γ_1.

For the ensuing developments, it is useful to recall Betti's integral identities of linear elasticity (see [26]). To this end, let D be a finite homogeneous elastic body with boundary ∂D of class $C^{1,\alpha}$, and let $\boldsymbol{n}$ denote the unit outward normal on ∂D.

With such a premise Betti's third formula for vector fields $\boldsymbol{u}, \boldsymbol{v} \in C^2(D) \cap C^1(\overline{D})$ can be written as

$$\int_D [\boldsymbol{v}(\boldsymbol{\xi}) \cdot \boldsymbol{L}\boldsymbol{u}(\boldsymbol{\xi}) - \boldsymbol{u}(\boldsymbol{\xi}) \cdot \boldsymbol{L}\boldsymbol{v}(\boldsymbol{\xi})] dV_{\boldsymbol{\xi}} = \int_{\partial D} [\boldsymbol{v}(\boldsymbol{\xi}) \cdot \boldsymbol{t}(\boldsymbol{\xi}; \boldsymbol{u}) - \boldsymbol{u}(\boldsymbol{\xi}) \cdot \boldsymbol{t}(\boldsymbol{\xi}; \boldsymbol{v})] dS_{\boldsymbol{\xi}}, \tag{3.4}$$

where $\boldsymbol{t}(\boldsymbol{\xi}; \boldsymbol{u})$ is given by (2.3) and $\boldsymbol{L}$ is the Lamé operator as examined before.

To formulate the counterpart of (3.1) in terms of an alternative near-field operator that entails integration over the receiver surface, it is essential to show that the scattered tensor in (2.10) is reciprocal. This result is established next.

THEOREM 3.1 (reciprocity). *For the scattering by a cavity, the following symmetry holds:*

$$\widetilde{\boldsymbol{U}}(\boldsymbol{\xi}, \boldsymbol{x}) = [\widetilde{\boldsymbol{U}}(\boldsymbol{x}, \boldsymbol{\xi})]^{\mathrm{T}}, \quad \boldsymbol{x}, \boldsymbol{\xi} \in \Omega^-. \tag{3.5}$$

Proof. Let Ω_{C} be fixed, and let $\tilde{\boldsymbol{u}}^k(\boldsymbol{\zeta}, \boldsymbol{x})$ and $\tilde{\boldsymbol{u}}^j(\boldsymbol{\zeta}, \boldsymbol{\xi})$ be the scattered fields at $\boldsymbol{\zeta} \in \Omega^-$ due to point forces acting respectively at $\boldsymbol{x} \in \Omega^-$ in the kth coordinate direction and $\boldsymbol{\xi} \in \Omega^-$ in the jth coordinate direction. Next, select R so that $\boldsymbol{\zeta}, \boldsymbol{x}, \boldsymbol{\xi} \in \Omega_{\mathrm{R}}^-$ as well. On the basis of Betti's third formula (3.4) applied to Ω_{R}^-, homogeneous Navier equations in terms of $\tilde{\boldsymbol{u}}^k$ and $\tilde{\boldsymbol{u}}^j$ over Ω^-, the homogeneous Neumann condition in (2.5), and the radiation condition (2.7), it can be shown in the limit as $R \to \infty$ that

$$\int_\Gamma [\tilde{\boldsymbol{u}}^j(\boldsymbol{\zeta}, \boldsymbol{\xi}) \cdot \tilde{\boldsymbol{t}}^k(\boldsymbol{\zeta}, \boldsymbol{x}) - \tilde{\boldsymbol{u}}^k(\boldsymbol{\zeta}, \boldsymbol{x}) \cdot \tilde{\boldsymbol{t}}^j(\boldsymbol{\zeta}, \boldsymbol{\xi})] dS_{\boldsymbol{\zeta}} = 0, \quad \boldsymbol{x}, \boldsymbol{\xi} \in \Omega^-. \tag{3.6}$$

Similarly, application of Betti's third formula and the use of the homogeneous Navier equations in terms of Green's functions $\hat{\boldsymbol{u}}^k(\boldsymbol{\zeta}, \boldsymbol{x})$ and $\hat{\boldsymbol{u}}^j(\boldsymbol{\zeta}, \boldsymbol{\xi})$ over the interior domain Ω_{C} yield the identity

$$\int_\Gamma [\hat{\boldsymbol{u}}^j(\boldsymbol{\zeta}, \boldsymbol{\xi}) \cdot \hat{\boldsymbol{t}}^k(\boldsymbol{\zeta}, \boldsymbol{x}) - \hat{\boldsymbol{u}}^k(\boldsymbol{\zeta}, \boldsymbol{x}) \cdot \hat{\boldsymbol{t}}^j(\boldsymbol{\zeta}, \boldsymbol{\xi})] dS_{\boldsymbol{\zeta}} = 0, \quad \boldsymbol{x}, \boldsymbol{\xi} \in \Omega^-, \tag{3.7}$$

where, owing to the vanishing right-hand side, the boundary normal $\boldsymbol{n}$ (implicit to $\hat{\boldsymbol{t}}^j$ and $\hat{\boldsymbol{t}}^k$) can be taken as oriented toward the interior of Ω_{C} for consistency with (3.6). For $\boldsymbol{x}, \boldsymbol{\xi} \in \Omega^-$, one can next write the boundary integral representations

$$\begin{aligned} \tilde{u}_j^k(\boldsymbol{\xi}, \boldsymbol{x}) &= \int_\Gamma [\hat{\boldsymbol{u}}^j(\boldsymbol{\zeta}, \boldsymbol{\xi}) \cdot \tilde{\boldsymbol{t}}^k(\boldsymbol{\zeta}, \boldsymbol{x}) - \hat{\boldsymbol{t}}^j(\boldsymbol{\zeta}, \boldsymbol{\xi}) \cdot \tilde{\boldsymbol{u}}^k(\boldsymbol{\zeta}, \boldsymbol{x})] dS_{\boldsymbol{\zeta}}, \\ \tilde{u}_k^j(\boldsymbol{x}, \boldsymbol{\xi}) &= \int_\Gamma [\hat{\boldsymbol{u}}^k(\boldsymbol{\zeta}, \boldsymbol{x}) \cdot \tilde{\boldsymbol{t}}^j(\boldsymbol{\zeta}, \boldsymbol{\xi}) - \hat{\boldsymbol{t}}^k(\boldsymbol{\zeta}, \boldsymbol{x}) \cdot \tilde{\boldsymbol{u}}^j(\boldsymbol{\zeta}, \boldsymbol{\xi})] dS_{\boldsymbol{\zeta}}, \quad \boldsymbol{x}, \boldsymbol{\xi} \in \Omega^-, \end{aligned} \tag{3.8}$$

of the scattered field; see, e.g., [2]. On subtracting (3.8b) from the sum of (3.6), (3.7), and (3.8a), one finds that

$$\tilde{u}_j^k(\boldsymbol{\xi}, \boldsymbol{x}) - \tilde{u}_k^j(\boldsymbol{x}, \boldsymbol{\xi}) = \int_\Gamma [\boldsymbol{u}^j(\boldsymbol{\zeta}, \boldsymbol{\xi}) \cdot \boldsymbol{t}^k(\boldsymbol{\zeta}, \boldsymbol{x}) - \boldsymbol{u}^k(\boldsymbol{\zeta}, \boldsymbol{x}) \cdot \boldsymbol{t}^j(\boldsymbol{\zeta}, \boldsymbol{\xi})] dS_{\boldsymbol{\zeta}}, \quad \boldsymbol{x}, \boldsymbol{\xi} \in \Omega^-, \tag{3.9}$$

where $\boldsymbol{u}^j(\boldsymbol{\zeta}, \boldsymbol{\xi}) = \hat{\boldsymbol{u}}^j + \tilde{\boldsymbol{u}}^j$ and $\boldsymbol{t}^j(\boldsymbol{\zeta}, \boldsymbol{\xi}) = \hat{\boldsymbol{t}}^j + \tilde{\boldsymbol{t}}^j$ denote respectively the *total* displacement and traction vectors at $\boldsymbol{\zeta} \in \Gamma$ due to a point source acting at $\boldsymbol{\xi} \in \Omega^-$ in the jth coordinate direction. By virtue of the fact that $\boldsymbol{t}^j(\boldsymbol{\zeta}, \cdot) = \boldsymbol{t}^k(\boldsymbol{\zeta}, \cdot) \equiv \boldsymbol{0}$ for $\boldsymbol{\zeta} \in \Gamma$ according to (2.5), the right-hand side of (3.9) vanishes identically, which, through (2.10), completes the proof. ∎

One of the key steps in establishing the rationale for (3.1) is the proof that (3.2) represents a scattered field in Ω^-. The following theorem and its lemma aim to establish an analogous result for the sought "source-friendly" variant of (3.1).

THEOREM 3.2. *Let $\Gamma_2 \subset \Sigma$ be a surface of finite extent, and let $\boldsymbol{h} \in L_2(\Gamma_2)$. Then the single-layer potential*

$$\boldsymbol{v}(\boldsymbol{x}) = \int_{\Gamma_2} [\widehat{\boldsymbol{U}}(\boldsymbol{\xi}, \boldsymbol{x})]^{\mathrm{T}} \cdot \boldsymbol{h}(\boldsymbol{\xi})\, dS_{\boldsymbol{\xi}} = \int_{\Gamma_2} \hat{\boldsymbol{u}}^k(\boldsymbol{x}, \boldsymbol{\xi})\, h_k(\boldsymbol{\xi})\, dS_{\boldsymbol{\xi}}, \quad \boldsymbol{x} \in \Omega, \tag{3.10}$$

is a radiating solution of the homogeneous Navier equation in Ω so that

$$\begin{aligned} \boldsymbol{L}\boldsymbol{v}(\boldsymbol{x}) + \rho\omega^2 \boldsymbol{v}(\boldsymbol{x}) &= \boldsymbol{0}, \quad \boldsymbol{x} \in \Omega, \\ \boldsymbol{t}(\boldsymbol{x}; \boldsymbol{v}) = \boldsymbol{0}, \quad \boldsymbol{x} \in \Sigma \backslash \Gamma_2, &\qquad \mathcal{R}(\boldsymbol{v}) = 0, \quad \boldsymbol{x} \in \Omega, \end{aligned} \tag{3.11}$$

where $\boldsymbol{t}(\boldsymbol{x}; \boldsymbol{v}) = \boldsymbol{n}(\boldsymbol{x}) \cdot \mathbf{C} : \boldsymbol{\nabla} \boldsymbol{v}(\boldsymbol{x})$ is the traction vector associated with $\boldsymbol{v}$, understood in the sense of the trace.

Proof. For $\boldsymbol{x} \in \Omega$ and $\boldsymbol{\xi} \in \Gamma_2$, $\boldsymbol{u}^k(\boldsymbol{x}, \boldsymbol{\xi})$ are regular since $\Omega \cap \Gamma_2 = \emptyset$, and (3.10) accordingly permits differentiation under the integral sign. With such a result, (3.11) follows directly from the fact that $\hat{\boldsymbol{u}}^k$ $(k = 1, 2, 3)$ satisfies the homogeneous Navier equation away from the source surface Γ_2. In a similar fashion, the homogeneous Neumann condition in (3.11) can be obtained by means of (3.10) and the limit of (2.8b) when the source point $\boldsymbol{y} \to \Gamma_2$. On the basis of (2.3) and (3.10), on the other hand, one finds that for any $\boldsymbol{y} \in \Omega$

$$\begin{aligned} &\int_{\Gamma_{\mathrm{R}}} \left\{ \hat{\boldsymbol{u}}^j(\boldsymbol{x}, \boldsymbol{y}) \cdot \boldsymbol{t}(\boldsymbol{x}; \boldsymbol{v}) - \hat{\boldsymbol{t}}^j(\boldsymbol{x}, \boldsymbol{y}) \cdot \boldsymbol{v}(\boldsymbol{x}) \right\} dS_{\boldsymbol{x}} \\ &= \int_{\Gamma_2} h_k(\boldsymbol{\eta}) \int_{\Gamma_{\mathrm{R}}} \left\{ \hat{\boldsymbol{u}}^j(\boldsymbol{x}, \boldsymbol{y}) \cdot \hat{\boldsymbol{t}}^k(\boldsymbol{x}, \boldsymbol{\eta}) - \hat{\boldsymbol{t}}^j(\boldsymbol{x}, \boldsymbol{y}) \cdot \hat{\boldsymbol{u}}^k(\boldsymbol{x}, \boldsymbol{\eta}) \right\} \mathrm{d}S_{\boldsymbol{x}}\, dS_{\boldsymbol{\eta}} \end{aligned} \tag{3.12}$$

over a hemispherical surface Γ_{R}, where R is taken sufficiently large so that $\Gamma_2 \subset \partial\Omega_{\mathrm{R}}$ and $\boldsymbol{y} \in \Omega_{\mathrm{R}}$. By virtue of (3.12), statement $\mathcal{R}(\boldsymbol{v}) = 0$ in (3.11) immediately follows from the fact that the displacement Green's function $\hat{\boldsymbol{u}}^k(\cdot, \boldsymbol{z})$ is a radiating elastodynamic solution in $\Omega \backslash \{\boldsymbol{z}\}$; see [19]. ∎

LEMMA 3.3. *For a given vector density $\boldsymbol{h} \in L_2(\Gamma_2)$, the radiating solution to the scattering problem for a cavity Ω_C in the semi-infinite reference solid Ω illuminated by the free field*

$$\mathring{\boldsymbol{v}}(\boldsymbol{x}) = (\boldsymbol{E}\boldsymbol{h})(\boldsymbol{x}) \equiv \int_{\Gamma_2} [\widehat{\boldsymbol{U}}(\boldsymbol{\xi}, \boldsymbol{x})]^{\mathrm{T}} \cdot \boldsymbol{h}(\boldsymbol{\xi})\, dS_{\boldsymbol{\xi}}, \quad \boldsymbol{x} \in \Omega, \tag{3.13}$$

is given by the scattered field

$$\tilde{\boldsymbol{v}}(\boldsymbol{x}) = \int_{\Gamma_2} [\widetilde{\boldsymbol{U}}(\boldsymbol{\xi}, \boldsymbol{x})]^{\mathrm{T}} \cdot \boldsymbol{h}(\boldsymbol{\xi})\, dS_{\boldsymbol{\xi}}, \quad \boldsymbol{x} \in \Omega^-, \tag{3.14}$$

where $\widehat{\boldsymbol{U}}$ and $\widetilde{\boldsymbol{U}}$ are given respectively by (2.9) *and* (2.10).

Proof. With the aid of (2.10) and the reciprocity property (3.5) established in Theorem 3.1, formula (3.14) can be rewritten as

$$\tilde{v}_j(\boldsymbol{x}) = \int_{\Gamma_2} \tilde{u}_j^k(\boldsymbol{x}, \boldsymbol{\xi})\, h_k(\boldsymbol{\xi})\, dS_{\boldsymbol{\xi}}, \qquad \boldsymbol{x} \in \Omega^-, \quad j = 1, 2, 3. \tag{3.15}$$

By virtue of (2.5), integral representation of the scattered field $\tilde{\boldsymbol{u}}^k$ in Ω^- in terms of the total field $\boldsymbol{u}^k$ over the cavity boundary Γ can be written as

$$\tilde{u}_j^k(\boldsymbol{x},\boldsymbol{\xi}) = -\int_\Gamma \hat{\boldsymbol{t}}^j(\boldsymbol{\eta},\boldsymbol{x})\cdot\boldsymbol{u}^k(\boldsymbol{\eta},\boldsymbol{\xi})\,dS_{\boldsymbol{\eta}}, \quad \boldsymbol{x}\in\Omega^-, \quad \boldsymbol{\xi}\in\Gamma_2; \tag{3.16}$$

see, e.g., [21]. By use of (3.16) in (3.15) and interchanging the order of integration, one finds

$$\tilde{v}_j(\boldsymbol{x}) = -\int_\Gamma \hat{\boldsymbol{t}}^j(\boldsymbol{\eta},\boldsymbol{x})\cdot\boldsymbol{v}(\boldsymbol{\eta})\,dS_{\boldsymbol{\eta}}, \quad \boldsymbol{x}\in\Omega^-, \tag{3.17}$$

where

$$\boldsymbol{v}(\boldsymbol{\eta}) = \int_{\Gamma_2} \boldsymbol{u}^k(\boldsymbol{\eta},\boldsymbol{\xi})\,h_k(\boldsymbol{\xi})\,dS_{\boldsymbol{\xi}}, \quad \boldsymbol{\eta}\in\Gamma. \tag{3.18}$$

From (3.17), it is seen that $\tilde{\boldsymbol{v}}(\boldsymbol{x})$ admits a representation in terms of a double-layer potential similar to that in (3.16). Since $\boldsymbol{x}\in\Omega^-$, it can be shown [27] using the radiating property of $\hat{\boldsymbol{u}}^k$ that the right-hand side of (3.17) is itself a *radiating* solution of the homogeneous Navier equation in Ω^- so that

$$\begin{aligned} \boldsymbol{L}\tilde{\boldsymbol{v}}(\boldsymbol{x}) + \rho\omega^2\tilde{\boldsymbol{v}}(\boldsymbol{x}) &= \boldsymbol{0}, \qquad \boldsymbol{x}\in\Omega^-, \\ \mathcal{R}(\tilde{\boldsymbol{v}}) &= 0, \qquad \boldsymbol{x}\in\Omega^-. \end{aligned} \tag{3.19}$$

On applying (2.3) to (3.13) and (3.15) and interchanging the order of integral and differential operators, one finds from (2.5) that

$$\boldsymbol{t}(\boldsymbol{y};\tilde{\boldsymbol{v}}) \;=\; \int_{\Gamma_2} \tilde{\boldsymbol{t}}^k(\boldsymbol{y},\boldsymbol{\xi})\,h_k(\boldsymbol{\xi})\,dS_{\boldsymbol{\xi}} \;=\; \begin{cases} \boldsymbol{0}, & \boldsymbol{y}\in\Sigma, \\ -\boldsymbol{t}(\boldsymbol{y};\mathring{\boldsymbol{v}}), & \boldsymbol{y}\in\Gamma, \end{cases} \tag{3.20}$$

in the limits as $\boldsymbol{x}\to\boldsymbol{y}\in\Sigma$ and $\boldsymbol{x}\to\boldsymbol{y}\in\Gamma$, respectively. Given the fact that $\Gamma=\partial\Omega_{\rm C}$ and $\Omega^-=\Omega\setminus(\Omega_{\rm C}\cup\Gamma)$, equations (3.19) and (3.20) indeed demonstrate that $\tilde{\boldsymbol{v}}$ is a radiating solution to the scattering problem for a cavity $\Omega_{\rm C}$ due to free field (3.13). ∎

4. Adjoint formulation of the LSM. With the foregoing developments, linear sampling equation (3.1) aiding the full waveform tomography of semi-infinite elastic solids can now be reformulated so that the featured integration is performed over the receiver surface Γ_2 in lieu of Γ_1. To this end, let the sampling point $\boldsymbol{z}\in\Omega$ be fixed. The idea is to establish an alternative near-field operator $\boldsymbol{G}\colon L_2(\Gamma_2)\to L_2(\Gamma_1)$ that synthesizes experimental observations in terms of the scattered tensor $\widetilde{\boldsymbol{U}}$, and to seek the vector density $\boldsymbol{h}_{\boldsymbol{z},\boldsymbol{d}}(\cdot)\equiv\boldsymbol{h}(\cdot;\boldsymbol{z},\boldsymbol{d})\in L_2(\Gamma_2)$ as a solution to the integral equation of the first kind

$$(\boldsymbol{G}\boldsymbol{h}_{\boldsymbol{z},\boldsymbol{d}})(\boldsymbol{x}) \;=\; \widehat{\boldsymbol{U}}(\boldsymbol{x},\boldsymbol{z})\cdot\boldsymbol{d}, \qquad \boldsymbol{x}\in\Gamma_1, \quad \boldsymbol{z}\in\Omega, \quad \boldsymbol{d}\in\mathbb{R}^3, \tag{4.1}$$

where $\boldsymbol{d}$ is a fixed polarization vector ($\|\boldsymbol{d}\|=1$) as examined earlier. Adopting fundamental hypotheses of the LSM, $\boldsymbol{G}$ must be designed so that, under certain restrictions on the excitation frequency ω, the following statements apply:

- For $\boldsymbol{z}\in\Omega_{\rm C}$, operator $\boldsymbol{G}$ is injective and has a dense range. The solution of (4.1) further has the property $\lim_{\boldsymbol{z}\to\boldsymbol{y}\in\Gamma}\|\boldsymbol{h}_{\boldsymbol{z},\boldsymbol{d}}\|_{L_2(\Gamma_2)}=\infty$.

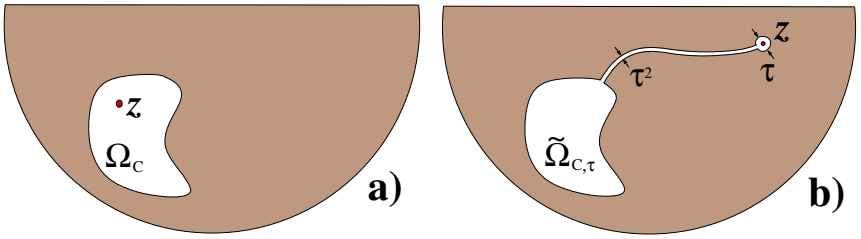


FIG. 4.1. *Sampling cases:* (a) $\boldsymbol{z}\in\Omega_C$, *"true" obstacle;* (b) $\boldsymbol{z}\in\Omega\setminus\overline{\Omega}_C$, *perturbed obstacle.*

- For $\boldsymbol{z}\in\Omega\setminus(\Omega_{\rm C}\cup\Gamma)$, there exists an approximate solution $\boldsymbol{h}^{\tau}_{\boldsymbol{z},\boldsymbol{d}}$ to (4.1) such that $\lim_{\tau\to 0}\|\boldsymbol{h}^{\tau}_{\boldsymbol{z},\boldsymbol{d}}\|_{L_2(\Gamma_2)}=\infty$, where τ is the approximation parameter. With reference to Figure 4.1, approximate solution $\boldsymbol{h}^{\tau}_{\boldsymbol{z},\boldsymbol{d}}$ is understood in the sense of a perturbed scatterer domain $\widetilde{\Omega}_{{\rm C},\tau}\ni z$ [32].

With such prerequisites, the *unboundedness* property of the sought vector density $\boldsymbol{h}_{\boldsymbol{z},\boldsymbol{d}}$ can then be used to reconstruct a hidden cavity $\Omega_{\rm C}$ by probing the subsurface volume of interest through an array of sampling points $\boldsymbol{z}$, and identifying $\Omega_{\rm C}$ via an assembly of points where $\|\boldsymbol{h}_{\boldsymbol{z},\boldsymbol{d}}\|_{L_2(\Gamma_2)}$ is bounded. As elucidated earlier, such an identification procedure would make sense even when the density of source points on Γ_1, used to illuminate the cavity, is limited.

To facilitate the ensuing developments, it is useful to make reference to the near-field operator $\boldsymbol{F}$ in (3.2) and introduce its adjoint counterpart $\boldsymbol{F}^*: L_2(\Gamma_2)\to L_2(\Gamma_1)$ by the ensuing proposition.

LEMMA 4.1. *For all* $\boldsymbol{g}\in L_2(\Gamma_1)$ *and* $\boldsymbol{e}\in L_2(\Gamma_2)$,

$$(\boldsymbol{F}\boldsymbol{g},\boldsymbol{e})_{L_2(\Gamma_2)} \;=\; (\boldsymbol{g},\boldsymbol{F}^*\boldsymbol{e})_{L_2(\Gamma_1)}, \tag{4.2}$$

where $\boldsymbol{F}$ *is defined by* (3.2) *and*

$$(\boldsymbol{F}^*\boldsymbol{e})(\boldsymbol{x}) := \int_{\Gamma_2}[\overline{\widetilde{\boldsymbol{U}}(\boldsymbol{\xi},\boldsymbol{x})}]^{\rm T}\cdot\boldsymbol{e}(\boldsymbol{\xi})\,dS_{\boldsymbol{\xi}},\quad \boldsymbol{x}\in\Gamma_1, \tag{4.3}$$

where an overbar symbol denotes complex conjugation.

Proof. The statement of the lemma in terms of (4.2) and (4.3) can be established using (3.2) and (3.3) and interchanging the order of integration. ∎

To arrive at a form of $\boldsymbol{G}$ that yields the required solvability and unboundedness properties in terms of $\boldsymbol{h}_{\boldsymbol{z},\boldsymbol{d}}$, one is tempted to employ the result of Lemma 4.1 and postulate the integral equation

$$\int_{\Gamma_2}[\overline{\widetilde{\boldsymbol{U}}(\boldsymbol{\xi},\boldsymbol{x})}]^{\rm T}\cdot\boldsymbol{e}(\boldsymbol{\xi})\,dS_{\boldsymbol{\xi}} \;=\; [\overline{\widehat{\boldsymbol{U}}(\boldsymbol{z},\boldsymbol{x})}]^{\rm T}\cdot\boldsymbol{d},\qquad \boldsymbol{x}\in\Gamma_1,\quad \boldsymbol{z}\in\Omega, \tag{4.4}$$

as a basis for the "source-friendly"alternative to (3.1). On employing the symmetry of the elastodynamic displacement Green's tensor (2.9) and letting $\boldsymbol{h}=\overline{\boldsymbol{e}}$, integral equation (4.4) can be conveniently rewritten as (4.1), where

$$(\boldsymbol{G}\boldsymbol{h}_{\boldsymbol{z},\boldsymbol{d}})(\boldsymbol{x}) := \int_{\Gamma_2}[\widetilde{\boldsymbol{U}}(\boldsymbol{\xi},\boldsymbol{x})]^{\rm T}\cdot\boldsymbol{h}_{\boldsymbol{z},\boldsymbol{d}}(\boldsymbol{\xi})\,dS_{\boldsymbol{\xi}},\quad \boldsymbol{x}\in\Gamma_1. \tag{4.5}$$

With reference to Lemma 3.3, the conjugation of (4.4) represents a key step in establishing (4.1) that features the near-field operator (4.5) as a *radiating* elastodynamic

field in the sense of (2.6) and thus enables a direct use of the results obtained in section 3. It is also useful to note that, for $\widetilde{\boldsymbol{U}} \in L_2(\Gamma_1 \times \Gamma_2)$, the near-field operator $\boldsymbol{G}$ is well defined, linear, and bounded from $L_2(\Gamma_2)$ into $L_2(\Gamma_1)$. The latter property can be demonstrated via the Cauchy-Schwarz inequality

$$\|\boldsymbol{G}\boldsymbol{h}\|^2_{L_2(\Gamma_1)} \leq \|\boldsymbol{h}\|^2_{L_2(\Gamma_2)} \left(\sum_{k=1}^{3} \sum_{j=1}^{3} \int_{\Gamma_1} \int_{\Gamma_2} \left| \tilde{u}^k_j(\boldsymbol{\xi}, \boldsymbol{x}) \right|^2 dS_{\boldsymbol{x}}\, dS_{\boldsymbol{\xi}} \right), \tag{4.6}$$

where $|\cdot|$ denotes the complex modulus. It can also be shown (see, e.g., [25]) that the linear integral operator $\boldsymbol{G}$ is *compact* from $L_2(\Gamma_2)$ into $L_2(\Gamma_1)$, thus rendering the linear equation (4.1) ill-posed.

4.1. Mathematical justification of the adjoint method. To validate the proposed developments, it is next necessary to establish the injectivity, denseness, and unboundedness theorems characterizing the solution of (4.1). For brevity, attention is herein focused on the "default" case when $\boldsymbol{z} \in \Omega_{\mathrm{C}}$. Following the approach taken in [32], situations with $\boldsymbol{z} \notin \Omega_{\mathrm{C}}$ can be effectively treated by considering the perturbed scatterer $\widetilde{\Omega}_{\mathrm{C},\tau}$ with a vanishing appendage (see Figure 4.1(b)) so that $\boldsymbol{z} \in \widetilde{\Omega}_{\mathrm{C},\tau}$ and investigating the behavior of such perturbed (4.1) as $\tau \to 0$.

THEOREM 4.2 (solvability). *Let Ω_C be a cavity. Then the equation*

$$(\boldsymbol{G}\boldsymbol{h}_{\boldsymbol{z},\boldsymbol{d}})(\boldsymbol{x}) \;=\; \widehat{\boldsymbol{U}}(\boldsymbol{x}, \boldsymbol{z}) \cdot \boldsymbol{d}, \qquad \boldsymbol{x} \in \Gamma_1, \quad \boldsymbol{z} \in \Omega_C, \tag{4.7}$$

where $\boldsymbol{G}$ is given by (4.5)*, possesses a solution $\boldsymbol{h}_{\boldsymbol{z},\boldsymbol{d}} \in L_2(\Gamma_2)$ if and only if there exists an elastodynamic solution $\mathring{\boldsymbol{v}}$ to the interior Neumann problem*

$$\begin{aligned} \boldsymbol{L}\mathring{\boldsymbol{v}}(\boldsymbol{x}) + \rho\,\omega^2 \mathring{\boldsymbol{v}}(\boldsymbol{x}) &= \boldsymbol{0}, \quad \boldsymbol{x} \in \Omega_C, \\ \boldsymbol{t}(\boldsymbol{x}; \mathring{\boldsymbol{v}}) + \widehat{\boldsymbol{T}}(\boldsymbol{x}, \boldsymbol{z}) \cdot \boldsymbol{d} &= \boldsymbol{0}, \quad \boldsymbol{x} \in \Gamma, \end{aligned} \tag{4.8}$$

that permits a representation in the form of the single-layer potential (3.13).

Proof. Let $\boldsymbol{h}_{\boldsymbol{z},\boldsymbol{d}} \in L_2(\Gamma_2)$ be a solution to the integral equation (4.7). From the results of Theorem 3.2 and Lemma 3.3, it follows directly that the free field $\mathring{\boldsymbol{v}} = \boldsymbol{E}\boldsymbol{h}$ and the induced scattered field $\tilde{\boldsymbol{v}} = \boldsymbol{G}\boldsymbol{h}$ are both radiating solutions in Ω^-. In addition, the single-layer potential (3.13) satisfies the homogeneous Navier equation in Ω_{C}, owing to the fact that $\Gamma_2 \cap \Omega_{\mathrm{C}} = \emptyset$. Accordingly, (4.8) is obtained from (4.7), the Holmgren's uniqueness theorem [17], and the Neumann boundary conditions (2.5) on Γ. Conversely let $\mathring{\boldsymbol{v}} = \boldsymbol{E}\boldsymbol{h}$, which solves the interior Neumann problem (4.8), be taken as a free field for the scattering by a cavity Ω_{C}. By virtue of Lemma 3.3 and the fact that $\boldsymbol{z} \in \Omega_{\mathrm{C}}$, both the (induced) scattered field $\tilde{\boldsymbol{v}}$ and the Green's function $\widehat{\boldsymbol{U}}(\cdot, \boldsymbol{z}) \cdot \boldsymbol{d}$ are radiating solutions in Ω^- satisfying the homogeneous Navier equation. Owing to (2.5), (4.8b), and the uniqueness of the solution to the scattering problem (2.4)–(2.6) (see [27]), the induced scattered field $(\boldsymbol{G}\boldsymbol{h})(\boldsymbol{x}) = \widehat{\boldsymbol{U}}(\boldsymbol{x}) \cdot \boldsymbol{d}$ in Ω^- and (4.7) follows in the limit as $\boldsymbol{x} \to \boldsymbol{y} \in \Gamma_1$. ∎

Unfortunately, the solution of the interior Neumann problem (4.8) may not permit representation $\mathring{\boldsymbol{v}} = \boldsymbol{E}\boldsymbol{h}$ with $\boldsymbol{h} \in L_2(\Gamma_2)$ in many situations. To examine the approximating characteristics of (3.13), assume that $D \subset \Omega$ is a bounded domain with boundary ∂D of class $C^{1,\alpha}$, and let $H^1(D)$ be a Sobolev space of vector fields equipped with the inner product

$$(\boldsymbol{v}, \boldsymbol{u})_{H^1(D)} \;=\; \theta \int_D \overline{\boldsymbol{v}}(\boldsymbol{\xi}) \cdot \boldsymbol{u}(\boldsymbol{\xi})\, dV_{\boldsymbol{\xi}} + \int_D \boldsymbol{\nabla}\overline{\boldsymbol{v}}(\boldsymbol{\xi}) \mathbin{:} \mathbf{C} \mathbin{:} \boldsymbol{\nabla}\boldsymbol{u}(\boldsymbol{\xi})\, dV_{\boldsymbol{\xi}}, \tag{4.9}$$

where $\mathbb{R} \ni \theta > 0$. With such definitions, one may introduce $\mathbb{H}(D)$ as a set of classical solutions to the homogeneous Navier equation

$$\mathbb{H}(D) = \big\{ \boldsymbol{u} \in C^2(D) \cap C^1(\overline{D}) : \boldsymbol{L}\boldsymbol{u} + \rho\omega^2 \boldsymbol{u} = \boldsymbol{0} \ \text{ in } D \big\},$$

whose closure, $\overline{\mathbb{H}(D)}$, is defined with respect to the norm $\|\boldsymbol{u}\|_{H^1(D)} = \sqrt{(\boldsymbol{u},\boldsymbol{u})_{H^1(D)}}$. Next, consider the single-layer integral operator $\boldsymbol{S} : L_2(\Gamma_2) \to \overline{\mathbb{H}(D)}$ given by

$$(\boldsymbol{S}\boldsymbol{h})(\boldsymbol{\xi}) := \int_{\Gamma_2} [\widehat{\boldsymbol{U}}(\boldsymbol{x},\boldsymbol{\xi})]^{\mathrm{T}} \cdot \boldsymbol{h}(\boldsymbol{x})\, dS_{\boldsymbol{x}}, \quad \boldsymbol{\xi} \in D. \tag{4.10}$$

For $\Gamma_2 \cap \overline{D} = \emptyset$, assumed in this study, $\boldsymbol{S}\boldsymbol{h} \in C^\infty\left(\overline{D}\right) \subset \{C^2(D) \cap C^1(\overline{D})\}$. By virtue of this result and the fact that $\boldsymbol{S}\boldsymbol{h}$ satisfies the homogeneous Navier equation in D according to Theorem 3.2, it immediately follows that $\boldsymbol{S}\boldsymbol{h} \in \mathbb{H}(D)$.

THEOREM 4.3 (range denseness). *The space of single-layer potentials* $\{\boldsymbol{S}\boldsymbol{h}, \boldsymbol{h} \in L_2(\Gamma_2)\}$ *given by* (4.10) *is dense in the space of classical solutions to the homogeneous Navier equation:* $\boldsymbol{L}\boldsymbol{u} + \rho\omega^2\boldsymbol{u} = \boldsymbol{0}$ *in* D *with respect to the* $H^1(D)$ *norm.*

Proof. By establishing the elements of the proof as in [32], it can be shown that $(\boldsymbol{S}\boldsymbol{h}, \boldsymbol{u})_{H^1(D)} = 0$ for all $\boldsymbol{h} \in L_2(\Gamma_2)$ requires $\boldsymbol{u} \equiv \boldsymbol{0}$ in D. ∎

As examined earlier, the adjoint variant of the linear sampling method revolves around the equation of the first kind (4.7), whose reciprocal solution norm can be used as a characteristic function of the scatterer. The key hypotheses in this approach, however, are that (i) (4.7) can be solved uniquely when $\widehat{\boldsymbol{U}} \cdot \boldsymbol{d} \in \text{Range}(\boldsymbol{G})$, (ii) (4.7) can be solved approximately (with arbitrary accuracy) when $\widehat{\boldsymbol{U}} \cdot \boldsymbol{d} \notin \text{Range}(\boldsymbol{G})$, and (iii) the solution $\boldsymbol{h}_{\boldsymbol{z},\boldsymbol{d}}$ behaves such that $\lim_{\boldsymbol{z} \to \boldsymbol{y} \in \Gamma} \|\boldsymbol{h}_{\boldsymbol{z},\boldsymbol{d}}\|_{L_2(\Gamma_2)} = \infty$. These requirements are established next.

THEOREM 4.4 (injectivity, approximation, and solution unboundedness). *Assume that* (i) $\boldsymbol{z} \in \Omega_C$ *is fixed and* $\boldsymbol{d} \in \mathbb{R}^3$ *with* $\|\boldsymbol{d}\| = 1$, (ii) Γ *is of class* $C^{1,\alpha}$, *and* (iii) $\rho\omega^2$ *is not a Neumann eigenvalue of* $-\boldsymbol{L}$ *in* Ω_C *with eigenfunction* $\boldsymbol{S}\boldsymbol{h}$ *given by* (4.10). *Then* $\boldsymbol{G}$ *is one-to-one, and for every* $\varepsilon > 0$ *there exists* $\boldsymbol{h}^\varepsilon(\cdot;\boldsymbol{z},\boldsymbol{d}) \in L_2(\Gamma_2)$ *such that*

$$\left\| \boldsymbol{G}\boldsymbol{h}^\varepsilon(\cdot;\boldsymbol{z},\boldsymbol{d}) - \widehat{\boldsymbol{U}}(\cdot,\boldsymbol{z}) \cdot \boldsymbol{d} \right\|_{L_2(\Gamma_1)} < \varepsilon. \tag{4.11}$$

For every fixed $\varepsilon > 0$ *and* $\boldsymbol{h}^\varepsilon$ *satisfying* (4.11), *one further has*

$$\lim_{\boldsymbol{z} \to \boldsymbol{y} \in \Gamma} \|\boldsymbol{h}^\varepsilon(\cdot;\boldsymbol{z},\boldsymbol{d})\|_{L_2(\Gamma_2)} = \infty. \tag{4.12}$$

Proof. A detailed proof of (4.11) and (4.12), which builds on the results from Theorems 4.2 and 4.3, is similar to that in [32] established for the treatment of (3.1) and is omitted here for brevity. ∎

5. Computational treatment and regularization. On the basis of the foregoing developments, elastic-wave reconstruction of impenetrable obstacles in a semi-infinite solid can be achieved by solving the integral equation (4.1) with the near-field operator $\boldsymbol{G}$ given by (4.5), a format that may be especially useful in situations involving a limited density of "illuminating" point sources distributed over $\Gamma_1 \subset \Sigma$ (see Figure 2.1). In this approach, the reference semi-infinite solid is sequentially probed in a pointwise fashion by placing a fictitious point source (acting in direction $\boldsymbol{d}$) over an array of sampling points $\boldsymbol{z} \in \mathcal{D} \subset \Omega$, where $\mathcal{D}$ is the subsurface region of interest. With

reference to (4.1), the unknown scatterer Ω_{C} can thus be reconstructed by solving the linear operator equation

$$\boldsymbol{G}\boldsymbol{h} = \boldsymbol{b}, \tag{5.1}$$

where $\boldsymbol{G}$ is given by (4.5), $\boldsymbol{h} = \boldsymbol{h}(\cdot; \boldsymbol{z}, \boldsymbol{d})$, and $\boldsymbol{b} = \widehat{\boldsymbol{U}}(\cdot, \boldsymbol{z}) \cdot \boldsymbol{d}$. As elucidated earlier, Fredholm integral equation of the first kind (5.1) constitutes an ill-posed mathematical problem in the sense of Hadamard [14, 24]. On citing the solvability of (5.1) as examined in section 4.1, a careful numerical treatment must be adopted next to obtain a stable solution in terms of $\boldsymbol{h}$.

5.1. Discretization. In practice the input data, herein synthesized in the form of the scattered tensor $\widetilde{\boldsymbol{U}}$, are monitored over a discrete set of control points located on the measurement surface Γ_2. Likewise, the time-harmonic excitation used to illuminate the obstacle is often provided by a finite number of "point" sources acting sequentially on the source surface Γ_1. To illustrate the physical relevance of the assumed illuminating field, it is worth noting that in quantitative ultrasound imaging, a point-like excitation of soft tissues can be achieved by way of the so-called acoustic radiation force [13, 16].

To consistently deal with such a discrete experimental input, let $\{E_k\}_{k=1}^{K}$ be a system of closed and nonoverlapping subsets of the receiver surface Γ_2 such that $\Gamma_2 = \bigcup_{k=1}^{K} E_k$. On assuming that each subset E_k can be parametrized by a mapping $E \to E_k$ that introduces local coordinates, $\boldsymbol{\eta} = (\eta^1, \eta^2) \in E$, over $E_k \subset \Gamma_2$, where E is a polygonal domain in $\mathbb{R}^2$, the interpolation formula for a Q-noded approximation E_k^a of a generic surface element $E_k \subset \Gamma_2$ can be written as

$$\boldsymbol{\xi}(\boldsymbol{\eta}) \;=\; \sum_{q=1}^{Q} \psi_q(\boldsymbol{\eta})\, \boldsymbol{\xi}^q, \qquad \boldsymbol{\xi} \in E_k^a, \quad \boldsymbol{\xi}^q \in E_k, \quad \boldsymbol{\eta} \in E.$$

Here $\psi_q(\boldsymbol{\eta})$ are the Lagrange interpolation polynomials (shape functions) for the Q-noded element E_k^a with parent domain E, and $\boldsymbol{\xi}^q$ are the nodal points on E_k. Accordingly $\Gamma_2^a = \bigcup_{k=1}^{K} E_k^a$ is an approximation of Γ_2 so that the scattered tensor $\widetilde{\boldsymbol{U}}$ and distribution $\boldsymbol{h}$ featured in (5.1) via (4.5) can be approximated over E_k^a as

$$\widetilde{\boldsymbol{U}}(\boldsymbol{\xi}(\boldsymbol{\eta}), \boldsymbol{x}) = \sum_{p=1}^{Q} \psi_p(\boldsymbol{\eta})\, \widetilde{\boldsymbol{U}}(\boldsymbol{\xi}^p, \boldsymbol{x}), \qquad \boldsymbol{h}_a(\boldsymbol{\xi}(\boldsymbol{\eta})) = \sum_{q=1}^{Q} \psi_q(\boldsymbol{\eta})\, \boldsymbol{h}^q, \qquad \boldsymbol{\xi} \in E_k^a \subset \Gamma_2^a,$$

where $\boldsymbol{h}^q = \boldsymbol{h}(\boldsymbol{\xi}^q)$ and $\boldsymbol{x} \in \Gamma_1$. In what follows, it is assumed that the values of the scattered tensor $\widetilde{\boldsymbol{U}}(\boldsymbol{\xi}, \boldsymbol{x})$ are sampled over N_{S} source points $\{\boldsymbol{x}^i\}_1^{N_{\mathrm{S}}}$ on Γ_1, and N_{O} observation points $\{\boldsymbol{\xi}^j\}_1^{N_{\mathrm{O}}}$ on Γ_2. In this setting, an approximation of the near-field operator $\boldsymbol{G}$ over Γ_1 can be written as

$$(\boldsymbol{G}_a \boldsymbol{h}_a)(\boldsymbol{x}) = \sum_{k=1}^{K} \sum_{q=1}^{Q} \sum_{p=1}^{Q} [\widetilde{\boldsymbol{U}}(\boldsymbol{\xi}^{p_k}, \boldsymbol{x})]^{\mathrm{T}} \cdot \boldsymbol{h}^{q_k} \int_{E_k^a} \psi_q(\boldsymbol{\eta})\, \psi_p(\boldsymbol{\eta})\, J\, d\eta^1 d\eta^2, \quad \boldsymbol{x} \in \Gamma_1, \tag{5.2}$$

where $J = J(\boldsymbol{\eta})$ is the Jacobian of transformation (5.2), while p_k and q_k are the respective global indices of the pth and qth element nodes on E_k^a. On the basis of (5.2) and a set of collocation points $\{\boldsymbol{x}^i\}_1^{N_{\mathrm{S}}} \subset \Gamma_1$, a discretized form of the near-field integral equation (5.1) can be written as

$$\boldsymbol{G}_a \boldsymbol{h}_a = \boldsymbol{b}_a, \tag{5.3}$$

where $\boldsymbol{h}_a = (\boldsymbol{h}^1, \boldsymbol{h}^2, \ldots, \boldsymbol{h}^{N_{\rm O}})^{\rm T}$ is a vector containing the nodal values of $\boldsymbol{h}$ on Γ_2, $\boldsymbol{b}_a = (\widehat{\boldsymbol{U}}(\boldsymbol{x}^1, \boldsymbol{z}) \cdot \boldsymbol{d}, \widehat{\boldsymbol{U}}(\boldsymbol{x}^2, \boldsymbol{z}) \cdot \boldsymbol{d}, \ldots, \widehat{\boldsymbol{U}}(\boldsymbol{x}^{N_{\rm S}}, \boldsymbol{z}) \cdot \boldsymbol{d})^{\rm T}$, and $\boldsymbol{G}_a \in \mathbb{C}^{3N_{\rm S} \times 3N_{\rm O}}$ is a finite-dimensional approximation of the near-field operator $\boldsymbol{G}$ following (5.2) wherein the surface integrals are approximated via a product Gauss–Legendre quadrature.

In view of the ill-posed nature of (5.1), a suitable regularization is necessary to obtain a stable approximate solution of (5.3). To this end, let ε be an a priori estimate of the measurement and numerical errors characterizing $\boldsymbol{G}_a$ so that

$$\|\boldsymbol{G} - \boldsymbol{G}_a\|_{L_2(\Gamma_1)} \leq \varepsilon, \qquad \varepsilon = \gamma \, \|\boldsymbol{G}_a\|_{L_2(\Gamma_1)}, \qquad \gamma > 0, \tag{5.4}$$

and let the right-hand side $\boldsymbol{b}$, poluted with numerical inaccuracies, be known up to an error δ, whereby

$$\|\boldsymbol{b} - \boldsymbol{b}_a\|_{L_2(\Gamma_1)} \leq \delta, \qquad \delta = \beta \, \|\boldsymbol{b}_a\|_{L_2(\Gamma_1)}, \qquad \beta > 0. \tag{5.5}$$

In the ensuing (regularized) solution of the discrete system (5.3), Euclidean norm $\|\cdot\|$ in $\mathbb{C}^N$ induced by the inner product

$$(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^{N} \bar{u}_i v_i, \quad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^N, \tag{5.6}$$

will be assumed, where N is an appropriate dimension.

5.2. Tikhonov regularization. The Tikhonov regularization method [14, 41] replaces (5.3) with an equation of the second kind:

$$\boldsymbol{G}_a^* \boldsymbol{G}_a \boldsymbol{h}_a^\alpha + \alpha \boldsymbol{h}_a^\alpha \, = \, \boldsymbol{G}_a^* \boldsymbol{b}_a, \qquad \boldsymbol{h}_a \in \mathbb{C}^{3N_{\rm O}}, \tag{5.7}$$

where $\boldsymbol{G}_a^*$ denotes the conjugate transpose of $\boldsymbol{G}_a$, $\alpha > 0$ is the regularization parameter, and $\boldsymbol{h}_a^\alpha$ defacto minimizes the functional $J_\alpha(\boldsymbol{h}_a) = \|\boldsymbol{G}_a \boldsymbol{h}_a - \boldsymbol{b}_a\|^2 + \alpha \|\boldsymbol{h}_a\|^2$.

On employing the singular value decomposition of $\boldsymbol{G}_a$, it can be shown [10, 24] that the regularized solution $\boldsymbol{h}_a^\alpha$ of (5.7) and its squared norm admit the representation

$$\boldsymbol{h}_a^\alpha = \sum_{\nu_j > 0} \frac{\nu_j}{\alpha + \nu_j^2} \, (\boldsymbol{u}_j, \boldsymbol{b}_a) \, \boldsymbol{v}_j, \qquad \|\boldsymbol{h}_a^\alpha\|^2 = \sum_{\nu_j > 0} \frac{\nu_j^2}{\left(\alpha + \nu_j^2\right)^2} \, |(\boldsymbol{u}_j, \boldsymbol{b}_a)|^2 \, . \tag{5.8}$$

Here $\boldsymbol{u}_i \in \mathbb{C}^{3N_{\rm S}}$ $(i = 1, 2, \ldots, 3N_{\rm S})$ and $\boldsymbol{v}_j \in \mathbb{C}^{3N_{\rm O}}$ $(j = 1, 2, \ldots, 3N_{\rm O})$ denote respectively the left and right singular vectors of $\boldsymbol{G}_a$, and $\nu_k \in \mathbb{R}$, $k = 1, 2, \ldots, p = \min\{3N_{\rm S}, 3N_{\rm O}\}$ are the singular values of $\boldsymbol{G}_a$ ordered so that $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_p \geq 0$.

A method for choosing an optimal regularization parameter $\alpha = \alpha^*$, for which $\boldsymbol{h}_a^\alpha$ "closely" approximates the solution of (5.3), is given by the Morozov's discrepancy principle [14, 24, 29, 41]. In its most general form, the discrepancy principle due to Morozov states that the residual $\|\boldsymbol{G}_a \boldsymbol{h}_a^\alpha - \boldsymbol{b}_a\|$ should be commensurate to the errors characterizing the estimates of $\boldsymbol{G}$ and $\boldsymbol{b}$. With reference to (5.4)–(5.5), this implies

$$\|\boldsymbol{G}_a \boldsymbol{h}_a^\alpha - \boldsymbol{b}_a\| \, = \, \varepsilon \|\boldsymbol{h}_a^\alpha\| + \delta. \tag{5.9}$$

On assuming that $\delta \ll \varepsilon$ (the right-hand side, $\boldsymbol{b} = \widehat{\boldsymbol{U}}(\cdot, \boldsymbol{z}) \cdot \boldsymbol{d}$, is a known analytic function of real variables), one can neglect numerical inaccuracies in the computation of the right-hand side $\boldsymbol{b}$ in (5.9) and define the discrepancy function as

$$\zeta(\alpha) = \|\boldsymbol{G}_a \boldsymbol{h}_a^\alpha - \boldsymbol{b}_a\|^2 - \varepsilon^2 \|\boldsymbol{h}_a^\alpha\|^2, \qquad \alpha > 0. \tag{5.10}$$

On the basis of (5.8) and decomposition $\boldsymbol{b}_a = \sum_{\nu_j>0} \boldsymbol{u}_j\,(\boldsymbol{u}_j, \boldsymbol{b}_a)$, the discrepancy function (5.10) and its derivative can be respectively rewritten as

$$\zeta(\alpha) = \sum_{\nu_j>0} \frac{\alpha^2 - \varepsilon^2\nu_j^2}{\left(\alpha+\nu_j^2\right)^2}\,|(\boldsymbol{u}_j, \boldsymbol{b}_a)|^2\,, \qquad \zeta'(\alpha) = \sum_{\nu_j>0} \frac{2\nu_j^2(\alpha+\varepsilon^2)}{\left(\alpha+\nu_j^2\right)^3}\,|(\boldsymbol{u}_j, \boldsymbol{b}_a)|^2\,.$$

It is readily seen that $\zeta'(\alpha)>0$ for $\alpha\in(0,\infty)$, and hence the discrepancy function $\zeta(\alpha)$ is a monotonically increasing function. By virtue of the limit from above, asymptotic behavior $\lim_{\alpha\downarrow 0}\zeta(\alpha) < 0$, and the monotonicity of ζ, it follows that $\zeta(\alpha)$ has a unique root $\alpha^\star$, satisfying $\zeta(\alpha^\star) = 0$, which can be computed using, e.g., a root-finding Newton method.

5.3. Preconditioned conjugate gradient method. In situations where the singular value decomposition of $\boldsymbol{G}_a$ is not practical, e.g., for "large" systems, the conjugate gradient (CG) method [22, 24] can be alternatively employed to solve (5.3) wherein the regularized iterative solution $\boldsymbol{h}_a^\kappa$ is found by minimizing the functional $J(\boldsymbol{h}_a) = \|\boldsymbol{G}_a\boldsymbol{h}_a - \boldsymbol{b}_a\|^2$. In the iteration procedure of the CG method, iteration number κ plays the role of the regularization parameter; accordingly, its optimal value, $\kappa=\kappa^\star$, is to be chosen by a suitable stopping rule. In this investigation, a preconditioned CG method proposed by Santos [37] for ill-conditioned systems will be used. Rooted in [3], this technique can be briefly described using the decomposition $\boldsymbol{G}_a\boldsymbol{G}_a^* = \boldsymbol{T} + \boldsymbol{D} + \boldsymbol{T}^*$, where $\boldsymbol{T}$ is strictly lower triangular, $\boldsymbol{D} = \mathrm{diag}(d_1, d_2, \ldots, d_{3N_\mathrm{S}})$, and $d_i \in \mathbb{R}$ is the diagonal element of $\boldsymbol{G}_a\boldsymbol{G}_a^* \in \mathbb{C}^{3N_\mathrm{S}\times 3N_\mathrm{S}}$. With such definitions, let

$$\boldsymbol{C}_\tau = (\boldsymbol{D} + \tau\,\boldsymbol{T})\boldsymbol{D}^{-1/2} \tag{5.11}$$

form a basis for the preconditioner, where $\tau\in[0,2]$ is a relaxation parameter. In this setting, the regularized solution $\boldsymbol{h}_a^\kappa$ of (5.3) can be found by minimizing the functional

$$J_C(\boldsymbol{h}_a) = \|\boldsymbol{C}_\tau^{-1}(\boldsymbol{G}_a\boldsymbol{h}_a - \boldsymbol{b}_a)\|^2, \quad \boldsymbol{h}_a\in\mathbb{C}^{3N_\mathrm{O}},$$

i.e., by solving the the normal equation

$$\boldsymbol{G}_a^*\boldsymbol{C}_\tau^{-*}\boldsymbol{C}_\tau^{-1}\boldsymbol{G}_a\boldsymbol{h}_a \;=\; \boldsymbol{G}_a^*\boldsymbol{C}_\tau^{-*}\boldsymbol{C}_\tau^{-1}\boldsymbol{b}_a, \tag{5.12}$$

where $\boldsymbol{C}_\tau^{-*} = (\boldsymbol{C}_\tau^{-1})^*$. A modification of the CG algorithm (PCCGNR) [37] for solving (5.12), wherein the "net" residual $\boldsymbol{\gamma}^\kappa = \boldsymbol{b}_a - \boldsymbol{G}_a\boldsymbol{h}_a^\kappa$ is computed at every iterate κ, can be written as follows

ALGORITHM 5.1.

Given $\boldsymbol{h}_a^0$:

Set $\boldsymbol{\gamma}^0 = \boldsymbol{b}_a - \boldsymbol{G}_a\boldsymbol{h}_a^0, \quad \boldsymbol{r}^0 = \boldsymbol{G}_a^*\boldsymbol{C}_\tau^{-*}\boldsymbol{C}_\tau^{-1}\boldsymbol{\gamma}^0, \quad \boldsymbol{p}^1 = \boldsymbol{r}^0.$

For $\kappa = 1, 2, \ldots$

$$\begin{aligned}
&\boldsymbol{g}^\kappa = \boldsymbol{G}_a\,\boldsymbol{p}^\kappa, \quad \boldsymbol{q}^\kappa = \boldsymbol{C}_\tau^{-1}\boldsymbol{g}^\kappa,\\
&\alpha_\kappa = \frac{\|\boldsymbol{r}^{\kappa-1}\|^2}{\|\boldsymbol{q}^\kappa\|^2},\\
&\boldsymbol{h}_a^\kappa = \boldsymbol{h}_a^{\kappa-1} + \alpha_\kappa\,\boldsymbol{p}^\kappa,\\
&\boldsymbol{\gamma}^\kappa = \boldsymbol{\gamma}^{\kappa-1} - \alpha_\kappa\,\boldsymbol{g}^\kappa, \quad \boldsymbol{r}^\kappa = \boldsymbol{G}_a^*\boldsymbol{C}_\tau^{-*}\boldsymbol{C}_\tau^{-1}\boldsymbol{\gamma}^\kappa,\\
&\beta_{\kappa+1} = \frac{\|\boldsymbol{r}^\kappa\|^2}{\|\boldsymbol{r}^{\kappa-1}\|^2},\\
&\boldsymbol{p}^{\kappa+1} = \boldsymbol{r}^\kappa + \beta_{\kappa+1}\,\boldsymbol{p}^\kappa.
\end{aligned}$$

For situations involving large numbers of sampling points $\boldsymbol{z}$, the inverse of the lower triangular matrix, $\boldsymbol{C}_\tau^{-1}$, and consequently $\boldsymbol{G}_a^*\boldsymbol{C}_\tau^{-*}\boldsymbol{C}_\tau^{-1}$, can be precomputed explicitly, as they are independent of $\boldsymbol{z}$. In the contrary cases involving only a limited number of sampling points, on the other hand, it may be beneficial to compute the products $\boldsymbol{C}_\tau^{-1}\boldsymbol{g}$ and $\boldsymbol{C}_\tau^{-*}\boldsymbol{w}$ without calculating the inverse of (5.11) explicitly. To this end, let $\boldsymbol{f}_i \in \mathbb{C}^{3N_{\rm O}}$ denote the ith row of $\boldsymbol{G}_a$, and let $d_i = (\boldsymbol{f}_i, \boldsymbol{f}_i)$ be the ith diagonal entry of $\boldsymbol{D} = \mathrm{diag}(d_1, d_2, \ldots, d_{3N_{\rm S}})$, where the inner product $(\cdot,\cdot)$ is given by (5.6). With such definitions, the components of $\boldsymbol{q} = \boldsymbol{C}_\tau^{-1}\boldsymbol{g}$ can be computed in a recursive fashion [3, 37] as

$$q_i = d_i^{-1/2}(g_i - \tau\,(\boldsymbol{a}_i, \boldsymbol{f}_i))\,, \qquad \boldsymbol{a}_{i+1} = \boldsymbol{a}_i + d_i^{-1/2}\,\overline{q}_i\,\boldsymbol{f}_i, \qquad i = 1, 2, \ldots, 3N_{\rm S},$$

where $\mathbb{C}^{3N_{\rm O}} \ni \boldsymbol{a}_1 = \boldsymbol{0}$ initializes the procedure. Similarly by letting $\mathbb{C}^{3N_{\rm O}} \ni \boldsymbol{a}_{3N_{\rm S}} = \boldsymbol{0}$, the components of $\boldsymbol{s} = \boldsymbol{C}_\tau^{-*}\boldsymbol{w}$ can be computed as

$$s_i = d_i^{-1/2} w_i - \tau\, d_i^{-1}(\boldsymbol{a}_i, \boldsymbol{f}_i), \qquad \boldsymbol{a}_{i-1} = \boldsymbol{a}_i + \overline{s}_i\,\boldsymbol{f}_i, \qquad i = 3N_{\rm S}, 3N_{\rm S}-1, \ldots, 1.$$

The selection of an optimal iteration number (regularization parameter) $\kappa = \kappa^\star$ in Algorithm 5.1 is rather heuristic. As mentioned in [40], a generalization of the discrepancy principle manifest in (5.10) to CG-type methods is still an open question. Nevertheless one may, by analogy to (5.10), introduce the discrepancy function as

$$\zeta(\kappa) \;=\; \|\boldsymbol{G}_a\,\boldsymbol{h}_a^\kappa - \boldsymbol{b}_a\|^2 - \varepsilon^2\|\boldsymbol{h}_a^\kappa\|^2, \qquad \kappa \in \mathbb{N}\cup\{0\}. \tag{5.13}$$

By virtue of (5.13), one can select the optimal iteration number $\kappa^\star$ as the number κ that corresponds to the minimum of $|\zeta(\kappa)|$, a quantity whose computation at every iterate is facilitated by the computation of the "net" residual $\boldsymbol{\gamma}^\kappa$ in Algorithm 5.1.

6. Results and discussion. On the basis of the foregoing developments, the task of reconstructing an obstacle $\Omega_{\rm C}$ in the semi-infinite solid Ω from near-field elastic waveforms (Figure 2.1) can be achieved by solving either (3.1) or its adjoint counterpart (4.1) over a sampled region $\mathcal{D} \supset \Omega_{\rm C}$ by means of the featured regularization methods. By introducing the grid of sampling points $\boldsymbol{z}^m \in \mathcal{D} \subset \Omega$ $(m = 1, 2, \ldots, M)$ spanning the region of interest, Ω is sequentially excited by the virtual point sources acting at $\boldsymbol{z}^m$ in direction $\boldsymbol{d}$, and $1/\|\boldsymbol{h}(\cdot; \boldsymbol{z}^m, \boldsymbol{d})\|_{L_2(\Gamma_2)}$ is plotted over the selected raster.

6.1. Testing configuration. With reference to Figure 6.1, consider the problem of reconstructing a dual cavity consisting of (i) a sphere of diameter 1.6 centered at $(-2, -2, 3)$, and (ii) an ellipsoid centered at $(2, 1, 3)$ whose axes, of lengths $(3, 1.6, 1.6)$, are aligned with the reference Cartesian frame $\{O; \xi_1, \xi_2, \xi_3\}$. The Lamé parameters and mass density of the elastic solid are taken as $\lambda = \frac{7}{3}$, $\mu = 1$, and $\rho = 1$, corresponding to a Poisson ratio of 0.35. On assuming that the source surface Γ_1 and the observation surface Γ_2 coincide, i.e., $\Gamma_1 = \Gamma_2 = \Pi$, synthetic observations of the scattered tensor $\widetilde{\boldsymbol{U}}$ are generated via an elastodynamic boundary element method [33] by assuming $N_{\rm S} \in \{8, 21, 40, 65, 96\}$ source points and $N_{\rm O} = 96$ receiver points regularly distributed over the square test area (dimensions 10×10). For quadrature purposes, both the source and the receiver grid are each associated with a uniform "mesh" of eight-node surface elements. In this setting, $N_{\rm S} = 8, 21, 40, 65, 96$ correspond respectively to the $k \times k$ mesh of surface elements, $k = 1, 2, 3, 4, 5$. From every source point $\boldsymbol{x}^k$ on the grid $(k = 1, 2, \ldots, N_{\rm S})$, the half-space is sequentially illuminated using time-harmonic force of magnitude $P = \sqrt{3}\mu a^2$ acting along the coordinate directions ξ_1, ξ_2,

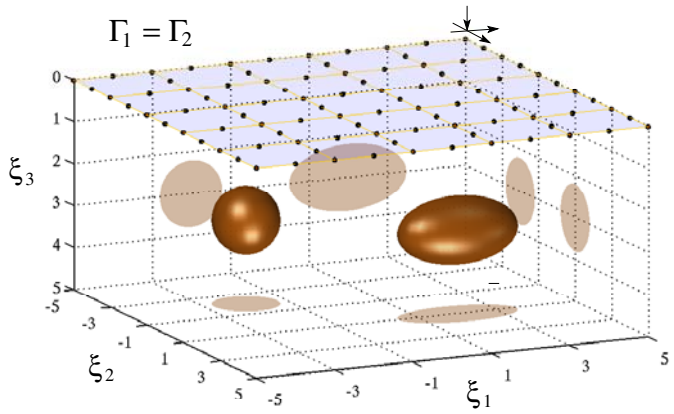


FIG. 6.1. *Dual cavity and testing grid (5×5 surface elements) in a semi-infinite elastic solid with* $N_S = N_O = 96$.

and ξ_3. For each $\boldsymbol{x}^k$, the synthetic data $\widetilde{\boldsymbol{U}}(\boldsymbol{\xi}^j, \boldsymbol{x}^k)$ are computed over $N_{\rm O} = 96$ receiver points $\boldsymbol{\xi}^j$, as examined earlier. To examine the effect of measurement uncertainties, synthetic observations $\widetilde{\boldsymbol{U}}$ in selected examples are corrupted as

$$\widetilde{\boldsymbol{U}}(\boldsymbol{\xi}^j, \boldsymbol{x}^k) \;:=\; (1 + \varrho\chi)\, \widetilde{\boldsymbol{U}}(\boldsymbol{\xi}^j, \boldsymbol{x}^k), \qquad \begin{matrix} j = 1, 2, \ldots, N_{\rm O}, \\ k = 1, 2, \ldots, N_{\rm S}, \end{matrix}$$

where ϱ is the noise amplitude and $\chi \in [-1, 1]$ is a uniform random variable.

With the above problem parameters, near-field equations (3.1) and (4.1) are discretized as examined in section 5 and solved, assuming $\boldsymbol{d} = \frac{1}{\sqrt{3}}(1, 1, 1)^{\rm T}$, for the densities $\boldsymbol{g}_{\boldsymbol{z},\boldsymbol{d}}$ and $\boldsymbol{h}_{\boldsymbol{z},\boldsymbol{d}}$, repectively, over a grid of uniformly spaced sampling points in the horizontal plane $\xi_3 = 3$ and vertical planes $\xi_2 = -2, 1$. For completeness, representative numerical results are computed using both Tikhonov regularization (TR) and the preconditioned conjugate gradient (PCG) methods, assuming $\tau = 0.2$ for the relaxation parameter. With reference to the discrepancy functions (5.10) and (5.13), an estimate ε of the "measurement" and numerical errors characterizing the near-field operator is computed as $\varepsilon = \gamma \|\boldsymbol{G}_a\|$, where $\|\boldsymbol{G}_a\|$ is given by the maximum singular value of $\boldsymbol{G}_a$ for the TR method, and by the Frobenius norm of $\boldsymbol{G}_a$ for the PCG method.

To facilitate the comparison of results, each sectional distribution of the reciprocal solution density norm is accompanied by (i) an intersection with the boundary of the "true" obstacle indicated via a dark (red) solid line, and (ii) a white dashed isoline corresponding to a fraction, $\mathcal{R}$, of the peak (maximum) value in the plot. For clarity, the selected threshold level is also featured on the color bar accompanying each graph.

6.2. Adjoint versus direct sampling method. By setting $\gamma = 10^{-7}$ in (5.4) and assuming no extraneous noise on the measurements $\widetilde{\boldsymbol{U}}$ (i.e., $\varrho = 0$), Figures 6.2 and 6.3 depict the contour plots of $1/\|\boldsymbol{g}(\cdot; \boldsymbol{z}, \boldsymbol{d})\|_{L_2(\Pi)}$ (top row) and $1/\|\boldsymbol{h}(\cdot; \boldsymbol{z}, \boldsymbol{d})\|_{L_2(\Pi)}$ (bottom row) computed using the TR method for $\omega = 4$ and $N_{\rm O} = 96$ under the *decreasing number* of source points, namely $N_{\rm S} \in \{96, 65, 40, 21, 8\}$. In particular, the top panels in each figure are computed via direct formulation (3.1), while their bottom companions are associated with the adjoint formula (4.1). In Figure 6.2, which compares the methods for $N_{\rm S} \in \{96, 65, 40\}$, the dashed isolines are computed

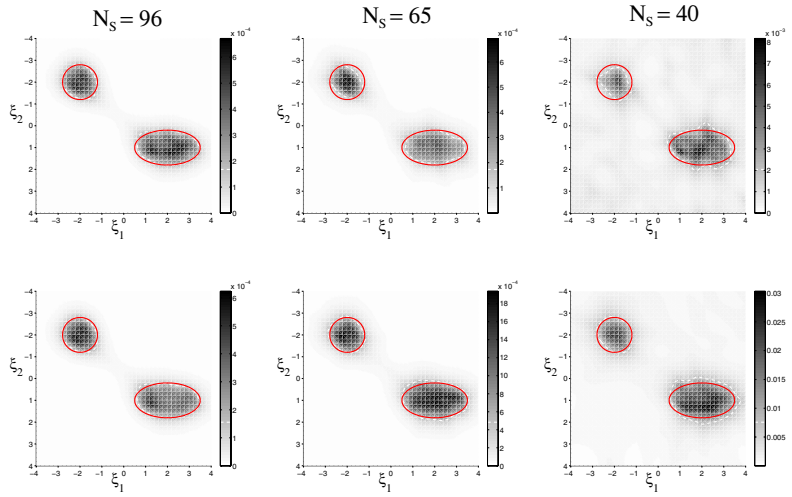


FIG. 6.2. *Images of dual cavity in the $\xi_3 = 3$ plane under decreasing number of source points—$5 \times 5, 4 \times 4$, and 3×3 grids of surface elements: direct (top row) versus adjoint formulation (bottom row) with $N_O = 96$, $\omega = 4$, and $\varrho = 0$.*

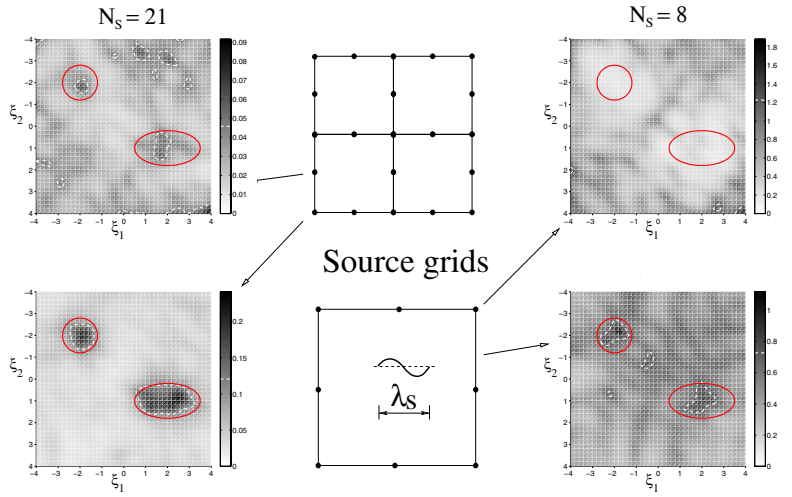


FIG. 6.3. *Images of dual cavity in the $\xi_3 = 3$ plane under decreasing number of source points—2×2 and 1×1 grids of surface elements: direct (top row) versus adjoint formulation (bottom row) with $N_O = 96$, $\omega = 4$, and $\varrho = 0$.*

assuming $\mathcal{R} = 0.25$, a threshold value which was found to perform well (in terms of obstacle reconstruction) when dealing with "high-quality" observations. Owing to the fact that $\Gamma_1 = \Gamma_2$, the images in Figure 6.2 for $N_S = N_O = 96$ resulting respectively from (3.1) and (4.1) should be identical. Indeed, the two distributions are similar, and the apparent (minor) differences reflect the finite accuracy of three-dimensional boundary element simulations used to generate the synthetic measurements. Under reduced N_S, however, the top images stemming from (3.1) became progressively more smeared than their adjoint counterparts. In particular the inspection of results for

$N_S = 40$ indicates that, despite an apparent similarity of the respective isolines, the gray tones are notably more localized inside the support of the defect for the (bottom) "adjoint" image computed from (4.1), indicating higher quality of reconstruction. This trend is further highlighted in Figure 6.3 dealing with "severely limited" obstacle illumination where $N_S \in \{21, 8\}$. As a point of reference, the two sets of results also feature the isolines corresponding respectively to $\mathcal{R} \in \{0.50, 0.65\}$ (applied to compensate for image deterioration), as well as the schematics of respective source grids with the shear wave length plotted to scale. As can be seen from the display, the adjoint formulation of the linear sampling method continues to outperform its "direct" counterpart with the differences becoming more pronounced with decreasing N_S.

With reference to the results in Figures 6.2 and 6.3, it is noted that the rank of the discretized operators $\boldsymbol{F}_a$ in (3.1) and $\boldsymbol{G}_a$ in (4.1) was found, as expected, to be $(3N_S)^2$ for all configurations examined. Here the factor of 3 appears due to the fact that the obstacle is sequentially illuminated by a point force in each coordinate direction from every source location $\boldsymbol{x}^k$, $k = 1, 2, \ldots, N_S$. In this sense both (3.1) and (4.1) operate on the same data set (for a given N_S), and the respective drawbacks of these two methods, when used in the context of limited obstacle illumination, can be described as that of *underintegration* versus *undercollocation*. From the numerical results in Figures 6.2 and 6.3, it follows that the adjoint variant (4.1) of the LSM makes better use of such limited experimental data where the scattered tensor $\widetilde{\boldsymbol{U}}(\boldsymbol{\xi}, \boldsymbol{x})$ is undersampled with respect to its second (i.e., source) argument. This conclusion is uniformly supported by the results from a number of other testing configurations whose results are herein omitted for brevity. From the practical point of view, the above results further indicate that a *comparative* defect reconstruction using *both* direct and adjoint formulation of the LSM may provide an effective tool for exposing an apparent undersampling of the scattered tensor $\widetilde{\boldsymbol{U}}$ with respect to either argument. As an example, such comparative study from Figures 6.2 and 6.3 arguably indicates that, for a given testing aperture as controlled by Γ_1 and Γ_2, the scattered tensor is undersampled with respect to its second argument for $N_S \leq 40$. While a similar conclusion for the configuration of interest (where $\Gamma_1 = \Gamma_2$) could be obtained using an independent argument of spatial aliasing [38], the above heuristic approach for exposing the undersampling of $\widetilde{\boldsymbol{U}}$ could be equally applied to more complex situations where the source and receiver surfaces are associated with distinct "viewing" apertures, i.e., situations where the conventional indicators of undersampling may be insufficient.

6.3. TR versus PCG. To diversify the computational treatment of the LSM, section 5.3 describes an application of the PCG algorithm to ill-conditioned linear systems featured in (5.3). By its nature, such an alternative method of solution, i.e., regularization, may be particularly useful in testing situations involving large numbers of sources and receivers (the issue of their density set aside), as in three-dimensional seismic imaging [38], where an application of the singular value decomposition in terms of TR may lead to substantial computational cost and inaccurate singular values. While the numerical simulation of such "large" sets of experimental data is beyond the scope of this study, one may pose a question as to the relative performance of the PCG and TR methods in the context of "small" experimental data sets such as those in Figures 6.2 and 6.3, where the TR method is expected to have an advantage owing to an explicit knowledge of the singular values of the system.

In the above setting, Figure 6.4 "borrows" the example with $N_S = 40$ from Figure 6.2, where $N_O = 96$, $\omega = 4$, and $\varrho = 0$, as a platform for the comparison between the

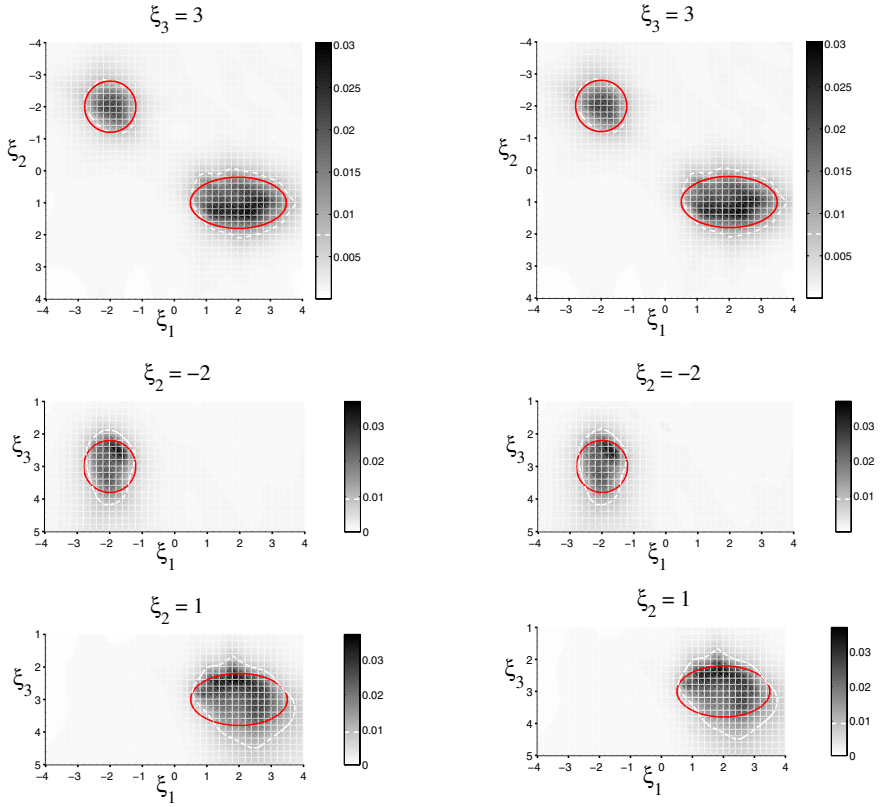


FIG. 6.4. *Images of dual cavity stemming from the adjoint formulation in the horizontal ($\xi_3=3$) and vertical ($\xi_2=-2,1$) planes computed using TR (left panels) and PCG (right panels). Problem parameters: $\omega=4$, $N_S=40$, $N_O=96$, and $\varrho=0$.*

TR method (left panels) and the PCG method (right panels) in terms of the adjoint formulation (4.1). Here the distributions of $1/\|\boldsymbol{h}(\cdot;\boldsymbol{z},\boldsymbol{d})\|_{L_2(\Pi)}$ are plotted both in the horizontal ($\xi_3=3$) and vertical ($\xi_2=-2,1$) sections. For ease of comparison, isolines of $1/\|\boldsymbol{h}\|_{L_2(\Pi)}$ corresponding to $\mathcal{R}=0.25$ are included as before. With reference to the discussion in section 6.2, application of the TR and PCG methods to a "slightly" undercollocated system (5.3) with $N_{\rm S}=40$ and $N_{\rm O}=96$ in Figure 6.4 yields distributions that are barely distinguishable despite the distinct computational treatments. To examine the effect of noise in the measurements, the latter comparison is repeated in Figure 6.5, but this time assuming $\varrho=0.04$, i.e., the 4%-level of experimental errors in (6.1). From the diagram where the isolines are computed for $\mathcal{T}=0.5$, one may observe that (i) the sectional images are relatively stable with respect to the measurement noise, and (ii) the TR and PCG methods still produce comparable results, with the PCG-based image in the $\xi_2=1$ (vertical) section providing a slightly better reconstruction of the "bottom" of the ellipsoid.

As a final illustration, the effect of diminishing $N_{\rm S}$ on the performance of the two regularization methods is illustrated in Figure 6.6 where $N_{\rm S}=21$, $N_{\rm O}=96$, $\varrho=0$, and $\mathcal{T}=0.5$. In this case the edge (as expected) belongs to the TR results, even though the PCG method still performs satisfactorily. In terms of the computational effort, it is noted that for each horizontal section in Figure 6.5 that features $41^2=1681$ sampling

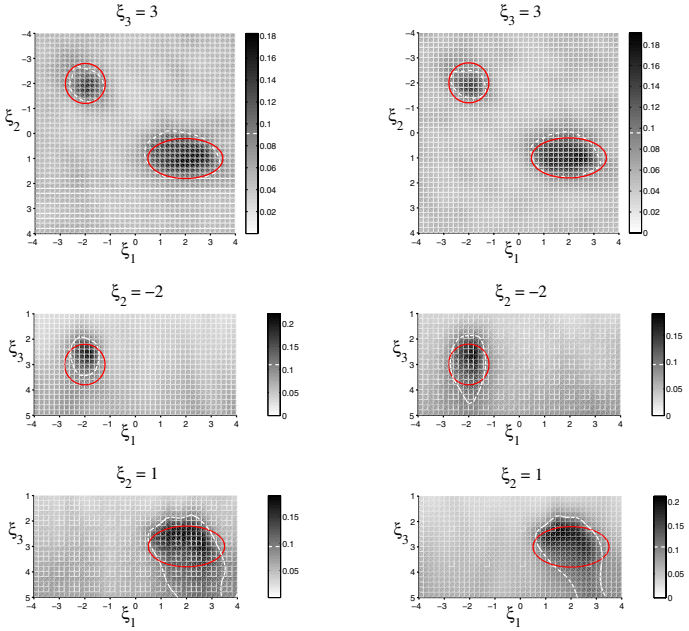


FIG. 6.5. *Images of dual cavity stemming from the adjoint formulation in the horizontal* ($\xi_3 = 3$) *and vertical* ($\xi_2 = -2, 1$) *planes computed using TR (left panels) and PCG (right panels). Problem parameters:* $\omega = 4$, $N_S = 40$, $N_O = 96$, *and* $\varrho = 0.04$.

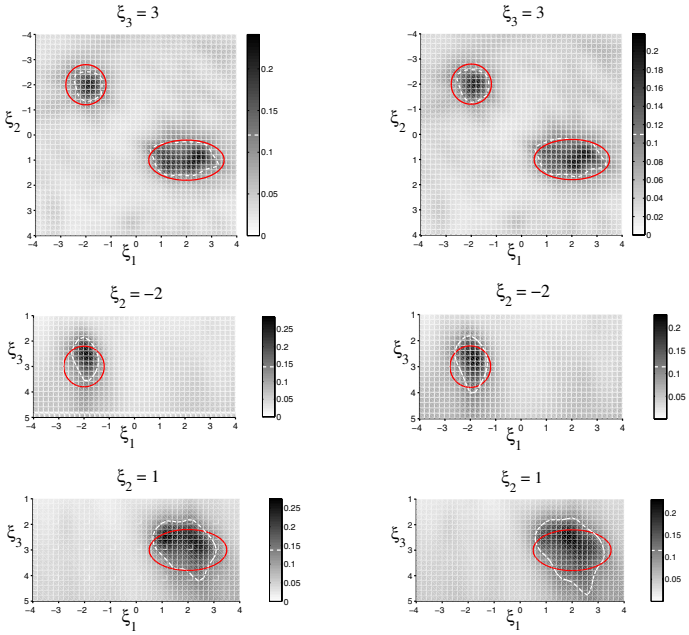


FIG. 6.6. *Images of dual cavity stemming from the adjoint formulation in the horizontal* ($\xi_3 = 3$) *and vertical* ($\xi_2 = -2, 1$) *planes computed using TR (left panels) and PCG (right panels) methods. Problem parameters:* $\omega = 4$, $N_S = 21$, $N_O = 96$, *and* $\varrho = 0$.

points, the computation of $1/\|\boldsymbol{h}\|_{L_2(\Pi)}$ took approximately 19 minutes for the TR method and 21 minutes for the PCG method on a Linux system with a 2.4 GHz Opteron processor. Here one should bear in mind that the majority of computational effort in each case is spent on the calculation of the half-space Green's function [18] featured on the right-hand side of (4.1). For significantly larger experimental systems, however, the computational advantage is naturally expected to shift toward the PCG method.

7. Summary. In this study, three-dimensional inverse scattering problem involving near-field elastodynamic reconstruction of impenetrable obstacles in a semi-infinite solid is examined by way of the linear sampling method (LSM). To cater to active imaging configurations characterized by a limited density of illuminating sources, an adjoint formulation of the near-field LSM is established that features a linear integral equation of the first kind involving integration over the *measurement* (as opposed to the source) surface. To diversify the computational treatment of ill-posed systems involving a significant number of experimental observations, a finite-dimensional approximation of the featured integral equation is solved by alternative means of Tikhonov regularization and a preconditioned conjugate gradient method. Computational details of the imaging technique, including evaluation of the featured integrals as well as the implementation of regularization strategies, are highlighted. Numerical results indicate that the adjoint variant of the LSM outperforms its predecessor (the so-called direct formulation) in situations involving a limited density of illuminating sources. From the practical standpoint, it is also found that a combined defect reconstruction by alternative means of the adjoint and direct sampling methods provides a rational basis for exposing an apparent undersampling of the experimental input, synthesized via the so-called scattered tensor. The results further indicate that the CG method, while designed primarily for the treatment of "large" systems involving a significant amount of experimental observations, may perform satisfactorily even for "small" systems that are characterized by a subpar number of illuminating sources.

Acknowledgments. The support provided by University of Minnesota Supercomputing Institute during the course of this investigation is kindly acknowledged.

REFERENCES

[1] T. Arens, *Linear sampling methods for 2D inverse elastic wave scattering*, Inverse Problems, 17 (2001), pp. 1445–1464.
[2] M. Bonnet, *BIE and material differentiation applied to the formulation of obstacle inverse problems*, Engrg. Anal. Bound. Elem., 15 (1995), pp. 121–136.
[3] A. Björck and T. Elfving, *Accelerated projection methods for computing pseudoinverse solutions of systems of linear equations*, BIT, 19 (1979), pp. 145–163.
[4] C. Bunks, F. M. Saleck, S. Zaleski, and G. Chavent, *Multiscale seismic waveform inversion*, Geophys., 60 (1995), pp. 1457–1473.
[5] A. Charalambopoulos, D. Gintides, and K. Kiriaki, *The linear sampling method for the transmission problem in three-dimensional linear elasticity*, Inverse Problems, 18 (2002), pp. 547–558.
[6] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, Wiley, New York, 1983.
[7] D. Colton and A. Kirsch, *A simple method for solving inverse scattering problems in the resonance region*, Inverse Problems, 12 (1996), pp. 383–393.
[8] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer, Berlin, 1998.

[9] D. Colton and P. Monk, *A linear sampling method for the detection of leukemia using microwaves*, SIAM J. Appl. Math., 58 (1998), pp. 926–941.
[10] D. Colton, K. Giebermann, and P. Monk, *Regularized sampling method for solving three-dimensional inverse scattering problems*, SIAM J. Sci. Comput., 21 (2000), pp. 2316–2330.
[11] D. Colton, J. Coyle, and P. Monk, *Recent developments in inverse acoustic scattering theory*, SIAM Rev., 42 (2000), pp. 369–414.
[12] M. M. Doyley, S. Srinivasan, E. Dimidenko, N. Soni, and J. Ophir, *Enhancing the performance of model-based elastography by incorporating additional a priori information in the modulus image reconstruction process*, Phys. Med. Biol., 51 (2006), pp. 95–112.
[13] E. Ebbini, *Phase-coupled two-dimensional speckle tracking algorithm*, IEEE Trans. Ultrasonics Ferroelectrics Freq. Control, 53 (2006), pp. 972–989.
[14] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.
[15] A. C. Eringen and E. S. Suhubi, *Elastodynamics*, Academic Press, New York, 1975.
[16] M. Fatemi and J. F. Greenleaf, *Probing the dynamics of tissue at low frequencies with the radiation force of ultrasound*, Phys. Med. Biol., 45 (1998), pp. 1449–1464.
[17] P. R. Garabedian, *Partial Differential Equations*, AMS, Chelsea, RI, 1998.
[18] B. B. Guzina and R. Y. S. Pak, *On the analysis of wave motions in a multi-layered solid*, Quart. J. Mech. Appl. Math., 54 (2001), pp. 13–37.
[19] B. B. Guzina, S. Nintcheu Fata, and M. Bonnet, *On the stress-wave imaging of cavities in a semi-infinite solid*, Int. J. Solids Structures, 40 (2003), pp. 1505–1523.
[20] B. B. Guzina and M. Bonnet, *Topological derivative for the inverse scattering of elastic waves*, Quart. J. Mech. Appl. Math., 57 (2004), pp. 161–179.
[21] B. B. Guzina and I. Chikichev, *From imaging to material characterization: A generalized concept of topological sensitivity*, J. Mech. Phys. Solids, 55 (2007), pp. 245–279.
[22] M. Hanke, *Conjugate Gradient Type Methods for Ill-posed Problems*, Pitman Res. Notes in Math., Longman Scientific & Technical, Harlow, UK, 1995.
[23] F. E. Kennedy, M. M. Doyley, E. E. Van Houten, J. B. Weaver, and K. D. Paulsen, *Determination of in-vivo elastic properties of soft tissue using magnetic resonance elastography*, ASME Adv. Bioengineering, 55 (2003), pp. 327–328.
[24] A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, Springer, Berlin, 1996.
[25] R. Kress, *Linear Integral Equation*, Springer, Berlin, 1999.
[26] V. D. Kupradze, *Three Dimensional Problems of the Mathematical Theory of Elasticity and Thermoelasticity*, North–Holland, Amsterdam, 1979.
[27] A. I. Madyarov and B. B. Guzina, *A radiation condition for layered elastic media*, J. Elasticity, 82 (2006), pp. 73–98.
[28] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, London, 2000.
[29] V. A. Morozov, *Methods for Solving Incorrectly Posed Problems*, Springer, Berlin, 1984.
[30] J. C. Nédélec, *Acoustic and Electromagnetic Equations*, Springer, Berlin, 2001.
[31] S. Nintcheu Fata, B. B. Guzina, and M. Bonnet, *Computational framework for the BIE solution to inverse scattering problems in elastodynamics*, Comp. Mech., 32 (2003), pp. 370–80.
[32] S. Nintcheu Fata and B. B. Guzina, *A linear sampling method for near-field inverse problems in elastodynamics*, Inverse Problems, 20 (2004), pp. 713–736.
[33] R. Y. S. Pak and B. B. Guzina, *Seismic soil-structure interaction analysis by direct boundary element methods*, Int. J. Solids Structures, 36 (1999), pp. 4743–4766.
[34] G. Pelekanos and V. Sevroglou, *Inverse scattering by penetrable objects in two-dimensional elastodynamics*, J. Comput. Appl. Math., 151 (2003), pp. 129–140.
[35] R. E. Plessix, Y. H. De Roeck, and G. Chavent, *Waveform inversion of reflection seismic data for kinematic parameters by local optimization*, SIAM J. Sci. Comput., 20 (1999), pp. 1033–1052.
[36] R. G. Pratt, F. C. Gao, C. Zelt, and A. Levander, *The complementary nature of traveltime and waveform tomography*, in Proceedings of the International Conference on Sub-basalt Imaging, Cambridge, UK, 2002; see J. Conference Abstracts, 7 (2002), pp. 181–183 (electronic).
[37] R. J. Santos, *Preconditioning conjugate gradient with symmetric algebraic reconstruction technique (ART) in computerized tomography*, Appl. Numer. Math., 47 (2003), pp. 255–263.
[38] R. E. Sheriff and L. P. Geldart, *Exploration Seismology*, Cambridge University Press, London, 1995.

[39] R. Sinkus, J. Lorenzen, D. Schrader, M. Lorenzen, M. Dargatz, and D. Holz, *High-resolution tensor MR elastography for breast tumour detection*, Phys. Med. Biol., 45 (2000), pp. 1649–1664.

[40] A. Tacchino, J. Coyle, and M. Piana, *Numerical validation of the linear sampling method*, Inverse Problems, 18 (2002), pp. 511–527.

[41] A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov, and A. G. Yagola, *Numerical Methods for the Solution of Ill-Posed Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995.

SIAM J. APPL. MATH.
Vol. 67, No. 5, pp. 1353–1378

INVERSE SOURCE PROBLEM IN NONHOMOGENEOUS BACKGROUND MEDIA*

ANTHONY J. DEVANEY†, EDWIN A. MARENGO†, AND MEI LI†

Abstract. The scalar wave inverse source problem (ISP) of determining an unknown radiating source from knowledge of the field it generates outside its region of localization is investigated for the case in which the source is embedded in a nonhomogeneous medium with known index of refraction profile $n(\mathbf{r})$. It is shown that the solution to the ISP having minimum energy (the so-called minimum energy source) can be obtained via a simple method of constrained optimization. This method is applied to the special case when the nonhomogeneous background is spherically symmetric ($n(\mathbf{r}) = n(r)$), and it yields the minimum energy source in terms of a series of spherical harmonics and radial wave functions that are solutions to a Sturm–Liouville problem. The special case of a source embedded in a spherical region of constant index is treated in detail, and results from computer simulations are presented for this case.

Key words. inverse source problem, nonhomogeneous media, inhomogeneous media, antenna substrate, scattering, reciprocity

AMS subject classifications. 45Q05, 78A46, 78A40, 78A50, 35-02

DOI. 10.1137/060658618

1. Introduction.

We consider in three-dimensional space the fundamental inverse source problem (ISP) of determining an unknown scalar source ρ to the inhomogeneous Helmholtz equation

$$[\nabla^2 + k_0^2 n^2(\mathbf{r})]U(\mathbf{r}) = -\rho(\mathbf{r}) \tag{1.1}$$

(where ∇^2 denotes the Laplacian operator) that radiates a scalar field U which is specified everywhere *outside* the support volume τ of the source. In this equation, k_0 is a constant wavenumber, and $n(\mathbf{r})$ is an index of refraction distribution that depends on position $\mathbf{r} \in R^3$ and that is assumed to go to unity for sufficiently large r. We will assume throughout this paper that the source volume τ is a sphere, centered at the origin and having a radius a. The ISP then consists of computing a source ρ that generates a prescribed *exterior* field U, whose value is specified for $\mathbf{r} \notin \tau$.

There are a number of treatments of the scalar ISP of interest in this work as well as of the full vector electromagnetic inverse problem for the free space case where the index distribution $n(\mathbf{r})$ is constant (equal to unity) throughout space [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. Most of these treatments make use of the fact that the source's radiation pattern (see below) determines, in principle, the field everywhere outside the source volume [11]. Using this fact, the ISP can be cast in terms of the radiation pattern: determine a source ρ that generates a prescribed radiation pattern. It is also well known [12, 13, 14] that there exist an infinity of sources that radiate fields that vanish identically outside their support volumes so that the ISP does not possess a unique

*Received by the editors May 1, 2006; accepted for publication (in revised form) March 28, 2007; published electronically July 11, 2007. This work was supported by the Air Force Office of Scientific Research under grant FA9550-06-01-0013, and is affiliated with CenSSIS, the Center for Subsurface Sensing and Imaging Systems, under the Engineering Research Centers Program of the National Science Foundation (award EEC-9986821).

http://www.siam.org/journals/siap/67-5/65861.html

†Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 (Tonydev2@aol.com, emarengo@ece.neu.edu, mei_li_cdsp@yahoo.com).

solution; i.e., an infinity of solutions can be obtained by adding any one of these *nonradiating sources* to any given solution [15, 16]. Thus, in order to obtain a unique solution to the ISP it is necessary to add constraints that the source must satisfy in addition to yielding a specified radiation pattern. An important and tractable choice of constraint is source energy $\mathcal{E}$, defined to be the L^2 norm of the source over the source volume τ:

$$\mathcal{E} = \int_\tau d^3r \, |\rho(\mathbf{r})|^2. \tag{1.2}$$

The solution to the ISP that minimizes the source energy as defined in (1.2) is usually termed the *minimum energy solution* [2, 3, 6, 7, 8, 9, 10]. It is orthogonal to the nonradiating sources [3] and is the pseudoinverse of the ISP [5, 6]. Physically, this source is also related to the real image field generated by a point-reference hologram of the field recorded on a closed surface completely surrounding the source volume [2, 3, 17].

The source energy as defined in (1.2) is an important measure of the realizability of a source generating a given radiation pattern with the given spatial resource or source volume τ. For the free space case one finds that the energy of the minimum energy source depends critically on the product $x = k_0 a$ of the (free space) wavenumber k_0 and the source radius a [2, 6]. For a given radiation pattern this energy $\mathcal{E}(x)$ is small for $x > l_0$, where l_0 is a parameter that characterizes the radiation pattern and that increases with increasing fine detail in the pattern. However, it is found that $\mathcal{E}(x)$ increases exponentially with decreasing x below the critical value l_0. This exponential increase of source energy indicates that the given radiation pattern cannot be physically realized by *any* source having that specific $k_0 a$ product: it is necessary to either decrease the wavelength (increase k_0) or increase the source radius. This is analogous to the well-known result in antenna theory that states that reactive energy and the quality Q of an antenna increase exponentially with decreasing $k_0 a$ if one attempts to achieve superdirectivity [18, 19]. Furthermore, in the associated full vector treatment the source energy is also a measure of the current levels of the antenna structure, which ideally should be small to cope with ohmic losses in realistic lossy antenna material. In particular, the ability of a source (antenna) to radiate a prescribed field with reduced source energy or "resources" is an indication of efficiency, so that the constraint of minimizing the source energy for a given radiation field is of interest not only for the theoretical treatment of the inverse source problem and of the related field realizability question but also for the practical antenna synthesis problem (see, e.g., the antenna characterizations in [20], the sensitivity factor in [21], and similar factors used in [22, 23, 24]).

As far as the authors of this paper know there are only two treatments of the ISP for the case where the background index distribution $n(\mathbf{r})$ in which the source is embedded is nonhomogeneous [25, 26]. The work in [26] generalizes the main results in [25] to lossy media. Of particular interest to the present research, which focuses on lossless media, is [25], which shows that the minimum energy solution satisfies an integral equation whose kernel is the imaginary part of the outgoing wave Green function of the inhomogeneous Helmholtz equation (1.1). Using this fact, it is shown in [25] that the minimum energy solution can be expanded into a series of eigenfunctions of this integral equation with expansion coefficients that can be determined from the radiation pattern. The paper shows that the formalism reduces to the known theory of the ISP when the index $n = 1$ (homogeneous medium case), but no examples of the general theory were provided.

In this paper we consider the case of a source embedded in a known nonhomogeneous real background index (corresponding to a lossless medium) and solve the minimum energy ISP using a simple method of constrained optimization. Besides providing a simpler formulation of the problem than that used in [25], the method yields a solution that is directly implemented without the need of first computing a Green function for the Helmholtz equation (1.1) and then computing the eigenfunctions of the imaginary part of this quantity. The special case of a spherically symmetric index $n(\mathbf{r}) = n(r)$ is treated in detail, and it is shown that the minimum energy solution to the ISP has exactly the same mathematical form as the solution for the constant index case, but with the spherical Bessel functions replaced by radial wave functions that are solutions to a Sturm–Liouville problem. In particular, these radial wave functions are the radial wave scattering functions obtained in the scattering of an incident plane wave from the spherically symmetric index distribution $n(r)$. This latter problem has been studied extensively in quantum mechanical [27], optical [28], and electromagnetic [29] scattering, and there exist a number of index distributions for which the scattering wave functions have been computed and that can be used to compute the minimum energy solution to the ISP.

Motivation for the research presented in this paper is provided in part by the possibility of optimally selecting the source region index of refraction distribution $n(\mathbf{r})$ to achieve some specified radiation pattern that would otherwise not be realistically possible (due to prohibitive values of required current level, or other engineering constraints) for a source embedded in free space. This possibility has attracted research from time to time in the antenna community, being of interest a variety of antenna-embedding materials or substrates, including plasmas [30], nonmagnetic dielectrics [31, 32, 33, 34, 35, 36], magneto-dielectrics [37, 38, 39, 40], and, more recently, double negative metamaterials, which are receiving much recent attention by a number of groups as antenna performance-enhancing substrates [41, 42, 43, 44, 45, 46, 47, 48, 49]. The envisaged property is miniaturization of antennas by controlling electric size (via larger wavenumber), but other effects are involved, particularly when metamaterials are used.

To arrive from first wave theoretic principles at a non–device specific understanding of the practical possibilities opened by antenna-embedding substrates, we emphasize in the present work the fundamental minimum energy source yielding a given radiation pattern, rather than particular devices, as has been the focus of the aforementioned presentations in this area. Also, the present treatment concerns the scalar inverse problem, which is a simplification of the full vector electromagnetic case. Rigorous treatment of the metamaterials, which are generally bi-anisotropic media, requires the full vector formulation [50, 51] and is left for future work. Thus in the present work attention is restricted to the pertinent properties associated with the index of refraction $n(\mathbf{r}) \geq 0$ of natural "positive" materials whose key aspects can be treated within the scalar formulation, but the general approach can be extended also to the vector case along the lines of, e.g., [52], where both source energy and reactive power constraints have been considered in the formulation of the inverse problem for sources embedded in free space.

The key observation is that, as outlined earlier and as shown in section 3 of this paper, in the free space case the minimum energy source energy increases exponentially with decreasing $x = k_0 a$ below a critical point that is determined by the fine detail that is desired in the radiation pattern. The question then is whether this limitation can be mitigated by embedding the source in a nonhomogeneous background medium. In effect we create a new "effective source" that consists of the actual physi-

cal source interacting with the nonhomogeneous background. In the simplest case we can consider a source embedded in a cavity with partially reflecting walls. This cavity will, of course, have a pronounced effect on the radiated field and, possibly, can aid in achieving desired properties of the radiation pattern.

In this paper we limit our attention, for the most part, to source regions characterized by a spherically symmetric index of refraction distribution $n(\mathbf{r}) = n(r)$, although many of our results can be generalized to sources embedded in cavities and nonspherically symmetric index distributions. The realizability of a given radiation pattern is investigated in some detail by examining the dependence of minimum source energy on the index of refraction profile of the source region. It is found that this energy depends critically on the (weighted) L^2 norm of the radial wave functions taken over the source region. This fact suggests that, by proper choice of the index of refraction profile $n(r)$, the energy can be minimized for any given radiation pattern; i.e., an index distribution can be selected that results in a source having minimum energy for a given prescribed radiation pattern.

The L^2 norms of the radial wave functions are found to be dependent on "resonant" properties of the source index distribution and are also related in a one-to-one fashion with the different angular modes of the radiation pattern. These facts suggest the interesting possibility of exploiting the "resonances" of the source index distribution to selectively control the shape and form of the radiation pattern. This possibility is briefly considered in the computer simulation study.

The final section of the paper treats the simple example of a source embedded in a homogeneous sphere whose constant index of refraction differs from that of the background medium. This is the simplest example of a spherically symmetric index of refraction distribution, and the scattering wave functions are well known (the so-called Mie scattering problem [28, 29]) and easily computed. The minimum energy source is computed for this case, and results from a computer simulation study that examines the dependence of the energy of the minimum energy source on the index of refraction of the source region are presented. The source energy depends in a nonlinear manner on the value of the source region refraction index. The examples considered reveal that, as desired, there are values of the index of refraction for which the improvement of the source-embedded case relative to the free space case is significant for the entire effectively radiating multipole spectrum pertinent to the same source region in free space.

2. Problem formulation. We introduce the scattering potential defined according to the equation

$$V(\mathbf{r}) = k_0^2[1 - n^2(\mathbf{r})]$$

and rewrite (1.1) in a form that we will use in the development to follow. In particular we find that

$$[\nabla^2 + k_0^2 - V(\mathbf{r})]U(\mathbf{r}) = -\rho(\mathbf{r}), \tag{2.1}$$

where here and in the remainder of the paper both the scattering potential V and the source ρ are assumed to vanish outside the source region τ (thus $n(\mathbf{r}) = 1$ for $\mathbf{r} \notin \tau$).

The outgoing wave solution to (2.1) is the unique field radiated by the source obeying Sommerfeld's radiation condition (e.g., see [53, Chapter 2], [54, Chapter 3]), which is the one having the asymptotic behavior of an outgoing spherical wave

$$U(r\mathbf{s}) = f(\mathbf{s})\frac{e^{ik_0 r}}{r} + O\left(\frac{1}{r^2}\right) \tag{2.2}$$

as $k_0 r \to \infty$ uniformly in the direction specified by the unit vector $\mathbf{s}$. In the above equation, the quantity $f(\mathbf{s})$ is the source's far field radiation pattern which is seen to correspond to the limit

$$f(\mathbf{s}) = \lim_{r\to\infty} \left[r e^{-ik_0 r} U(r\mathbf{s}) \right]. \tag{2.3}$$

In the following, we will bear in mind (2.2), (2.3) but will usually express the associated far field asymptotic behavior simply as

$$U(r\mathbf{s}) \sim f(\mathbf{s}) \frac{e^{ik_0 r}}{r} \quad \text{as } k_0 r \to \infty, \tag{2.4}$$

with the understanding that throughout the paper all the far field approximations hold to within $O(1/r^2)$. It is well known that the radiation pattern specified for all directions $\mathbf{s}$ uniquely determines the field U everywhere outside the source region τ [11]; i.e., knowledge of the radiated field everywhere outside τ is equivalent to knowledge of the radiation pattern $f(\mathbf{s})$ specified for all directions $\mathbf{s}$.

The ISP consists of determining a source distribution $\rho(\mathbf{r})$ that radiates a given field U for $\mathbf{r} \notin \tau$. Because the ISP requires that the field radiated by the source be specified only outside τ, the problem does not possess a unique solution, because of the possible presence of nonradiating sources [12, 13, 14] within τ. A nonradiating source generates a field that vanishes identically outside τ and hence can be added to any given solution to the ISP to yield a different solution. Also, because the radiation pattern uniquely determines the field everywhere outside τ, the ISP is equivalent to the problem of determining a source that generates a prescribed radiation pattern $f(\mathbf{s})$ for all observation directions $\mathbf{s}$.

Most treatments of the ISP cast the problem in terms of the radiation pattern but require only that the source generate the radiation pattern to within a given accuracy defined by the integral squared error

$$E = \int_{4\pi} d\Omega_s \, |\hat{f}(\mathbf{s}) - f(\mathbf{s})|^2, \tag{2.5}$$

where f is the prescribed radiation pattern and $\hat{f}$ the radiation pattern actually generated by the source, while $d\Omega_s$ denotes the solid-angular differential element. More precisely, the desired radiation pattern is approximated by a finite series of spherical harmonics $Y_l^m(\mathbf{s})$,

$$f(\mathbf{s}) \approx \hat{f}(\mathbf{s}) = \sum_{l=0}^{L} \sum_{m=-l}^{l} \alpha_{l,m} Y_l^m(\mathbf{s}), \tag{2.6}$$

and the source is required only to generate the approximate radiation pattern $\hat{f}$. Here we have used the unit vector $\mathbf{s}$ having polar angle θ and azimuthal angle ϕ to denote the θ, ϕ arguments of the spherical harmonics. Because the spherical harmonics are orthonormal and complete over the unit sphere, the approximated radiation pattern satisfies (2.5) with an error E given by

$$E = \sum_{l=L+1}^{\infty} \sum_{m=-l}^{l} |\alpha_{l,m}|^2,$$

where the expansion coefficients (multipole moments) $\alpha_{l,m}$, $l > L$, are the higher order (neglected) expansion coefficients of the ideal radiation pattern.

Besides requiring only that the source generate the radiation pattern within a finite error, most treatments of the ISP also require that the source minimize the source energy defined by (cf. (1.2))

$$\mathcal{E} = \int_\tau d^3r\, |\rho(\mathbf{r})|^2. \tag{2.7}$$

We will show (see also [3]) that minimizing the source energy leads to a unique solution of the ISP, namely, the *minimum energy source*, which we designate by ρ_{ME}. This solution has the distinct advantage of being the most efficient source that solves the ISP for a given scattering potential $V(\mathbf{r})$ (corresponding to a given background index of refraction $n(\mathbf{r})$). Since the minimum energy source and, hence, the minimum source energy depend on the scattering potential, an interesting question arises as to the dependence of the source energy on the background index distribution and, in particular, on which index distributions lead to lowest source energies. This question provides much of the motivation for studying the ISP in nonhomogeneous backgrounds, and the derived results shed light onto the possibility of designing very efficient sources (e.g., antennas) that are embedded in such backgrounds.

Our goal in this paper is to develop the formalism for solving the ISP as defined above and to evaluate the formalism in a set of computer simulations. We will first treat the case of a source embedded in free space, and then extend the free space theory to the general case of a source embedded in an inhomogeneous background medium. The free space case is important in that it provides a benchmark of performance as well as a frame of reference for the general theory.

3. Free space case. The minimum energy ISP as defined above has been solved within both the scalar wave formulation under consideration here [1, 2, 3, 4, 5, 7] and the electromagnetic wave formulation [6, 8, 9, 10] in the special case where the scattering potential $V(\mathbf{r})$ vanishes; i.e., when the source is embedded in free space. We will review the scalar wave free space case here, where, however, we will employ a solution methodology to find the minimum energy source somewhat different from the one adopted in earlier work. We will use this same procedure for the general case of nonvanishing scattering potentials later in the paper.

The outgoing wave solution to (2.1) is given in terms of an outgoing wave Green function $G(\mathbf{r}, \mathbf{r}')$ (obeying (2.1) for $\rho(\mathbf{r}) = -\delta(\mathbf{r} - \mathbf{r}')$ and an asymptotic condition of the form (2.4)) by the expression

$$U(\mathbf{r}) = -\int_\tau d^3r'\, \rho(\mathbf{r}')G(\mathbf{r}, \mathbf{r}'), \tag{3.1}$$

where (as indicated earlier) the source volume τ is a sphere of radius a centered at the origin. In the particular free space case where the scattering potential $V = 0$, the outgoing wave Green function G is given by

$$G(\mathbf{r}, \mathbf{r}') = -\frac{1}{4\pi}\frac{e^{ik_0|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r} - \mathbf{r}'|}, \tag{3.2}$$

from which it is easy to show that

$$G(r\mathbf{s}, \mathbf{r}') \sim -\frac{1}{4\pi}e^{-ik_0\mathbf{s}\cdot\mathbf{r}'}\frac{e^{ik_0r}}{r} \tag{3.3}$$

as $k_0r \to \infty$ in the direction $\mathbf{s}$ (which has the required form (2.4)). Using the above result, we conclude from (2.4) and (3.1) that the radiation pattern is given by

$$f(\mathbf{s}) = \frac{1}{4\pi}\int_\tau d^3r'\,\rho(\mathbf{r}')e^{-ik_0\mathbf{s}\cdot\mathbf{r}'}. \tag{3.4}$$

We can obtain an expansion of the radiation pattern in a series of spherical harmonics by using the well-known expansion

$$e^{-ik_0\mathbf{s}\cdot\mathbf{r}'} = 4\pi\sum_{l=0}^{\infty}\sum_{m=-l}^{l}(-i)^l j_l(k_0r')Y_l^{m*}(\hat{\mathbf{r}}')Y_l^m(\mathbf{s}), \tag{3.5}$$

where j_l denotes the spherical Bessel function of the first kind of order l, Y_l^m are the spherical harmonics of degree l and order m, $\hat{\mathbf{r}}'$ denotes the unit vector in the $\mathbf{r}'$ direction, and $*$ denotes the complex conjugate. Upon substituting (3.5) into (3.4), we find that

$$f(\mathbf{s}) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\alpha_{l,m}Y_l^m(\mathbf{s}), \tag{3.6}$$

where the expansion coefficients (multipole moments) $\alpha_{l,m}$ are given by

$$\begin{aligned}\alpha_{l,m} &= \int_{4\pi} d\Omega_s\, f(\mathbf{s})Y_l^{m*}(\mathbf{s})\\ &= (-i)^l\int_\tau d^3r'\,\rho(\mathbf{r}')j_l(k_0r')Y_l^{m*}(\hat{\mathbf{r}}').\end{aligned} \tag{3.7}$$

3.1. Minimum energy source. The minimum energy solution to the ISP is required to satisfy (3.7) for some given set of multipole moments $\alpha_{l,m}$, $l = 0, 1, \ldots, L$, and also to minimize the source energy defined according to (2.7). Computing the minimum energy source can be cast as a problem of constrained minimization, where the generalized Lagrangian is given by

$$\mathcal{L} = \mathcal{E} + \sum_{l=0}^{L}\sum_{m=-l}^{l}C_{l,m}\left[\alpha_{l,m}^* - i^l\int_\tau d^3r\,\rho^*(\mathbf{r})j_l(k_0r)Y_l^m(\hat{\mathbf{r}})\right] + \text{c.c.},$$

where $\mathcal{E}$ is the source energy defined in (2.7), c.c. stands for the complex conjugate of the second term on the right-hand side (r.h.s.) of the equation, and the $C_{l,m}$ are a set of Lagrange multipliers to be determined. On expressing the source energy in terms of ρ and ρ^* and taking the first variation of the above Lagrangian, we obtain

$$\delta\mathcal{L} = \int_\tau d^3r\,\delta\rho^*(\mathbf{r})\left[\rho(\mathbf{r}) - \sum_{l=0}^{L}\sum_{m=-l}^{l}C_{l,m}i^l j_l(k_0r)Y_l^m(\hat{\mathbf{r}})\right] + \text{c.c.},$$

which, when set equal to zero, yields the solution

$$\rho_{ME}(\mathbf{r}) = \begin{cases}\sum_{l=0}^{L}\sum_{m=-l}^{l}C_{l,m}i^l j_l(k_0r)Y_l^m(\hat{\mathbf{r}}) & \text{if } r < a,\\ 0 & \text{if } r > a.\end{cases}$$

The Lagrange multipliers $C_{l,m}$ are determined from the condition that the source generate the multipole moments according to (3.7). We find that

$$C_{l,m} = \frac{\alpha_{l,m}}{\sigma_l^2}, \tag{3.8}$$

where

$$\sigma_l^2 = \int_0^a r^2 dr\, j_l^2(k_0 r). \tag{3.9}$$

On making use of the above expression for the Lagrange multipliers, we finally conclude that the minimum energy solution to the free space ISP is given by

$$\rho_{ME}(\mathbf{r}) = \begin{cases} \sum_{l=0}^{L} \sum_{m=-l}^{l} i^l \frac{\alpha_{l,m}}{\sigma_l^2} j_l(k_0 r) Y_l^m(\hat{\mathbf{r}}) & \text{if } r < a, \\ 0 & \text{if } r > a. \end{cases} \tag{3.10}$$

3.2. Source energy. The source energy $\mathcal{E}$ is readily computed using the minimum energy source given in (3.10). We find from (2.7) that

$$\mathcal{E}_{ME} = \int_\tau d^3 r\, |\rho_{ME}(\mathbf{r})|^2 = \sum_{l=0}^{L} \sum_{m=-l}^{l} \frac{|\alpha_{l,m}|^2}{\sigma_l^2}, \tag{3.11}$$

where we have added the subscript "ME" to denote the energy of the minimum energy source. Now, it is easy to show that the quantities σ_l^2 depend critically on the product $k_0 a$ of the free space wavenumber with the source radius a. In particular, these quantities can be shown to be given by

$$\sigma_l^2 = \int_0^a r^2 dr\, |j_l(k_0 r)|^2 = \frac{a^3}{2}[j_l{}^2(k_0 a) - j_{l-1}(k_0 a) j_{l+1}(k_0 a)] \tag{3.12}$$

and to decrease exponentially to zero for $l > k_0 a$. It then follows that the largest value L of the index l allowed in the approximation (2.6) is $L = k_0 a$ if we want to maintain low source energy. Values of $L \gg k_0 a$ will lead to extremely high source energy and unstable source distributions.

To illustrate the remarks made above concerning the behavior of the quantities σ_l^2 on the index l and the product $k_0 a$ we show in Figure 3.1 semilog plots of σ_l^2 as a function the index l for various values of $x = k_0 a$. It is seen from these plots that these quantities decay exponentially to zero for $l \gg x$, so that at wavenumber k_0 a source of radius a can only efficiently radiate a radiation pattern whose maximum l value is $L = k_0 a$. Similar behavior is exhibited in Figure 3.2, which shows semilog plots of $\sigma_l^2(x)$ as a function of x for values of $l = 10$, 20, and 30.

We computed the energy of the minimum energy source for a model radiation pattern $f(\theta)$ having multipole coefficients $\alpha_{l,m}$ given by

$$\alpha_{l,m} = \begin{cases} \frac{1}{\sqrt{L+1}} & \text{if } m = 0 \text{ and } l \le L, \\ 0 & \text{otherwise.} \end{cases} \tag{3.13}$$

This radiation pattern is circularly symmetric about the z axis (is independent of ϕ since the nonzero multipoles correspond to $m = 0$), has an effective beam width inversely related to the cut-off value L, and has unit energy; i.e.,

$$\int_{4\pi} \sin\theta d\theta d\phi\, |f(\theta)|^2 = \sum_{l=0}^{L} \frac{1}{L+1} = 1.$$

We show plots of the model radiation pattern as a function of angle θ in Figure 3.3 for values of the parameter L equal to $L = 10$, $L = 20$, and $L = 30$. It is clear

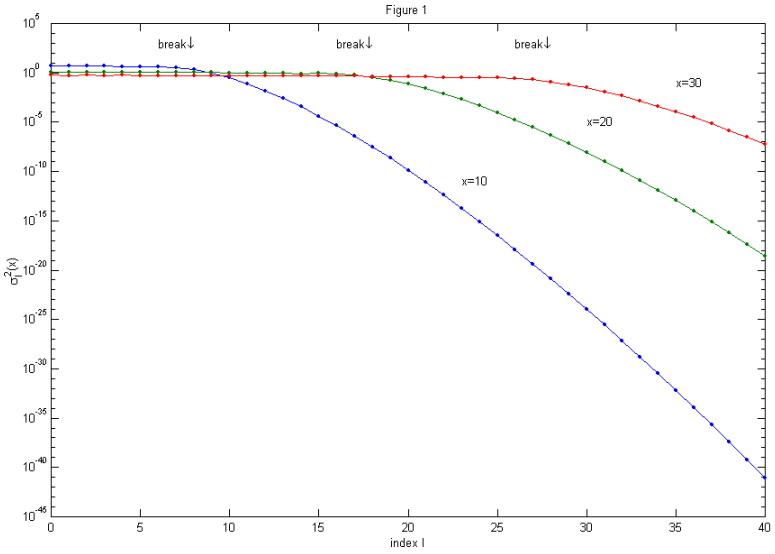


FIG. 3.1. *Behavior of σ_l^2 as a function of index l for three different values of $x = k_0 a$ equal to* 10, 20, *and* 30. *Plots indicate an exponential decay of these quantities for $l \gg x$. The break points are seen to occur when $x \approx l$.*

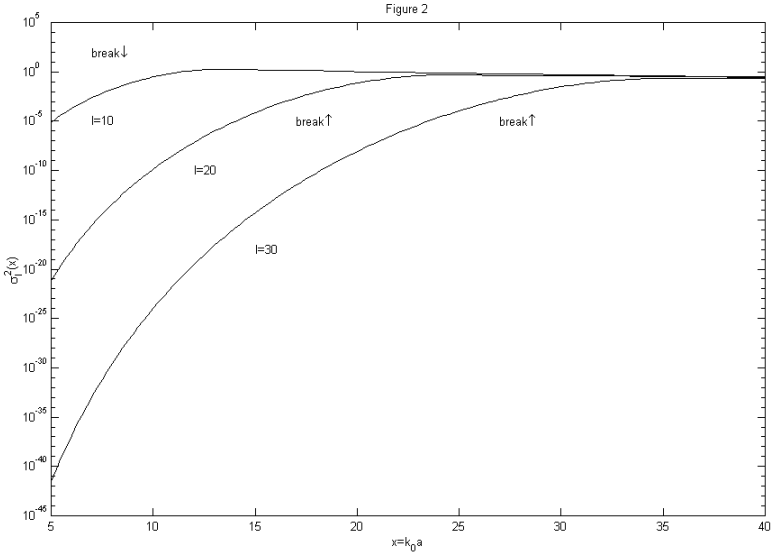


FIG. 3.2. *σ_l^2 as a function of $x = k_0 a$ for values of $l_0 = 10$, $l_0 = 20$, and $l_0 = 30$. Plots indicate an exponential growth of these quantities for $x \ll l_0$. The break points are seen to occur when $x \approx l$.*

from these plots that the larger the value of L, the narrower the radiation pattern and, hence, the higher the directivity of the source. Using the coefficients given in (3.13), we computed the source energy using (3.11) with the σ_l^2 given by (3.12) and

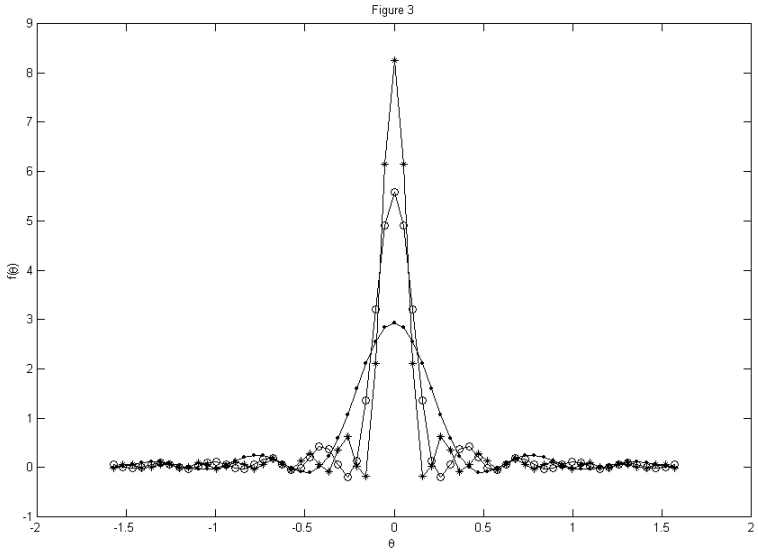


FIG. 3.3. *Plots of the model radiation pattern for* $L = 10$ *(.),* $L = 20$ *(o), and* $L = 30$ *(*).*

with three different L values of $L = 10, 20, 30$. It was found that, as expected, the source energy becomes extremely large if we try to achieve an L value that exceeds the critical value $L = k_0 a$. This is, of course, due to the fact that the quantities σ_l^2 become extremely small when $k_0 a \ll l$, as is indicated in Figure 3.2.

4. Nonhomogeneous backgrounds. The outgoing wave solution of (2.1) for a nonhomogeneous index distribution $n(\mathbf{r})$ of support τ, which behaves as in (2.2), (2.3), (2.4), can be expressed in terms of the outgoing wave Green function for the background medium via (3.1) where, however, the Green function is no longer the free space Green function defined in (3.2), but instead is the total outgoing wave Green function corresponding to the total medium comprised of free space plus the nonhomogeneous index distribution $n(\mathbf{r})$ or its associated scattering potential $V(\mathbf{r})$. In the ISP under consideration in this paper, the field is given outside the source region τ only; that is, the field for a source of support τ is given by (3.1), where $\mathbf{r}' \in \tau$ and $\mathbf{r} \notin \tau$. This particular situation enables us to formulate the forward mapping pertinent here (from a masked source that is confined within τ to an exterior field that is prescribed for $\mathbf{r} \notin \tau$ only) via a conceptually simple and insightful approach which borrows from standard scattering theory and the reciprocity property of the outgoing wave Green function, which can be readily shown for the formally self-adjoint partial differential operator $[\nabla^2 + k_0^2 - V(\mathbf{r})]$ using standard Green function theory (e.g., see [55, Chapters 9 and 10]). Two equivalent versions of the general methodology are outlined next. The main objective is to generalize the statement made in connection with the free space case in the far field mapping equation (3.4) for the more general case of nonhomogeneous backgrounds. The generalization of the derived expression for the far field mapping will be applied to the particular spherically symmetric and piecewise constant background cases later in the paper.

The starting point is provided by the familiar *Lippmann–Schwinger integral equation* (e.g., see [56, eqs. (8.4), (8.5), (10.12), (10.13)], [27, pp. 178–179, 263–265], [53,

p. 5], [57, pp. 60–61]), which in the present formulation and notation yields

$$G(\mathbf{r},\mathbf{r}') = G_0(\mathbf{r},\mathbf{r}') + \int_\tau d^3r''\, G_0(\mathbf{r},\mathbf{r}'')V(\mathbf{r}'')G(\mathbf{r}'',\mathbf{r}'), \tag{4.1}$$

where G denotes the total outgoing wave Green function of the total medium comprised of free space plus the generally nontrivial scattering potential V, and G_0 denotes the free space Green function defined in (3.2) (corresponding to $V = 0$), in particular,

$$G_0(\mathbf{r},\mathbf{r}') = -\frac{1}{4\pi}\frac{e^{ik_0|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|},$$

where (see also (3.3), where, again, the G in (3.3) corresponds to the G_0 of the present section)

$$G_0(r\mathbf{s},\mathbf{r}') = -\frac{1}{4\pi}e^{-ik_0\mathbf{s}\cdot\mathbf{r}'}\frac{e^{ik_0r}}{r} + O\left(\frac{1}{r^2}\right) \quad \text{as } k_0r \to \infty, \tag{4.2}$$

in the direction of the unit vector $\mathbf{s}$, so that from the discussion in (2.2), (2.3), and (2.4) (which holds for a general source ρ) one finds that the far field radiation pattern of a point source $-\delta(\mathbf{r}-\mathbf{r}')$ in free space is $-\frac{1}{4\pi}e^{-ik_0\mathbf{s}\cdot\mathbf{r}'}$. We shall recall this basic result later.

Due to reciprocity, the result (4.1) can be rewritten also as

$$G(\mathbf{r},\mathbf{r}') = G_0(\mathbf{r},\mathbf{r}') + \int_\tau d^3r''\, G(\mathbf{r},\mathbf{r}'')V(\mathbf{r}'')G_0(\mathbf{r}'',\mathbf{r}'). \tag{4.3}$$

We will borrow from both (4.1) and (4.3) in the following.

By making use of the asymptotic result (4.2), it is not difficult to show from (4.1) or its equivalent, (4.3), that the Green function G behaves asymptotically as

$$G(r\mathbf{s},\mathbf{r}') = -\frac{1}{4\pi}\psi^+(\mathbf{r}';-k_0\mathbf{s})\frac{e^{ik_0r}}{r} + O\left(\frac{1}{r^2}\right) \tag{4.4}$$

as $k_0r \to \infty$, where we have introduced the quantity $\psi^+(\mathbf{r};-k_0\mathbf{s})$ defined by

$$\begin{aligned}\psi^+(\mathbf{r};-k_0\mathbf{s}) &= e^{-ik_0\mathbf{s}\cdot\mathbf{r}} + \int_\tau d^3r'\, e^{-ik_0\mathbf{s}\cdot\mathbf{r}'}V(\mathbf{r}')G(\mathbf{r}',\mathbf{r}) \\ &= e^{-ik_0\mathbf{s}\cdot\mathbf{r}} + \int_\tau d^3r'\, G(\mathbf{r},\mathbf{r}')V(\mathbf{r}')e^{-ik_0\mathbf{s}\cdot\mathbf{r}'}.\end{aligned} \tag{4.5}$$

Note from (4.4) that the Green function $G(\mathbf{r},\mathbf{r}')$ does in fact behave as $k_0r \to \infty$, as we have required in (2.2), (2.3), and (2.4). From the same equations one also notes that the far field radiation pattern $f(\mathbf{s})$ in (2.2), (2.3), (2.4), corresponding to the field U obeying (2.1), applies to the *general* source ρ, while, on the other hand, the far field radiation pattern $-\frac{1}{4\pi}\psi^+(\mathbf{r}';-k_0\mathbf{s})$ in (4.4) corresponds to that of the total field produced by the *particular* Dirac-delta point source at $\mathbf{r}'$ in the same medium. In other words, the quantity $-\frac{1}{4\pi}\psi^+(\mathbf{r}';-k_0\mathbf{s})$ is the far field radiation pattern for the particular case of a point source at $\mathbf{r}'$. This quantity is seen from (4.5) to consist of the sum of two terms: the first term, $-\frac{1}{4\pi}e^{-ik_0\mathbf{s}\cdot\mathbf{r}'}$, exists even if $V = 0$ and is the far field radiation pattern for the point source in free space. (It will account for an incident field in the discussion to follow.) This is, in fact, what we have discussed

before in (4.2). On the other hand, the second term, the integral in (4.5), corresponds to a scattered field contribution, as we shall elaborate further next.

The quantity $\psi^+(\mathbf{r}; -k_0\mathbf{s})$ as defined by the formulation above (i.e., (4.4), (4.5)) is customarily termed a *scattering wave function* (e.g., see [56, pp. 3, 4, 164, 172]). This scattering wave function is the unique total (incident plus scattered) field for the scattering potential $V(\mathbf{r})$ under excitation by the incident plane wave $e^{-ik_0\mathbf{s}\cdot\mathbf{r}}$ in the direction defined by the unit vector $-\mathbf{s}$, under Sommerfeld's radiation condition for the scattered field, which translates into the requirement that the respective scattered field, that is, the integral term in (4.5), behaves like an outgoing wave at infinity. Thus the scattering wave function $\psi^+(\mathbf{r}; -k_0\mathbf{s})$ is the solution to the homogeneous Helmholtz equation

$$[\nabla^2 + k_0^2 - V(\mathbf{r})]\psi^+(\mathbf{r}; -k_0\mathbf{s}) = 0, \tag{4.6}$$

which obeys an asymptotic condition of the form

$$\psi^+(r\hat{\mathbf{r}}; -k_0\mathbf{s}) \sim e^{-ik_0\mathbf{s}\cdot\mathbf{r}} + g(\hat{\mathbf{r}}; -k_0\mathbf{s})\frac{e^{ik_0 r}}{r} \tag{4.7}$$

as $k_0 r \to \infty$, where g is the so-called scattering amplitude associated with the scattering potential V whose role in scattering problems is similar to that of the source radiation pattern f (of (2.4)) in radiation problems (e.g., see [56, eq. (10.19)], [58]). Note from (4.4), (4.5) that

$$\psi^+(r\hat{\mathbf{r}}; -k_0\mathbf{s}) \sim e^{-ik_0\mathbf{s}\cdot\mathbf{r}} - \frac{1}{4\pi}\frac{e^{ik_0 r}}{r}\int_\tau d^3r'\, \psi^+(\mathbf{r}'; -k_0\mathbf{s})V(\mathbf{r}')e^{-ik_0\mathbf{s}\cdot\mathbf{r}'} \quad \text{as } k_0 r \to \infty, \tag{4.8}$$

so that from (4.7) the scattering amplitude

$$g(\hat{\mathbf{r}}; -k_0\mathbf{s}) = -\frac{1}{4\pi}\int_\tau d^3r'\, \psi^+(\mathbf{r}'; -k_0\mathbf{s})V(\mathbf{r}')e^{-ik_0\mathbf{s}\cdot\mathbf{r}'}. \tag{4.9}$$

Thus the scattering wave function $\psi^+(\mathbf{r}; -k_0\mathbf{s})$ corresponds to the total (incident plus scattered) field that results when an incident plane wave propagating in the $-\mathbf{s}$ direction scatters off the inhomogeneous index of refraction distribution $n(\mathbf{r})$. Note that in the limit when the scattering potential vanishes this scattering wave function simply reduces to the incident plane wave.

Now, by substituting the asymptotic result (4.4) into (2.2), (2.3), (2.4), and (3.1) we find that the radiation pattern $f(\mathbf{s})$ of a general source ρ embedded in the nonhomogeneous background characterized by index of refraction $n(\mathbf{r})$ or, equivalently, scattering potential $V(\mathbf{r})$, is given by

$$f(\mathbf{s}) = \frac{1}{4\pi}\int_\tau d^3r'\, \rho(\mathbf{r}')\psi^+(\mathbf{r}'; -k_0\mathbf{s}), \tag{4.10}$$

which is simply the free space result (3.3) with the plane wave $\exp(-ik_0\mathbf{s}\cdot\mathbf{r}')$ replaced by the scattering wave function $\psi^+(\mathbf{r}'; -k_0\mathbf{s})$. For a given V, this scattering wave function can be obtained by solving the scattering problem posed by (4.6), (4.7). The ISP for given nonhomogeneous background media then reduces to determining a source ρ that satisfies (4.10) for all observation directions $\mathbf{s}$, and where the scattering wave functions ψ^+ are to be determined a priori for the pertinent scattering potential V by addressing the scattering problem in (4.6), (4.7). Clearly, in the special case

when $V = 0$ the scattering wave function reduces to the plane wave (refer to (4.5)), and (4.10) reduces to the free space result (3.4), as expected.

Prior to engaging in the particular cases of spherically symmetric and piecewise constant backgrounds, we wish to outline an alternative description of the formulation above based on reciprocity; that is, for the total Green function in the total medium characterized by the scattering potential V, $G(\mathbf{r}, \mathbf{r}') = G(\mathbf{r}', \mathbf{r})$. Such a description is very useful in elucidating the connection between radiation and scattering problems (e.g., see [59]), and may thus facilitate application of the present ISP research to other areas, such as scattering and inverse scattering problems. Without loss of generality, in the rest of this paragraph the point $\mathbf{r}'$ will be taken to lie in the spherical volume τ of radius a centered about the origin (the source region). Also, take $\mathbf{r}$ to be a point in a spherical surface of radius $R > a$ centered about the origin. The idea is that the problem of computing the far field produced at the field point $R\mathbf{s}$ (for large $k_0 R$, and where $\mathbf{s}$ is a unit vector) by a point source at the point $\mathbf{r}'$ in the source region τ is, due to reciprocity considerations, equivalent to the problem of computing the (near) field produced at point $\mathbf{r}'$ due to a far zone point source at $R\mathbf{s}$, in particular, $G(R\mathbf{s}, \mathbf{r}') = G(\mathbf{r}', R\mathbf{s})$. Conveniently, the latter problem essentially reduces to the familiar scattering problem under plane wave excitation, and this is the basis of the preceding formulation as well as of the complementary analysis to be given next. The field generated by a given point source at $\mathbf{r}' \in \tau$ at the field point $\mathbf{r} = R\mathbf{s}$ in the far field direction defined by the unit vector $\mathbf{s}$ obeys, from (2.2), (2.3), (2.4), the asymptotic form

$$G(R\mathbf{s}, \mathbf{r}') = f(\mathbf{s}; \mathbf{r}') \frac{e^{ik_0 R}}{R} + O\left(\frac{1}{R^2}\right) \tag{4.11}$$

as $k_0 R \to \infty$, where the respective radiation pattern $f(\mathbf{s}; \mathbf{r}')$ of the total field radiated by the point source at $\mathbf{r}'$ depends on $\mathbf{r}'$ in a way to be clarified in the following (again, note that the radiation pattern in (4.11) is, apart from a factor $-\frac{1}{4\pi}$, the scattering wave function $\psi^+(\mathbf{r}'; -k_0\mathbf{s})$ of the preceding development). On the other hand, the total field produced at the point $\mathbf{r}' \in \tau$ due to a point source at $R\mathbf{s}$ can be decomposed into the sum of an incident field, corresponding to the radiation in free space, that is, the free space Green function component $G_0(\mathbf{r}', R\mathbf{s})$, plus a scattered field corresponding to scattering of that incident field by the medium characterized by scattering potential V. This can be expressed formally by borrowing from (4.1), (4.3), in particular,

$$\begin{aligned} G(\mathbf{r}', R\mathbf{s}) &= G_0(\mathbf{r}', R\mathbf{s}) + \int_\tau d^3 r\, G_0(\mathbf{r}', \mathbf{r}) V(\mathbf{r}) G(\mathbf{r}, R\mathbf{s}) \\ &= G_0(\mathbf{r}', R\mathbf{s}) + \int_\tau d^3 r\, G(\mathbf{r}', \mathbf{r}) V(\mathbf{r}) G_0(\mathbf{r}, R\mathbf{s}), \end{aligned} \tag{4.12}$$

where the integrals define the scattered field component, and where from the outgoing wave nature of both G_0 and G,

$$G(\mathbf{r}', R\mathbf{s}) = -\frac{1}{4\pi} e^{-ik_0 \mathbf{s}\cdot\mathbf{r}'} \frac{e^{ik_0 R}}{R} + \frac{e^{ik_0 R}}{R} \int_\tau d^3 r\, G_0(\mathbf{r}', \mathbf{r}) V(\mathbf{r}) f(\mathbf{s}; \mathbf{r}) + O\left(\frac{1}{R^2}\right) \tag{4.13}$$

as $k_0 R \to \infty$ (where we have used (4.11) and $G(\mathbf{r}', R\mathbf{s}) = G(R\mathbf{s}, \mathbf{r}')$), or, equivalently (from the second of the equations in (4.12)), as

$$G(\mathbf{r}', R\mathbf{s}) = -\frac{1}{4\pi} e^{-ik_0 \mathbf{s}\cdot\mathbf{r}'} \frac{e^{ik_0 R}}{R} - \frac{1}{4\pi} \frac{e^{ik_0 R}}{R} \int_\tau d^3 r\, G(\mathbf{r}', \mathbf{r}) V(\mathbf{r}) e^{-ik_0 \mathbf{s}\cdot\mathbf{r}} + O\left(\frac{1}{R^2}\right) \tag{4.14}$$

as $k_0 R \to \infty$; furthermore, using $G(\mathbf{r}', R\mathbf{s}) = G(R\mathbf{s}, \mathbf{r}')$, one recovers from these developments (4.11), where $f(\mathbf{s}; \mathbf{r}') = -\frac{1}{4\pi}\psi^+(\mathbf{r}'; -k_0\mathbf{s})$ obeys

(4.15)
$$f(\mathbf{s}; \mathbf{r}') = -\frac{1}{4\pi}\psi^+(\mathbf{r}'; -k_0\mathbf{s}) = -\frac{1}{4\pi}\left[e^{-ik_0\mathbf{s}\cdot\mathbf{r}'} + \int_\tau d^3r\, G_0(\mathbf{r}', \mathbf{r})V(\mathbf{r})\psi^+(\mathbf{r}; -k_0\mathbf{s})\right]$$

or, equivalently,

$$f(\mathbf{s}; \mathbf{r}') = -\frac{1}{4\pi}\psi^+(\mathbf{r}'; -k_0\mathbf{s}) = -\frac{1}{4\pi}\left[e^{-ik_0\mathbf{s}\cdot\mathbf{r}'} + \int_\tau d^3r\, G(\mathbf{r}', \mathbf{r})V(\mathbf{r})e^{-ik_0\mathbf{s}\cdot\mathbf{r}}\right], \tag{4.16}$$

which agrees with the previous formulation ((4.4), (4.5), (4.6), (4.7)). It follows that, as explained in connection with those results, the scattering wave function $\psi^+(\mathbf{r}', -k_0\mathbf{s})$ is the total (incident plus scattered) field at $\mathbf{r}'$ due to the interaction of an incident plane wave in the direction $-\mathbf{s}$ with the scattering potential V. Finally, an alternative way of arriving at these results is via the so-called mixed reciprocity relation [57, pp. 61–62, particularly eq. (2.2.6); see also p. 42], which states that the value at $\mathbf{s}$ of the far field radiation pattern corresponding to the *scattered field* component of the field generated in the nonhomogeneous medium due to a point source excitation at $\mathbf{r}'$ is, apart from a multiplicative factor $(-\frac{1}{4\pi})$, equal to the value at $\mathbf{r}'$ of the field that is scattered by the same medium due to an incident plane wave traveling in the direction $-\mathbf{s}$. In the notation of this paper, the far field radiation pattern of the scattered field component of the field generated in the nonhomogeneous medium due to a point source at $\mathbf{r}'$ is, as has been discussed before (see (4.1), (4.2), (4.3), (4.4), (4.5)), precisely the integration term in (4.5), (4.15), and (4.16), which is the field scattered by a plane wave traveling in the direction $-\mathbf{s}$, as expected. Thus our findings are consistent with this standard relation from the literature [57]. Yet, we must point out a difference in notation between [57] and the present paper. In our notation (and without loss of generality), the Green function or fundamental solution applicable to free space,

$$G_0(\mathbf{r}, \mathbf{r}') = -\frac{1}{4\pi}\frac{e^{ik_0|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|},$$

has a negative sign since we define the Green function (or fundamental solution) for a point source $-\delta(\mathbf{r}-\mathbf{r}')$, while [57] considers the fundamental solution for a point source $\delta(\mathbf{r}-\mathbf{r}')$. (This is explained in [57, p. 8]; see also [53, p. 16], [54, Chapter 3].) Our entire formulation incorporates with no loss of generality this particular choice, as is obvious in our form of the Green function integral (3.1). Thus the result (2.2.6) in [57, p. 61] involves the factor $\gamma_3 = \frac{1}{4\pi}$ (refer to [57, p. 42]), which is equivalent, in the present usage for the fundamental solution (with the added negative sign), to our factor $-\frac{1}{4\pi}$, and this completes the picture (the two theories yield the same final result, as desired). Let us consider special cases next.

4.1. Spherically symmetric backgrounds. In the remainder of the paper we will restrict our attention to the case of spherically symmetric index distributions $n(\mathbf{r}) = n(r)$. The scattering wave functions then satisfy (cf. (4.6))

$$[\nabla^2 + k_0^2 - V(r)]\psi^+(\mathbf{r}; -k_0\mathbf{s}) = 0, \tag{4.17}$$

where the eigenfunctions ψ^+ are required to satisfy the boundary condition (4.7). Because of the spherical symmetry of the scattering potential V, the wavefield ψ^+

can depend only on the magnitude r of the field point vector $\mathbf{r}$ and the polar angle γ formed between the direction of propagation $-\mathbf{s}$ of the incident plane wave and $\mathbf{r}$. If we then take the incident wave direction to be the positive z axis and express the Helmholtz operator in spherical polar coordinates, (4.17) can be written in the form

$$\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)+\frac{1}{r^2\sin\gamma}\frac{\partial}{\partial\gamma}\left(\sin\gamma\frac{\partial}{\partial\gamma}\right)-V(r)+k_0^2\right]\psi^+(r,\gamma)=0, \tag{4.18}$$

where γ is the polar angle formed between the positive z axis and the field point vector $\mathbf{r}$ and where we have used the fact that the field must be independent of the azimuthal angle ϕ. The boundary condition (4.7) becomes

$$\psi^+(r,\gamma)\sim e^{ik_0r\cos\gamma}+g(\gamma)\frac{e^{ik_0r}}{r}, \tag{4.19}$$

where we have set $-k_0\mathbf{s}\cdot\mathbf{r}=k_0z=k_0r\cos\gamma$ and where $g(\gamma)$ is the scattering amplitude.

We can expand the scattering wave function $\psi^+(r,\gamma)$, the incident plane wave $\exp(ik_0r\cos\gamma)$, and the scattering amplitude $g(\gamma)$ into a series of Legendre polynomials as follows [55, 60]:

$$\begin{aligned}\psi^+(r,\gamma)&=\sum_l i^l(2l+1)\psi_l(r)P_l(\cos\gamma),\\ e^{ik_0r\cos\gamma}&=\sum_l i^l(2l+1)j_l(k_0r)P_l(\cos\gamma),\\ g(\gamma)&=\sum_l i^l(2l+1)A_lP_l(\cos\gamma),\end{aligned}$$

where we have introduced the factors $i^l(2l+1)$ into the expansions for the scattering wave function and the scattering amplitude for later notational convenience. In these equations j_l is the spherical Bessel function of the first kind of order l, and the A_l are expansion coefficients of the scattering amplitude that depend on the specific form of the scattering potential $V(r)$. On substituting the first of these equations into (4.18), we find that the radially dependent coefficients $\psi_l(r)$ satisfy the equation

$$\left[\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d}{dr}\right)-\frac{l(l+1)}{r^2}-V(r)+k_0^2\right]\psi_l(r)=0, \tag{4.20}$$

where we have used the fact that

$$\frac{1}{\sin\gamma}\frac{\partial}{\partial\gamma}\left(\sin\gamma\frac{\partial}{\partial\gamma}\right)P_l(\cos\gamma)=-l(l+1)P_l(\cos\gamma).$$

The asymptotic behavior of the radial functions $\psi_l(r)$ is obtained by substituting the expansions for the scattering wave function, the incident plane wave, and the scattering amplitude into (4.19). We find that

$$\psi_l(r)\sim j_l(k_0r)+A_l\frac{e^{ik_0r}}{r}. \tag{4.21}$$

Besides satisfying the boundary condition (4.21), we also require that the radial functions be everywhere continuous with continuous first derivatives.

Once the radial functions $\psi_l(r)$ are computed, the scattering wave function $\psi^+(r,\gamma)$ corresponding to an incident plane wave propagating along the z axis is given by the expansion in Legendre polynomials above. However, the defining equation for the radiation pattern (4.10) requires that we have the scattering wave functions for all directions $-\mathbf{s}$ of the incident plane wave. This can be easily accomplished by using the addition theorem for spherical harmonics [55, 60],

$$P_l(\cos\gamma) = \frac{4\pi}{2l+1}\sum_{m=-l}^{l}(-1)^l Y_l^{m*}(\hat{\mathbf{r}})Y_l^m(\mathbf{s}),$$

where now γ is the angle formed between the *arbitrary* incident wave direction $-\mathbf{s}$ and the field direction $\hat{\mathbf{r}}$. On using the addition theorem, we obtain

$$\psi^+(\mathbf{r};-k_0\mathbf{s}) = 4\pi\sum_{l=0}^{\infty}\sum_{m=-l}^{l}(-i)^l\psi_l(r)Y_l^{m*}(\hat{\mathbf{r}})Y_l^m(\mathbf{s}), \tag{4.22}$$

which is the generalization of (3.5) to spherically symmetric nonhomogeneous index of refraction distributions.

4.2. Minimum energy source. Upon substituting the expansion (4.22) into (4.10), we obtain (3.6) where, however, the multipole moments are now given by

$$\begin{aligned}\alpha_{l,m} &= \int_{4\pi} d\Omega_s\, f(\mathbf{s})Y_l^{m*}(\mathbf{s})\\ &= (-i)^l\int_\tau d^3r\,\rho(\mathbf{r})\psi_l(r)Y_l^{m*}(\hat{\mathbf{r}}),\end{aligned} \tag{4.23}$$

which is the generalization of (3.7) to the case where the source is embedded in a nonhomogeneous but spherically symmetric index of refraction profile. The generalized equation (4.23) is seen to result from (3.7) under the replacement of the spherical Bessel functions j_l by the radial functions ψ_l. The minimum energy solution to the ISP is required to satisfy (4.23) for some given set of multipole moments $\alpha_{l,m}$, $l = 0, 1, \ldots, L$, and also to minimize the source energy defined according to (2.7).

As in the free space case, the problem of computing the minimum energy source can be cast as one of constrained minimization, where the generalized Lagrangian is now given by

$$\mathcal{L} = \mathcal{E} + \sum_{l=0}^{L}\sum_{m=-l}^{l} C_{l,m}\left[\alpha_{l,m}^* - i^l\int_\tau d^3r\,\rho^*(\mathbf{r})\psi_l^*(r)Y_l^m(\hat{\mathbf{r}})\right] + \text{c.c.},$$

where, as before, $\mathcal{E}$ is the source energy defined in (2.7), c.c. stands for the complex conjugate of the second term on the r.h.s. of the equation, and the $C_{l,m}$ are a set of Lagrange multipliers to be determined. On expressing the source energy in terms of ρ and ρ^* and taking the first variation of the above Lagrangian, we obtain

$$\delta\mathcal{L} = \int_\tau d^3r\,\delta\rho^*(\mathbf{r})\left[\rho(\mathbf{r}) - \sum_{l=0}^{L}\sum_{m=-l}^{l} C_{l,m}i^l\psi_l^*(r)Y_l^m(\hat{\mathbf{r}})\right] + \text{c.c.},$$

which, when set equal to zero, yields the solution

$$\rho_{ME}(\mathbf{r}) = \begin{cases} \sum_{l=0}^{L}\sum_{m=-l}^{l} C_{l,m}i^l\psi_l^*(r)Y_l^m(\hat{\mathbf{r}}) & \text{if } r < a,\\ 0 & \text{if } r > a.\end{cases}$$

The Lagrange multipliers $C_{l,m}$ are determined from the condition that the source generate the multipole moments according to (4.23). We find that

$$C_{l,m} = \frac{\alpha_{l,m}}{\Sigma_l^2}, \tag{4.24}$$

where

$$\Sigma_l^2 = \int_0^a r^2 dr\, |\psi_l(r)|^2. \tag{4.25}$$

On making use of the above expression for the Lagrange multipliers, we finally conclude that the minimum energy solution to the ISP for spherically symmetric background index distributions is given by

$$\rho_{ME}(\mathbf{r}) = \begin{cases} \sum_{l=0}^{L} \sum_{m=-l}^{l} (i)^l \frac{\alpha_{l,m}}{\Sigma_l^2} \psi_l^*(r) Y_l^m(\hat{\mathbf{r}}) & \text{if } r < a, \\ 0 & \text{if } r > a. \end{cases} \tag{4.26}$$

The source energy is found to be given by the free space formula (3.11), where, however, the σ_l^2 are replaced by the Σ_l^2 defined in (4.25). As in the free space case, the source energy is seen to depend inversely on the Σ_l^2. Although these quantities are strictly positive they can become extremely small, leading to extremely high source energy and associated instability in the minimum energy source. Thus, it is of interest to maximize these quantities especially for large values of the index l, which is associated with fine detail (high resolution) in the radiation pattern.

The energy of the minimum energy source is obtained by substituting (4.26) into the source energy definition given in (1.2). We obtain the same expression as was obtained in the free space case (3.11), where, however, the σ_l^2 are replaced by the Σ_l^2. It is clear that the source energy is minimized by maximizing the Σ_l^2, which, in turn, is equivalent to maximizing the weighted L^2 norm of the radial functions ψ_l over the interval $[0, a]$. Since the radial functions ψ_l are solutions to a Sturm–Liouville problem, the energy minimization problem reduces to finding scattering potentials $V(r)$ whose corresponding Sturm–Liouville problem has solutions with maximum norm over this interval. This problem, although simple to state, appears to be nontrivial, and the authors offer no simple recipe for computing optimum potentials at this time. However, in the following section we will treat a simple class of potentials that illustrates the dependence of source energy on selection of V.

5. Piecewise constant backgrounds. In this section we consider the special case where the scattering potential $V(r)$ is constant throughout the source region; i.e.,

$$V(r) = \begin{cases} k_0^2 - k^2 & \text{if } r < a, \\ 0 & \text{if } r > a. \end{cases} \tag{5.1}$$

This is certainly a spherically symmetric scattering potential, so that the scattering wave functions can be expanded in the form of (4.22), where the radial functions $\psi_l(r)$ satisfy (4.20) with $V(r)$ given in (5.1) above. Thus we find that

$$\left[\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d}{dr}\right) - \frac{l(l+1)}{r^2} + k^2\right]\psi_l(r) = 0 \quad \text{if } 0 < r < a,$$
$$\left[\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d}{dr}\right) - \frac{l(l+1)}{r^2} + k_0^2\right]\psi_l(r) = 0 \quad \text{if } a < r < \infty,$$

together with the boundary condition from (4.21):

$$\psi_l(r) \sim j_l(k_0 r) + A_l \frac{e^{ik_0 r}}{r} \quad \text{if } r > a. \tag{5.2}$$

We also require that the radial functions be finite and continuous with a continuous first derivative. The set of differential equations together with the boundary and continuity conditions allow us to obtain a unique solution for the radial function.

5.0.1. Radial function. The radial function has the general form

$$\psi_l(r) = Aj_l(kr) + Bh_l(kr) \quad \text{if } 0 < r < a,$$
$$\psi_l(r) = Cj_l(k_0 r) + Dh_l(k_0 r) \quad \text{if } a < r < \infty,$$

where j_l is the spherical Bessel function of the first kind and h_l the spherical Hankel function of the first kind. The requirement that the radial function be finite at the origin $r = 0$ requires that $B = 0$, while the boundary condition (5.2) requires that the constant $C = 1$. The remaining constants A and D are determined by the continuity requirements applied at the boundary $r = a$. These conditions are

$$Aj_l(ka) = j_l(k_0 a) + Dh_l(k_0 a),$$
$$Aj_l{}'(ka) = \frac{k_0}{k}[j_l{}'(k_0 a) + Dh_l{}'(k_0 a)],$$

from which we obtain the solution

$$A = \frac{k_0}{k} \frac{j_l{}'(k_0 a)h_l(k_0 a) - j_l(k_0 a)h_l{}'(k_0 a)}{j_l{}'(ka)h_l(k_0 a) - \frac{k_0}{k} j_l(ka)h_l{}'(k_0 a)},$$
$$D = -\frac{j_l{}'(ka)j_l(k_0 a) - \frac{k_0}{k} j_l(ka)j_l{}'(k_0 a)}{j_l{}'(ka)h_l(k_0 a) - \frac{k_0}{k} j_l(ka)h_l{}'(k_0 a)}.$$

The expression for the constant A can be further simplified by using the Wronskian relation for spherical Bessel functions

$$j_l'(k_0 a)h_l(k_0 a) - j_l(k_0 a)h_l'(k_0 a) = \frac{-i}{k_0^2 a^2}.$$

We find that

$$A = \frac{\frac{-i}{k_0 k a^2}}{j_l{}'(ka)h_l(k_0 a) - \frac{k_0}{k} j_l(ka)h_l{}'(k_0 a)}. \tag{5.3}$$

5.1. Source energy. The quantities Σ_l^2 are found using (4.25) to be given by

$$\begin{aligned} {\Sigma_l}^2 &= \int_0^a r^2 dr\, |\psi_l(r)|^2 \\ &= |A|^2 \int_0^a r^2 dr\, |j_l(kr)|^2 \\ &= T_l(k, k_0)\sigma_l^2(k), \end{aligned} \tag{5.4}$$

where $\sigma_l^2(k)$ is the free space quantity defined in (3.12) *but with* k_0 *replaced by* k and

$$T_l(k, k_0) = |A|^2 = \frac{1}{k_0^2 a^4 |k j_l'(ka)h_l(k_0 a) - k_0 j_l(ka)h_l'(k_0 a)|^2}. \tag{5.5}$$

In the limit when $k \to k_0$ we have that

$$T_l(k, k_0) \to T(k_0, k_0) = \frac{1}{k_0^4 a^4 |j_l'(k_0 a) h_l(k_0 a) - j_l(k_0 a) h_l'(k_0 a)|^2} = 1,$$

where we have used the Wronskian between the spherical Bessel and spherical Hankel functions. It then follows that $\Sigma_l^2 \to \sigma_l^2(k_0)$ in this limit, as required.

The quantities $T_l(k, k_0)$ appearing in the expression for the Σ_l^2 have a simple interpretation: they are the magnitude square of the transmission coefficients relating the amplitudes of the outgoing multipole fields radiated by the source ρ evaluated on the exterior of the source region to the amplitude of the outgoing wave multipole fields radiated by the source on the interior of the source region. In particular, at the interior of the boundary at $r = a$ we can express the field radiated by the source in the form of a superposition of outgoing and standing wave solutions to the homogeneous Helmholtz equation with wavenumber $k = nk_0$, while outside this sphere the field is a superposition of outgoing wave solutions to the Helmholtz equation with wavenumber k_0. Because the total field and normal derivative must be continuous across the boundary, we find that for each multipole mode we require

$$\begin{aligned} h_l(ka) + r_l j_l(ka) &= t_l h_l(k_0 a), \\ h_l'(ka) + r_l j_l'(ka) &= \frac{k_0}{k} t_l h_l'(k_0 a), \end{aligned}$$

where r_l and t_l are reflection and transmission coefficients and the primes denote derivatives. Solving for the transmission coefficients t_l, we find that

$$\begin{aligned} t_l &= \frac{j_l'(ka) h_l(ka) - j_l(ka) h_l'(ka)}{j_l'(ka) h_l(k_0 a) - \frac{k_0}{k} j_l(ka) h_l'(k_0 a)} \\ &= \frac{\frac{-i}{k^2 a^2}}{j_l'(ka) h_l(k_0 a) - \frac{k_0}{k} j_l(ka) h_l'(k_0 a)}, \end{aligned} \tag{5.6}$$

which is seen to be identical to the coefficient A obtained earlier so that $|t_l|^2 = |A|^2 = T_l$, as indicated.

The interpretation of the quantities T_l as being the magnitude square of the transmission coefficients from the field modes in the interior of the source region to the field modes outside this region makes perfect sense in view of the formula (5.4) for the quantities Σ_l^2. In particular, to minimize source energy $\mathcal{E}$ (i.e., to obtain an efficient source) we wish to maximize the Σ_l^2, which, in turn, requires us to maximize the T_l or, equivalently, maximize the amount of energy transmitted from the source interior to the source exterior. As we will find in our simulations presented below, the $T_l(k, k_0)$ vary *inversely* with index value n, so that low source energy is obtained by selecting n to be small. On the other hand, the Σ_l^2 and hence the source energy also depend on the free space quantities σ_l^2 evaluated at the source region wavenumber k, and, as is easily confirmed from the results presented in section 3, the free space quantities $\sigma_l^2(k)$ *increase* with k and, hence, index n, at fixed source radius a. Thus, the two quantities entering into the expression (5.4) for Σ_l^2 have opposite dependencies on source index n, and it is necessary to carefully evaluate the relative importance of each quantity in order to select an index value that leads to small source energy.

To illustrate, we show in Figures 5.1–5.3 plots of the free space quantities $\sigma_l^2(k)$, the modulus square of the transmission coefficient $T_l(k, k_0)$, and finally the $\Sigma_l^2 =$

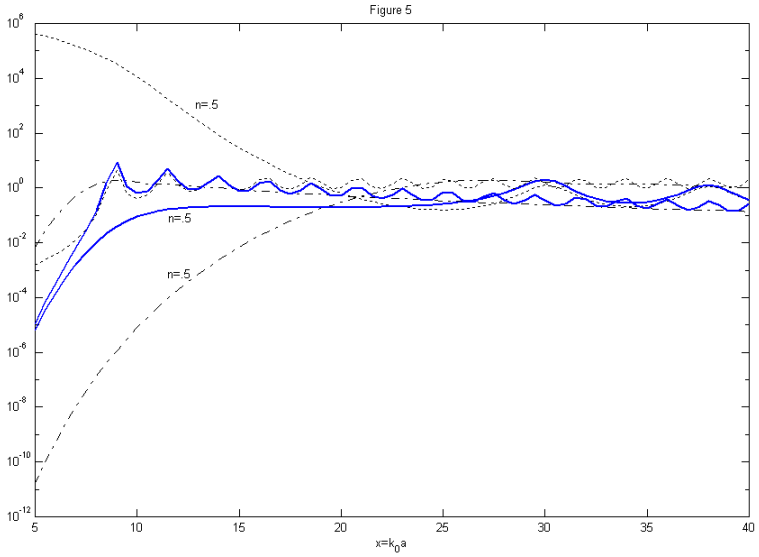


FIG. 5.1. *Plots of* $\sigma_l^2(k = nk_0)$ *(dash-dot),* $T_l(k = nk_0, k_0)$ *(dotted), and* $\Sigma_l^2 = T_l\sigma_l^2$ *(solid) for* $l = 10$ *and* $n = .5$ *and* $n = 1.5$. *It is seen from the plots that the larger* n *value yields larger* Σ_l^2 *around and below the critical point.*

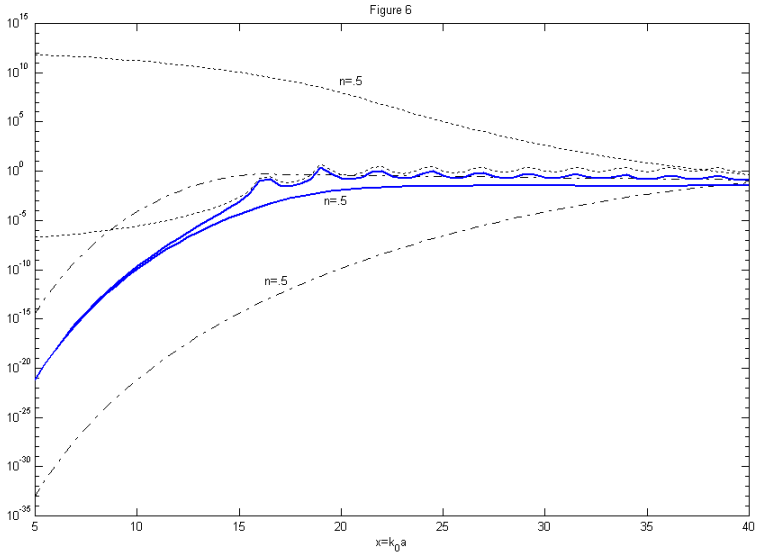


FIG. 5.2. *Plots of* $\sigma_l^2(k = nk_0)$ *(dash-dot),* $T_l(k = nk_0, k_0)$ *(dotted), and* $\Sigma_l^2 = T_l\sigma_l^2$ *(solid) for* $l = 20$ *and* $n = .5$ *and* $n = 1.5$. *It is seen from the plots that the larger* n *value yields larger* Σ_l^2.

$T_l(k, k_0)\sigma_l^2(k)$ plotted as a function of the product $x = k_0a$ of the free space wavenumber with the source radius $a = 10$ and for two values of the source region index n ($n = .5$, $n = 1.5$), for l values of $l = 10, 20, 30$. The following conclusions can be

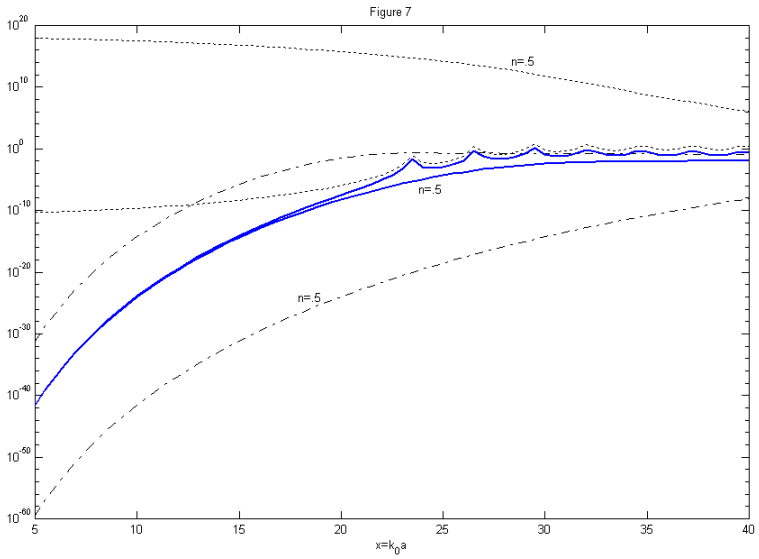


FIG. 5.3. *Plots of* $\sigma_l^2(k = nk_0)$ *(dash-dot),* $T_l(k = nk_0, k_0)$ *(dotted), and* $\Sigma_l^2 = T_l\sigma_l^2$ *(solid) for* $l = 30$ *and* $n = .5$ *and* $n = 1.5$. *It is seen from the plots that the larger* n *value yields larger* Σ_l^2.

drawn from these plots:

- For fixed a and fixed l the free space quantities $\sigma_l^2(k = nk_0)$ *increase* with increasing index n for any given free space wavenumber k_0.
- For fixed a and fixed l the quantities $T_l(k = nk_0, k_0)$ oscillate with k_0. The oscillations indicate the presence of resonances of the scattering functions within the source volume. The T_l are *decreasing* functions of index n at any given free space wavenumber k_0.
- The $\Sigma_l^2 = T_l\sigma_l^2$ oscillate with respect to k_0 due to the resonances of the scattering states in the source region and are also dependent on the source region index n. *For* k_0 *values below the critical point* $k_0a = l$, *the growth of the free space quantity* $\sigma_l^2(k = nk_0)$ *with respect to index* n *tends to outweigh the decay of* $T_l(k = nk_0, k_0)$ *with respect to* n, *with the result that the product* Σ_l^2 *is an increasing function of* n.

A more in-depth look at the behavior of the Σ_l^2 as a function of k_0, n, and l can be obtained from Figures 5.4–5.6. These figures show composite plots of Σ_l^2 for $n = .5, 1, 1.5$ for three different l values ($l = 10, 20, 30$). It is clear from these plots that by making the source region index $n > 1$ it is possible to increase the Σ_l^2 beyond their free space values $\sigma_l^2(k_0)$ and thereby obtain sources which have lower energy than those embedded in free space.

We conclude from the above results that, like the free space quantities $\sigma_l^2(k_0)$, the Σ_l^2 decay exponentially to zero for $l \gg k_0a$. However, by proper selection of the index n of the source region it is possible to obtain higher values of these quantities around and below the critical point $k_0a = l$ and, hence, lower source energy than can be obtained in the free space case for the same radiation pattern. This result follows from the fact that the radiation pattern expansion coefficients $\alpha_{l,m}$ *are independent of the source region index distribution* so that minimum source energy is obtained by

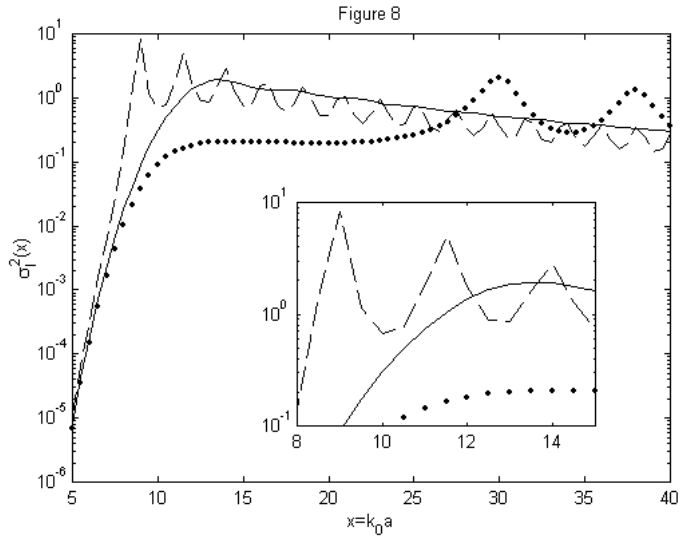


FIG. 5.4. *Plots of* $\Sigma_l^2 = T_l\sigma_l^2$ *for* $l = 10$ *and* $n = .5$ *(dotted),* $n = 1$ *(solid), and* $n = 1.5$ *(dashed). It is seen from the plots that the larger n value yields larger* Σ_l^2 *around and below the critical point* $k_0a = l$.

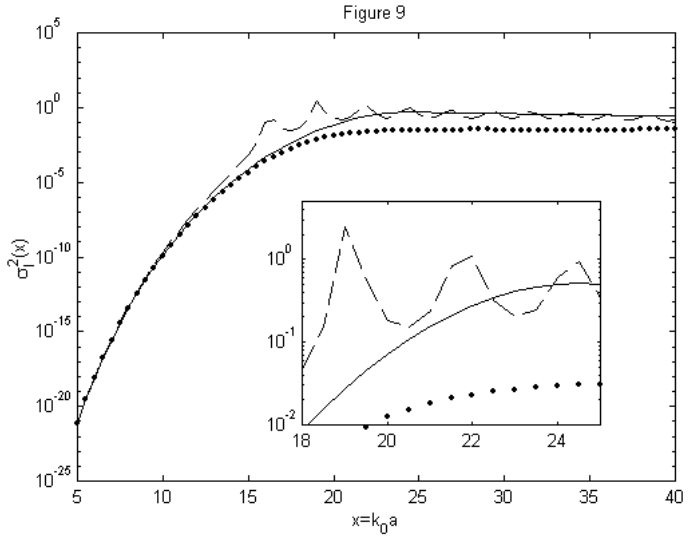


FIG. 5.5. *Plots of* $\Sigma_l^2 = T_l\sigma_l^2$ *for* $l = 20$ *and* $n = .5$ *(dotted),* $n = 1$ *(solid), and* $n = 1.5$ *(dashed). It is seen from the plots that the larger n value yields larger* Σ_l^2 *around and below the critical point* $k_0a = l$.

simply maximizing the Σ_l^2.

We computed the source energy for the model radiation pattern employed in the free space examples of section 3.2. Using the coefficients given in (3.13), we computed the source energy using (3.11) with the Σ_l^2 given by (5.4). We show in Figure 5.7 plots of the source energies as a function of $x = k_0a$ for three different values of the

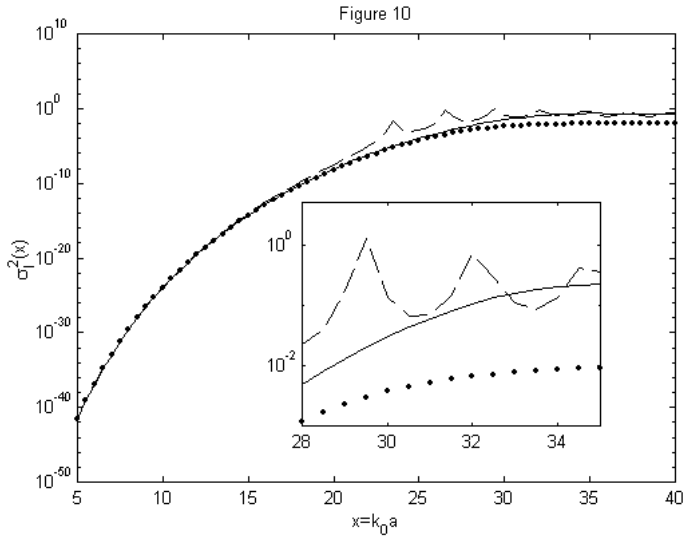


FIG. 5.6. *Plots of* $\Sigma_l^2 = T_l\sigma_l^2$ *for* $l = 30$ *and* $n = .5$ *(dotted),* $n = 1$ *(solid), and* $n = 1.5$ *(dashed). It is seen from the plots that the larger* n *value yields larger* Σ_l^2 *around and below the critical point* $k_0a = l$.

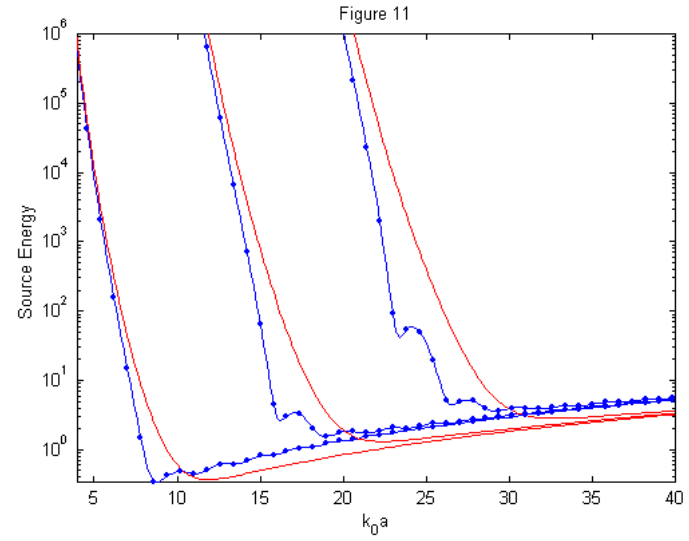


FIG. 5.7. *Plots of source energy for* $a = 10$, $n = 1$ *(solid), and* $n = 1.5$ *(with asterisks) and for* $L = 10, 20, 30$. *It is seen from the plots that the larger* n *value yields smaller energy and, hence, a more efficient source up to the critical point* $k_0a = L$.

cut-off parameter L and for a source radius of $a = 10$ and index value of $n = 1.5$. We also show for comparison the plots of the source energy for a source embedded in free space. It is seen that, as expected, the source energy becomes extremely large if we try to achieve an L value that exceeds the critical value $L = k_0a$. This is, of course, due to the fact that the quantities Σ_l^2 become extremely small when $l > k_0a$.

6. Summary and conclusions. We have developed the basic theory of the inverse source problem for compactly supported sources embedded in an inhomogeneous index profile $n(\mathbf{r})$. Most of our results pertain, in particular, to spherically symmetric index distributions $n(\mathbf{r}) = n(r)$, although the underlying formalism is applicable to general nonsymmetric distributions. For the class of spherically symmetric index profiles we showed how to construct the so-called minimum energy source that generates a given radiation pattern subject to the constraint that the source's L^2 norm over the source region is minimum. It was found that the energy of the minimum energy source depended on the index profile $n(r)$, and we examined this dependence using computer simulations for the case of piecewise constant valued profiles that are unity outside the (spherical) source volume and constant within the source volume and for a "model" radiation pattern characterized by a "resolution parameter" L that was inversely related to the effective angular width of the radiation pattern. The simulations showed that, in general, the source energy increases exponentially when the wavenumber source radius product $k_0 a \gg L$, independent of whether or not the source is embedded in a background medium. However, it was found that by embedding the source in a spherical region having constant index $n > 1$, the source energy can be made smaller than that obtained for a source in vacuum over moderate ranges of the wavenumber source radius product $k_0 a$ in the immediate vicinity of the critical value $k_0 a = L$. We conclude that embedding sources in "designer" background distributions may lead to significant improvement in source efficiency, particularly for resonant antennas. This conclusion has been established here from a new source-inversion point of view, which serves as a theoretical framework of reference for ongoing efforts in this direction within the antenna and optical communities, particularly in connection with novel magneto-dielectrics and metamaterials for enhanced radiation. Currently we are working on the generalization of the research reported in this work to the full vector electromagnetic case including the practical reactive power constraints. We plan to report this ongoing research elsewhere.

Acknowledgments. The authors would like to thank Drs. Arje Nachman and Richard Albanese for helpful comments on the material presented in the paper.

REFERENCES

[1] N. Bleistein and J. K. Cohen, *Nonuniqueness in the inverse source problem in acoustics and electromagnetics*, J. Math. Phys., 18 (1977), pp. 194–201.

[2] R. P. Porter and A. J. Devaney, *Generalized holography and computational solutions to inverse source problems*, J. Opt. Soc. Amer., 72 (1982), pp. 1707–1713.

[3] R. P. Porter and A. J. Devaney, *Holography and the inverse source problem*, J. Opt. Soc. Amer., 72 (1981), pp. 327–330.

[4] M. Bertero, C. De Mol, and E. R. Pike, *Linear inverse problems with discrete data.* I: *General formulation and singular system analysis*, Inverse Problems, 1 (1985), pp. 301–330.

[5] M. Bertero, *Linear inverse and ill-posed problems*, in Advances in Electronics and Electron Physics, Vol. 75, Academic Press, San Diego, 1989, pp. 1–120.

[6] E. A. Marengo and A. J. Devaney, *The inverse source problem of electromagnetics: Linear inversion formulation and minimum energy solution*, IEEE Trans. Antennas Propagat., 47 (1999), pp. 410–412.

[7] E. A. Marengo, A. J. Devaney, and R. W. Ziolkowski, *Inverse source problem and minimum energy sources*, J. Opt. Soc. Amer. A, 17 (2000), pp. 34–45.

[8] E. A. Marengo and R. W. Ziolkowski, *Nonradiating and minimum energy sources, and their fields: Generalized source inversion theory and applications*, IEEE Trans. Antennas Propagat., 48 (2000), pp. 1553–1562.

[9] J. C.-E. Sten, *Reconstruction of electromagnetic minimum energy sources in a prolate spheroid*, Radio Sci., 39 (2004), paper RS2020.
[10] J. C.-E. Sten and E. A. Marengo, *Inverse source problem in an oblate spheroidal geometry*, IEEE Trans. Antennas Propagat., 54 (2006), pp. 3418–3428.
[11] C. Muller, *Foundations of the Mathematical Theory of Electromagnetic Waves*, Springer-Verlag, New York, 1969.
[12] A. J. Devaney and E. Wolf, *Radiating and nonradiating classical current distributions and the fields they generate*, Phys. Rev. D (3), 8 (1973), pp. 1044–1047.
[13] K. Kim and E. Wolf, *Non-radiating monochromatic sources and their fields*, Opt. Comm., 59 (1986), pp. 1–6.
[14] B. J. Hoenders and H. A. Ferwerda, *The non-radiating component of the field generated by a finite monochromatic scalar source distribution*, Pure Appl. Opt. J. Opt. A, 7 (1998), pp. 1201–1211.
[15] A. J. Devaney and G. Sherman, *Nonuniqueness in inverse source and scattering problems*, IEEE Trans. Antennas Propagat., 30 (1982), p. 1034–1037.
[16] B. J. Hoenders, *The uniqueness of inverse problems*, in Inverse Source Problems in Optics, H. P. Baltes, ed., Springer, Berlin, 1978, pp. 41–82.
[17] K. J. Langenberg, *Applied inverse problems for acoustic, electromagnetic and elastic wave scattering*, in Basic Methods of Tomography and Inverse Problems, P. C. Sabatier, ed., Adam Hilger, Bristol, UK, 1987, pp. 128–467.
[18] L. J. Chu, *Physical limitations of omni-directional antennas*, J. Appl. Phys., 19 (1948), pp. 1163–1175.
[19] R. F. Harrington, *On the gain and beamwidth of directional antennas*, IRE Trans. Ant. Propagat., 6 (1958), pp. 219–225.
[20] T. S. Angell, A. Kirsch, and R. E. Kleinman, *Antenna control and optimization using generalized characteristic modes*, Proc. IEEE, 79 (1991), pp. 1559–1568.
[21] D. Liu, R. J. Garbacz, and D. M. Pozar, *Antenna synthesis and solution of inverse problems by regularization methods*, IEEE Trans. Antennas Propagat., 38 (1990), pp. 862–868.
[22] O. M. Bucci, G. D'Elia, G. Mazzarella, and G. Panariello, *Antenna pattern synthesis: A new general approach*, Proc. IEEE, 82 (1994), pp. 358–371.
[23] G. A. Deschamps and H. S. Cabayan, *Antenna synthesis and solution of inverse problems by regularization methods*, IEEE Trans. Antennas Propagat., 20 (1972), pp. 268–274.
[24] Y. T. Lo, S. W. Lee, and Q. W. Lee, *Optimization of directivity and signal-to-noise ratio of an arbitrary antenna array*, Proc. IEEE, 54 (1966), pp. 1033–1045.
[25] A. J. Devaney and R. P. Porter, *Holography and the inverse source problem. Part* II: *Inhomogeneous media*, J. Opt. Soc. Amer., 2 (1985), pp. 2006–2011.
[26] L. Tsang, A. Ishimaru, R. P. Porter, and D. Rouseff, *Holography and the inverse source problem.* III. *Inhomogeneous attenuative media*, J. Opt. Soc. Amer. A, 4 (1987), pp. 1783–1787.
[27] R. G. Newton, *Scattering Theory of Waves and Particles*, Dover Publications, Mineola, NY, 2002.
[28] M. Born and E. Wolf, *Principles of Optics*, 6th ed., Pergamon Press, New York, 1983.
[29] J. D. Jackson, *Classical Electrodynamics*, Wiley, New York, 1975.
[30] H. R. Raemer, *Radiation from linear electric or magnetic antennas surrounded by a spherical plasma shell*, IRE Trans. Ant. Propagat., 10 (1962), pp. 69–78.
[31] D. Lamensdorf, *An experimental investigation of dielectric-coated antennas*, IEEE Trans. Antennas Propagat., 15 (1967), pp. 767–771.
[32] D. Lamensdorf and C.-Y. Ting, *An experimental and theoretical study of the monopole embedded in a cylinder of anisotropic dielectric*, IEEE Trans. Antennas Propagat., 16 (1968), pp. 342–349.
[33] S. A. Long, M. W. McAllister, and L. C. Shen, *The resonant cylindrical dielectric cavity antenna*, IEEE Trans. Antennas Propagat., 31 (1983), pp. 406–412.
[34] K. W. Leung, *Complex resonance and radiation of hemispherical dielectric-resonator antenna with a concentric conductor*, IEEE Trans. Microwave Theory Tech., 49 (2001), pp. 524–531.
[35] J. R. James and J. C. Vardaxoglou, *Investigation of properties of electrically-small spherical ceramic antennas*, Electron. Lett., 38 (2002), pp. 1160–1162.
[36] E. E. Altshuler, *Electrically small genetic antennas immersed in a dielectric*, Proc. IEEE Ant. Propagat. Soc. Symp., 3 (2004), pp. 2317–2320.
[37] R. C. Hansen and M. Burke, *Antennas with magneto-dielectrics*, Microwave Opt. Tech. Lett., 26 (2000), pp. 75–78.
[38] K. Buell, H. Mosallaei, and K. Sarabandi, *A substrate for small patch antennas providing tunable miniaturization factors*, IEEE Trans. Microwave Theory Tech., 54 (2006), pp. 135–146.

[39] H. Mosallaei and K. Sarabandi, *Magneto-dielectrics in electromagnetics: Concept and applications*, IEEE Trans. Antennas Propagat., 52 (2004), pp. 1558–1567.
[40] Y. Mano and S. Bae, *A small meander antenna by magneto-dielectric material*, in Proceedings of the IEEE International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications, Beijing, China, 2005, IEEE Press, Piscataway, NJ, 2005, Vol. 1, pp. 63–66.
[41] M. Thèvenot, C. Cheype, A. Reineix, and B. Jecko, *Directive photonic-bandgap antennas*, IEEE Trans. Microwave Theory Tech., 47 (1999), pp. 2115–2122.
[42] B. Temelkuran, M. Bayindir, E. Ozbay, R. Biswas, M. M. Sigalas, G. Tuttle, and K. M. Ho, *Photonic crystal-based resonant antenna with a very high directivity*, J. Appl. Phys., 87 (2000), pp. 603–605.
[43] S. Enoch, G. Tayeb, P. Sabouroux, N. Guérin, and P. Vincent, *A metamaterial for directive emission*, Phys. Rev. Lett., 89 (2002), paper 213902.
[44] A. Alu and N. Engheta, *Radiation from a traveling-wave current sheet at the interface between a conventional material and a metamaterial with negative permittivity and permeability*, Microwave Opt. Technol. Lett., 35 (2002), pp. 460–463.
[45] R. W. Ziolkowski and A. Kipple, *Application of double negative metamaterials to increase the power radiated by electrically small antennas*, IEEE Trans. Antennas Propagat., 51 (2003), pp. 2626–2640.
[46] D. Tonn and R. Bansal, *Practical considerations for increasing radiated power from an electrically small antenna by application of a double negative metamaterial*, in Proceedings of the IEEE International Antenna Propagation Symposium, Washington, DC, 2005, Vol. 2A, pp. 602–605.
[47] A. Erentok, P. L. Luljak, and R. W. Ziolkowski, *Characterization of a volumetric metamaterial realization of an artificial magnetic conductor for antenna applications*, IEEE Trans. Antennas Propagat., 53 (2005), pp. 160–172.
[48] G. Lovat, P. Burghignoli, F. Capolino, D. R. Jackson, and D. R. Wilton, *Analysis of directive radiation from a line source in a metamaterial slab with low permittivity*, IEEE Trans. Antennas Propagat., 54 (2006), pp. 1017–1030.
[49] B.-I. Wu, W. Wang, J. Pacheco, X. Chen, T. Grzegorczyk, and J. A. Kong, *A study of using metamaterials as antenna substrate to enhance gain*, Prog. Electromag. Res., 51 (2005), pp. 295–328.
[50] J. A. Kong, *Electromagnetic Wave Theory*, EMW Publishing, Cambridge, MA, 2005.
[51] X. Chen, T. M. Grzegorczyk, and J. A. Kong, *Optimization approach to the retrieval of the constitutive parameters of a slab of general bianisotropic medium*, Prog. Electromag. Res., 60 (2006), pp. 1–18.
[52] E. A. Marengo, A. J. Devaney, and F. K. Gruber, *Inverse source problem with reactive power constraints*, IEEE Trans. Antennas Propagat., 52 (2004), pp. 1586–1595.
[53] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer, Berlin, 1998.
[54] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, Wiley, New York, 1983.
[55] G. Arfken, *Mathematical Methods for Physicists*, Academic Press, San Diego, 1985.
[56] J. R. Taylor, *Scattering Theory*, Wiley, New York, 1972.
[57] R. Potthast, *Point Sources and Multipoles in Inverse Scattering Theory*, Chapman & Hall, Boca Raton, FL, 2001.
[58] A. J. Devaney and M. L. Oristaglio, *Inversion procedure for inverse scattering within the distorted-wave Born approximation*, Phys. Rev. Lett., 51 (1983), pp. 237–240.
[59] R. W. Ziolkowski and A. D. Kipple, *Reciprocity between the effects of resonant scattering and enhanced radiated power by electrically small antennas in the presence of nested metamaterial shells*, Phys. Rev. E (3), 72 (2005), paper 036602.
[60] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, McGraw–Hill, New York, 1953.

SIAM J. APPL. MATH.
Vol. 67, No. 5, pp. 1379–1395

GLOBAL DYNAMICS OF A PREDATOR-PREY MODEL WITH STAGE STRUCTURE FOR THE PREDATOR*

PAUL GEORGESCU† AND YING-HEN HSIEH‡

Abstract. The global properties of a predator-prey model with nonlinear functional response and stage structure for the predator are studied using Lyapunov functions and LaSalle's invariance principle. It is found that, under hypotheses which ensure the uniform persistence of the system and the existence of a unique positive steady state, a feasible a priori lower bound condition on the abundance of the prey population ensures the global asymptotic stability of the positive steady state. A condition which leads to the extinction of the predators is indicated. We also obtain results on the existence and stability of periodic solutions. In particular, when (4.2) fails to hold and the unique positive steady state E^* becomes unstable, the coexistence of prey and predator populations is ensured for initial populations not on the one-dimensional stable manifold of E^*, albeit with fluctuating population sizes.

Key words. predator-prey model, stage structure, global stability, uniform persistence, Lyapunov function

AMS subject classifications. 92D25, 92D30, 34D20, 34D23, 93D20

DOI. 10.1137/060670377

1. Introduction.

In classical models of Lotka–Volterra type it is assumed that all individuals of a single species have largely similar capabilities to hunt or reproduce. However, the life cycle of most, if not all, animals and insects consists of at least two stages, immature and mature, and the individuals in the first stage often can neither hunt nor reproduce, being raised by their mature parents. Furthermore, immediately recognizable morphological and behavioral differences may exist between these stages and other adaptive stages, such as dormancy stages for immediate survival purposes.

To study this situation theoretically, stage-structured models have attracted much attention in recent decades. Fundamental work towards a systematic approach to stage-structured model formulation has been made by Gurney, Nisbet, and Blythe [7], Nisbet and Gurney [27], and Nisbet, Gurney, and Metz [28]. Further progress has been made by Aiello and Freedman, who proposed and studied in their often quoted work [1] a single species model with stage structure and discrete delay, predicting the global attractivity of the positive steady state and thereby suggesting that the stage structure does not generate sustained oscillations, at least for a single species model. General consistency criteria to be satisfied by models which describe stage-structured ecological interactions have been laid out in Kuang [18] or Arditi and Michalski [2]. See also Liu, Chen, and Agarwal [24] for a recent survey on the dynamics of stage-structured population models with an emphasis on modeling issues.

Predator-prey models with stage structure for the predator have received considerable attention in recent years. See Wang [35] and Xiao and Chen [38] for global

*Received by the editors September 21, 2006; accepted for publication (in revised form) April 2, 2007; published electronically July 18, 2007. This research was supported by National Science Council (NSC-Taiwan) research grant NSC-94-2115-M-005-006, which funded the first author's visit to National Chung Hsing University as a NSC research fellow.

http://www.siam.org/journals/siap/67-5/67037.html

†Department of Mathematics, Technical University of Iaşi, Bd. Copou 11, 700506 Iaşi, Romania (vpgeo@go.com).

‡Corresponding author. Department of Public Health and Biostatistics Center, China Medical University, Taichung 404, Taiwan (hsieh@amath.nchu.edu.tw).

stability and persistence analysis of a stage-structured predator-prey model without delay terms. See also Wang and Chen [36], Wang et al. [37], and Gourley and Kuang [9] for stability analyses of staged predator-prey models with time delays due to gestation of the predator and crowding of the prey.

Apart from analyzing the stability of their delayed model, Gourley and Kuang [9] also discussed its oscillatory dynamics for a linear functional response of the mature predator and observed that sustained oscillations took place only for a limited interval of maturation delays. This happens since, for small delays, their model inherits the properties of the nondelayed (of Lotka–Volterra type) system. However, if the maturation delay is too long, then the highest possible recruitment rate to adulthood drops below the adult death rate and the predator population dies out.

As far as the asymptotic behavior of predator-prey systems is concerned, it is known from Poincaré–Bendixson theory that two-dimensional continuous time models can approach either an equilibrium state or a limit cycle with any type of chaotic behavior being excluded, while three- and higher-dimensional models can exhibit more complex behavior. In this regard, staged models may provide in some situations a richer dynamics which leads to a better understanding of the interactions within the biological system under consideration. Such models may also incorporate meaningful biological parameters, such as different death rates for mature and immature predators and various delay effects.

In [36], [35], [38] the following predator-prey model with stage structure for the predator has been considered:

$$\begin{cases} x'(t) = x(t)\,(r - ax(t)) - \dfrac{bx(t)}{1+mx(t)}y_2(t), \\ y_1'(t) = k\dfrac{bx(t)}{1+mx(t)}y_2(t) - (D+d_1)y_1(t), \\ y_2'(t) = Dy_1(t) - d_2y_2(t). \end{cases} \tag{1.1}$$

Here $x(t)$, $y_1(t)$, $y_2(t)$ are the densities of prey, respectively of immature and mature predators at time t. It is assumed that in the absence of the predators the prey grows according to a logistic law with intrinsic growth rate r and carrying capacity r/a, while predators feed on prey only and do not count towards the carrying capacity. It is also assumed that the immature predators are either raised by their parents or consume a resource which is available in abundance and for which they do not have to compete. As a consequence, neither crowding nor intraspecies competition terms are added into the equation which models the growth of the immature predator class. The function $x \mapsto bx/(1+mx)$ represents the Holling type 2 functional (behavioral) response of the mature predator, which describes how the consumption rate of the predator depends on prey density, b being the search rate and m being the search rate multiplied by the handling time; while the function $x \mapsto kbx/(1+mx)$ is the associated numerical (reproductive) response of the mature predator which quantifies the relation between the numerical growth of the predator class and the prey consumption, with k representing the conversion coefficient under the assumption that the reproduction rate of the mature predators is directly proportional to the amount of prey consumed. The constants d_1 and d_2 represent the death rates of immature and mature predators, and D denotes the rate at which immature predators become mature predators.

It was proved in Wang [35] that if the condition

$$d_2(D+d_1) < \frac{kbrD}{a+mr} \tag{1.2}$$

holds, then the system (1.1) is uniformly persistent and a unique positive steady state $E^* = (x^*, y_1^*, y_2^*)$ exists. Moreover, it is shown that if, in addition to (1.2), conditions

$$x^*(D+d_1+d_2)(a+2max^*-mr)\left(D+d_1+d_2+\frac{x^*(a+2max^*-mr)}{1+mx^*}\right) > \frac{by_2^*d_2(D+d_1)}{1+mx^*}, \tag{1.3}$$

$$a > b+\frac{bmy_2^*}{1+mx^*}, \quad D+d_1 > \frac{kbr}{a+mr}+\frac{kby_2^*}{1+mx^*}, \quad d_2 > D, \tag{1.4}$$

are also satisfied, then the positive steady state $E^* = (x^*, y_1^*, y_2^*)$ is globally asymptotically stable. The proof uses the theory of competitive systems as developed in Smith [33], with condition (1.3) being used to establish the local stability of E^*.

More recently, Xiao and Chen [38] noted that condition (1.4) contradicts condition (1.2), and showed that the positive steady state E^* is globally asymptotically stable if (1.2) and (1.3) hold, in addition to one of the following two conditions:

$$\text{(H1) } D+d_1 > r \text{ and } \underline{x} > \frac{r}{2a}; \quad \text{(H2) } D+d_1 < r \text{ and } \underline{x} > \frac{r+D+d_1}{2a}.$$

Here $\underline{x} > 0$ is the persistency constant for x, which satisfies $\underline{x} \le \liminf_{t\to\infty} x(t)$. The proof is again based on the theory of competitive systems and uses a result given by Li and Muldowney in [23], which amounts to the fact that for competitive and permanent systems which are defined on convex and bounded sets and have the property of stability of periodic orbits, the local asymptotic stability of a unique positive steady state implies its global asymptotic stability. Essentially, the proof in [38] amounts to showing that the system (1.1) has the property of the stability of periodic orbits under either (H1) or (H2), a fact which is established using a criterion of Muldowney [26] and the theory of additive compound matrices.

Consider the conditions (1.3), (H1), and (H2). It is clear that if the inequality $\underline{x} > (r+D+d_1)/(2a)$, which is required in (H2), can be weakened to $\underline{x} > r/(2a)$ and either (H1) or (H2) can be modified to cover the case $D+d_1 = r$, then (H1) and (H2) can be combined into a single condition $\underline{x} > r/(2a)$, where $r/(2a)$ is the prey population size at the inflection point of the logistic curve in a prey-only system. Moreover, condition (1.3), which a priori ensures the local stability of the positive steady state, was motivated by specifics of the method used for the proof, which roughly inputs local asymptotic stability and outputs global asymptotic stability under certain assumptions.

However, it is clear that once the global asymptotic stability of the positive steady state is proved, then its local asymptotic stability is superseded anyway. Moreover, we shall indicate in section 4 that in fact (1.3) is satisfied if $x^* > r/(2a)$ (and consequently if $\underline{x} > r/(2a)$), and so there is no need to assume (1.3) separately.

In this article, we will study the global dynamics of (1.1) by constructing a suitable Lyapunov function and using LaSalle's invariance principle rather than by using the theory of competitive systems, as has been done in [35] and [38]. This will enable us to obtain the global asymptotic stability of the positive steady state under weaker hypotheses than those used in Xiao and Chen [38] and by a simpler method. In our setting, the persistence condition $\underline{x} > r/(2a)$ used in [38] will appear in a natural way as a monotonicity condition. We will also discuss in section 4 the existence of periodic solutions, together with their stability. Finally, we will discuss the biological significance of our results and indicate possible extensions to the study of more comprehensive models in section 5.

2. The model and its well-posedness. In this section we analyze the global existence of the solutions of (1.1) and their positivity properties.

Let us define $n : [0,\infty) \to \mathbb{R}$ and $f : [0,\infty) \to [0,\infty)$ by $n(x) = x(r - ax)$ and $f(x) = bx/(1 + mx)$ for all $x \in [0,\infty)$. Using the newly defined functions n and f we can rewrite (1.1) as

$$\begin{cases} x' = n(x) - f(x)y_2, \\ y_1' = kf(x)y_2 - (D + d_1)y_1, \\ y_2' = Dy_1 - d_2y_2. \end{cases} \tag{2.1}$$

Note that n is strictly decreasing on $[r/(2a), +\infty)$, while f is strictly increasing on $[0,\infty)$.

First, it is easy to see that if $x(0)$, $y_1(0)$, $y_2(0) \geq 0$, then $x(t), y_1(t), y_2(t) \geq 0$ on their respective intervals of existence. For this purpose, we observe that the vector (R_1, R_2, R_3) points inside the closed set $Q_1 = [0,\infty)^3$ at all points of ∂Q_1, where R_1, R_2, R_3 are the right-hand sides appearing in (1.1), and so Nagumo's tangency conditions are satisfied and Q_1 is a positively invariant set for (1.1). See Pavel [29] for further reference on flow invariance problems for ODEs and abstract ODEs.

To prove that $Q_2 = (0,\infty)^3$ is also a positively invariant set for (1.1), suppose that $x(0), y_1(0), y_2(0) > 0$ and note first that $\frac{d}{dt}(y_2e^{d_2t}) = Dy_1e^{d_2t} \geq 0$, and so $t \mapsto y_2(t)e^{d_2t}$ is increasing. It follows that $y_2(t) \geq y_2(0)e^{-d_2t}$ for all t for which $y_2(t)$ is well defined, and hence y_2 remains strictly positive. Also, $\frac{d}{dt}(y_1e^{(D+d_1)t}) \geq 0$, consequently, $y_1(t) \geq y_1(0)e^{-(D+d_1)t}$ and y_1 remains strictly positive. To prove that x also remains strictly positive, suppose that $x(t_0) = 0$ for some $t_0 > 0$. Then one may find $\tilde{y_1}(0)$ and $\tilde{y_2}(0) > 0$ such that the solution which starts at $t = 0$ from $(0, \tilde{y_1}(0), \tilde{y_2}(0))$ also reaches $(0, y_1(t_0), y_2(t_0))$ at $t = t_0$. By the uniqueness property of (1.1), this solution should coincide with the solution which starts at $t = 0$ from $(x(0), y_1(0), y_2(0))$, which is an obvious contradiction.

We shall now show that x, y_1, y_2 are bounded on their intervals of existence, which in turn will imply by a standard continuability argument that they are defined on $[0,\infty)$. Denote $M_1 = \max(r/a, x(0))$ and $d = \min(d_1, d_2)$. Since $x' \leq x(r - ax)$, it follows that $x(t) \leq M_1$ for all t. That is, x is bounded and consequently defined on $[0,\infty)$. Let us consider the Lyapunov function

$$U(x, y_1, y_2) = x + (1/k)y_1 + (1/k)y_2.$$

We now compute the time derivative of U along the solutions of (1.1). One then has

$$\dot{U} = n(x) - \frac{d_1}{k}y_1 - \frac{d_2}{k}y_2,$$

which implies

$$\dot{U} + dU \leq (r + d)x.$$

Consequently,

$$U(x(t), y_1(t), y_2(t)) \leq U(x(0), y(0), z(0))e^{-dt} + \frac{M_1(r + d)}{d}\left(1 - e^{-dt}\right) \quad \text{for all } t.$$

This implies that y_1, y_2 are also bounded and consequently defined on $[0,\infty)$. Finally, we analyze the behavior of solutions which start with initial data (x_i, y_{1i}, y_{2i}) on the boundary of $(0,\infty)^3$.

If $x_i = 0$, then $(x(t), y_1(t), y_2(t)) \to (0,0,0)$ irrespective of the initial values $y_{1i}, y_{2i} \geq 0$. If $x_i > 0$, then $(x(t), y_1(t), y_2(t)) \to (r/a, 0, 0)$ for $y_{1i} = y_{2i} = 0$, while $(x(t), y_1(t), y_2(t))$ enters $(0,\infty)^3$ (and stays there) otherwise.

3. Global dynamics of the model. In this section we perform a global stability analysis for the system (1.1) regarding both the stability of the boundary equilibrium $(r/a, 0, 0)$ (i.e., the case in which the predator classes tend to extinction) and of the positive steady state (x^*, y_1^*, y_2^*) (i.e., the case in which the coexistence of both species is assured for all future time). As a result, we find sufficient conditions for the stability of the equilibria and establish the existence of a threshold parameter.

Let us denote $T = d_2(D + d_1)/D$ and $x_0 = r/a$. First, we give a condition for the extinction of the predators.

THEOREM 3.1. *Suppose that $T \geq kf(x_0)$. Then $(x_0, 0, 0)$ is globally asymptotically stable on $(0,\infty)^3$.*

Proof. Let us consider the Lyapunov function

$$U_1(x, y_1, y_2) = \int_{x_0}^{x} \frac{f(\tau) - f(x_0)}{f(\tau)} d\tau + \frac{1}{k} y_1 + \frac{1}{k} \frac{D + d_1}{D} y_2.$$

We now compute the time derivative of U_1 along the solutions of (1.1). One then has

$$\begin{aligned} \dot{U_1} &= \frac{f(x) - f(x_0)}{f(x)} (n(x) - f(x)y_2) + \frac{1}{k} (kf(x)y_2 - (D + d_1)y_1) \\ &\quad + \frac{1}{k} \frac{D + d_1}{D} (Dy_1 - d_2 y_2) \\ &= \frac{f(x) - f(x_0)}{f(x)} n(x) + \frac{1}{k} \left(kf(x_0) - \frac{(D + d_1)d_2}{D} \right) y_2. \end{aligned}$$

Since c is strictly increasing on $[0,\infty)$ and $\operatorname{sgn} n(x) = \operatorname{sgn}(x_0 - x)$ for $x \in (0,\infty)$, it is seen that $\dot{U_1} \leq 0$, with equality if and only if $x = x_0$ and either $y_2 = 0$ or $T = kf(x_0)$. In both cases, the only invariant subset $\tilde{M}$ within the set $M = \{(x, y_1, y_2); x = x_0\}$ is $\tilde{M} = \{(x_0, 0, 0)\}$.

Since $\dot{U_1} \leq 0$ on $(0,\infty)^3$ and the only possible ω-limit sets of $(x(t), y_1(t), y_2(t))$ on the boundary of $(0,\infty)^3$ are $\{(x_0, 0, 0)\}$ and $\{(0, 0, 0)\}$, our conclusion follows from LaSalle's invariance principle (see [22]). □

We now attempt to analyze the existence of the positive steady state E^* and the uniform persistence of the system (1.1). We recall that the system (1.1) is said to be *uniformly persistent* if there is $\varepsilon_0 > 0$ such that any solution of (1.1) which starts with $x(0), y_1(0), y_2(0) > 0$ satisfies

$$\liminf_{t\to\infty} x(t) \geq \varepsilon_0, \quad \liminf_{t\to\infty} y_1(t) \geq \varepsilon_0, \quad \liminf_{t\to\infty} y_2(t) \geq \varepsilon_0.$$

For other (weaker) types of persistence and criteria to establish the persistence of a given system, see Butler, Freedman, and Waltman [4], Freedman, Ruan, and Tang [6], and Hofbauer and So [11].

THEOREM 3.2. *Suppose that $T < kf(x_0)$. Then the positive steady state E^* exists, is unique, and the system (1.1) is uniformly persistent.*

Proof. Let us consider the Lyapunov function

$$U_2(x, y_1, y_2) = \frac{1}{k} y_1 + \frac{1}{k} \frac{D + d_1}{D} y_2.$$

We now compute the time derivative of U_2 along the solutions of (1.1). One then has

$$\begin{aligned}\dot{U_2} &= \frac{1}{k}(kf(x)y_2 - (D+d_1)y_1) + \frac{1}{k}\frac{D+d_1}{D}(Dy_1 - d_2y_2)\\ &= \left(f(x) - \frac{(D+d_1)d_2}{kD}\right)y_2.\end{aligned}$$

If $T < kf(x_0)$, then $\dot{U_2}$ is positive in all strictly positive points of a vicinity of $(x_0, 0, 0)$, and so $(x_0, 0, 0)$ is unstable. Since the only invariant subsets on the boundary of $(0,\infty)^3$ are $\{(x_0, 0, 0)\}$ and $\{(0, 0, 0)\}$ and their stable manifolds are also contained in the boundary of $(0,\infty)^3$, it follows from a result of Hofbauer and So [11] that the system (1.1) is uniformly persistent. Also see Margheri and Rebelo [25] for a slightly different approach towards showing the persistence of dynamical systems based on a result of Fonda [5], which establishes necessary and sufficient conditions for a given compact set S to be a uniform repeller.

To show the existence of E^*, we need to find positive solutions for the system

$$\begin{cases} x^*\,(r - ax^*) - \dfrac{bx^*}{1+mx^*}y_2^* = 0,\\ k\dfrac{bx^*}{1+mx^*}y_2^* - (D+d_1)y_1^* = 0,\\ Dy_1^* - d_2y_2^* = 0.\end{cases} \tag{3.1}$$

After some algebraic manipulations, one obtains

$$x^* = \frac{(D+d_1)d_2}{bkD - m(D+d_1)d_2}, \quad y_1^* = \frac{x^*(r-ax^*)k}{(D+d_1)}, \quad y_2^* = \frac{x^*(r-ax^*)kD}{(D+d_1)d_2}. \tag{3.2}$$

Since $d_2(D+d_1)/D < kbr/(a+mr)$, it follows that $bkD/((D+d_1)d_2) > (a+mr)/r$, and so $x^* < r/a$. From the above, it also follows that $bkD/((D+d_1)d_2) > m$, and hence $x^* > 0$. Consequently, x^*, y_1^*, y_2^* are all well defined and positive. We also remark that since the system (1.1) is uniformly persistent, it follows that there is an $\underline{x} > 0$ such that $\liminf_{t\to\infty} x(t) \geq \underline{x}$. □

From Theorems 3.1 and 3.2, combined with the remark about the behavior of the solutions starting on the boundary of $[0,\infty)^3$ which was made at the end of section 2, it also follows that $(0, 0, 0)$ is an unstable equilibrium and its stable manifold consists of the positive quadrant $\{(0, y_{1i}, y_{2i}); y_{1i}, y_{2i} \geq 0\}$. That is, our model predicts that the predator and the prey cannot simultaneously face extinction, with the sole exception of the case in which the size of the initial prey populations equals zero, justified by the fact that the predators feed on prey only and do not consume other resource, and therefore in the absence of prey they are condemned to extinction.

Having established the existence and uniqueness of the positive steady state E^*, we now turn our attention to its stability. For this purpose, we employ a condition on the persistence constant $\underline{x}$, which ensures that the size of the prey population remains ultimately higher than a certain value.

THEOREM 3.3. *Suppose that $T < kf(x_0)$ and $\underline{x} > r/(2a)$. Then the positive steady state E^* is globally asymptotically stable on $(0,\infty)^3$.*

Proof. Since $\underline{x} > r/(2a)$, it is seen that there is $t_0 \geq 0$ such that $x(t) > r/(2a)$ for all $t \geq t_0$ and also that $x^* > r/(2a)$. Let us consider the Lyapunov function

$$U_3(x, y_1, y_2) = \int_{x^*}^{x} \frac{f(\tau) - f(x^*)}{f(\tau)}d\tau + \frac{1}{k}\int_{y_1^*}^{y_1} \frac{\tau - y_1^*}{\tau}d\tau + \frac{1}{k}\frac{D+d_1}{D}\int_{y_2^*}^{y_2} \frac{\tau - y_2^*}{\tau}d\tau.$$

It is easily seen that $U_3(x, y_1, y_2) \geq 0$ and $U_3(x, y_1, y_2) = 0$ if and only if $x = x^*$, $y_1 = y_1^*$, $y_2 = y_2^*$. We now compute the time derivative of U_3 along the solutions of (1.1). One obtains that

$$\begin{aligned}\dot{U_3} &= \frac{f(x) - f(x^*)}{f(x)}\,(n(x) - f(x)y_2) + \frac{1}{k}\frac{y_1 - y_1^*}{y_1}\,(kf(x)y_2 - (D + d_1)y_1) \\ &\quad + \frac{1}{k}\frac{D + d_1}{D}\frac{y_2 - y_2^*}{y_2}\,(Dy_1 - d_2y_2) \\ &= n(x)\frac{f(x) - f(x^*)}{f(x)} + f(x^*)y_2 - \frac{D + d_1}{k}y_1^*\left(\frac{f(x)}{f(x^*)}\frac{y_2}{y_2^*}\frac{y_1^*}{y_1} + \frac{y_2^*}{y_2}\frac{y_1}{y_1^*} + \frac{f(x^*)}{f(x)} - 3\right) \\ &\quad + \frac{D + d_1}{k}y_1^*\frac{f(x^*)}{f(x)} - \frac{D + d_1}{k}y_1^* - \frac{D + d_1}{kD}d_2y_2.\end{aligned}$$

Since $f(x^*) = (D + d_1)d_2/(kD)$, this yields

$$\begin{aligned}\dot{U_3} &= n(x)\frac{f(x) - f(x^*)}{f(x)} - \frac{D + d_1}{k}y_1^*\left(\frac{f(x)}{f(x^*)}\frac{y_2}{y_2^*}\frac{y_1^*}{y_1} + \frac{y_2^*}{y_2}\frac{y_1}{y_1^*} + \frac{f(x^*)}{f(x)} - 3\right) \\ &\quad + \frac{D + d_1}{k}y_1^*\left(\frac{f(x^*)}{f(x)} - 1\right) \\ &= \frac{1}{f(x)}(n(x) - n(x^*))(f(x) - f(x^*)) \\ &\quad - \frac{D + d_1}{k}y_1^*\left(\frac{f(x)}{f(x^*)}\frac{y_2}{y_2^*}\frac{y_1^*}{y_1} + \frac{y_2^*}{y_2}\frac{y_1}{y_1^*} + \frac{f(x^*)}{f(x)} - 3\right).\end{aligned}$$

From the *AM-GM* inequality, it is clear that

$$\frac{f(x)}{f(x^*)}\frac{y_2}{y_2^*}\frac{y_1^*}{y_1} + \frac{y_2^*}{y_2}\frac{y_1}{y_1^*} + \frac{f(x^*)}{f(x)} \geq 3,$$

with equality if and only if

$$\frac{f(x)}{f(x^*)}\frac{y_2}{y_2^*}\frac{y_1^*}{y_1} = \frac{y_2^*}{y_2}\frac{y_1}{y_1^*} = \frac{f(x^*)}{f(x)} = 1,$$

that is, $x = x^*$ and $y_1/y_1^* = y_2/y_2^*$.

If $x(t) > r/(2a)$ for $t \geq t_0$, then since n is strictly decreasing on $[r/(2a), \infty)$ and f is strictly increasing on $[0, \infty)$, it follows that

$$\frac{1}{f(x)}(n(x) - n(x^*))(f(x) - f(x^*)) \leq 0,$$

with equality if and only if $x = x^*$. This implies that $\dot{U_3} \leq 0$, with equality if and only if $x = x^*$ and $y_1/y_1^* = y_2/y_2^*$. We now find the invariant subsets $\tilde{M}$ within the set

$$M = \left\{(x, y_1, y_2); x = x^*, \frac{y_1}{y_1^*} = \frac{y_2}{y_2^*}\right\}.$$

Since $x = x^*$ on $\tilde{M}$ and consequently $x' = n(x^*) - f(x^*)y_2$, it follows that $x' = f(x^*)(y_2 - y_2^*)$, and so $y_2 = y_2^*$. This implies $y_1 = y_1^*$, and consequently the only invariant set in M is $\tilde{M} = \{(x, y_1^*, y_2^*)\}$. From LaSalle's invariance principle one then obtains the desired conclusion. □

4. The local stability of the positive steady state and the existence of the periodic solutions. Suppose now that $T < kf(x_0)$ and consequently that the system (1.1) is persistent and the positive steady state E^* exists and is unique. As seen in Wang [35] and Xiao and Chen [38], it is possible to study the local stability of the positive steady state and the existence of the periodic solutions together with their orbital stability by using a result on the behavior of three-dimensional competitive systems established by Zhu and Smith in [39].

It is easy to see that the Jacobian of the system (1.1) at (x, y_1, y_2) is given by

$$J_{(1.1)}(x, y_1, y_2) = \begin{pmatrix} r - 2ax - \frac{b}{(1+mx)^2} y_2 & 0 & -\frac{bx}{1+mx} \\ k\frac{b}{(1+mx)^2} y_2 & -(D + d_1) & \frac{kbx}{1+mx} \\ 0 & D & -d_2 \end{pmatrix}.$$

Using the equilibrium relations (3.1), one finds that the characteristic equation of the system (1.1) at E^* is given by

$$\lambda^3 + \left[D + d_1 + d_2 + x^* \left(2a - \frac{rm + a}{1 + mx^*}\right)\right] \lambda^2 + x^* \left(2a - \frac{rm + a}{1 + mx^*}\right) (D + d_1 + d_2)\lambda + \frac{r - ax^*}{1 + mx^*} d_2(D + d_1) = 0. \tag{4.1}$$

Consequently, by the classical Routh–Hurwitz theorem, all roots of (4.1) have negative real parts if

$$\left[D + d_1 + d_2 + x^* \left(2a - \frac{rm + a}{1 + mx^*}\right)\right] x^* \left(2a - \frac{rm + a}{1 + mx^*}\right) (D + d_1 + d_2) > \frac{r - ax^*}{1 + mx^*} d_2(D + d_1), \tag{4.2}$$

and if the reverse of the above inequality is satisfied, then two of the characteristic roots have positive real parts. Note that since $(r - ax^*)/(1 + mx^*)d_2(D + d_1) > 0$, there is always a negative real root of (4.1). It is also important to note that (4.2) is satisfied if $x^* > r/(2a)$. Toward this goal, we remark that if $x^* > r/(2a)$, one has

$$x^*(2a(1 + mx^*) - (rm + a)) = x^* \left(2am \left(x^* - \frac{r}{2a}\right) + a\right) \geq ax^* \geq r - ax^*$$

and

$$\left[D + d_1 + d_2 + x^* \left(2a - \frac{rm + a}{1 + mx^*}\right)\right] (D + d_1 + d_2) > 4d_2(D + d_1),$$

from which (4.2) results immediately. It then follows that all equilibria E^* with $x^* > r/(2a)$ are locally asymptotically stable. Moreover, a quick inspection of our argument shows that E^* is also stable for some $x^* < r/(2a)$, provided that $x^* > r/(2a) - \tilde{c}/(2m)$, where

$$\tilde{c} = \left(1 + \frac{a}{mr} - \sqrt{1 - \left(1 - \frac{a}{mr}\right)^2 + 4\frac{a}{mr}\frac{d_2\,(D + d_1)}{(D + d_1 + d_2)^2}}\right) \frac{mr}{2a}. \tag{4.3}$$

In particular, this shows that the inequality (4.2), which has been a priori assumed in Xiao and Chen [38] (stated under the equivalent form (1.3)), does actually follow if

either (H1) or (H2) are assumed, since $\underline{x} > r/(2a)$ implies $x^* > r/(2a)$, and so there is no need to assume (4.2) separately. Also, it is perhaps interesting to remark that while the inequality $\underline{x} > r/(2a)$ ensures the global stability of E^*, a somewhat similar but weaker estimate $x^* > r/(2a)$ ensures its local stability. We do not know, however, whether or not the inequality $\underline{x} > r/(2a)$ is sharp, that is, if $r/(2a)$ is the smallest constant C with the property that $\underline{x} > C$ ensures the converge of the respective solution of (1.1) to E^*, under the condition $kf(r/a) > T$.

Consider now

$$C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad S = [0,\infty) \times (-\infty, 0] \times [0,\infty).$$

One then has

$$CJ_{(1.1)}(x, y_1, y_2)C = \begin{pmatrix} r - 2ax - \frac{b}{(1+mx)^2} y_2 & 0 & -\frac{bx}{1+mx} \\ -k\frac{b}{(1+mx)^2} y_2 & -(D + d_1) & -\frac{kbx}{1+mx} \\ 0 & -D & -d_2 \end{pmatrix}.$$

It is then seen that the matrix $CJ_{(1.1)}(x, y_1, y_2)C$ has negative off-diagonal entries for $(x, y_1, y_2) \in S$, and so the system (1.1) is competitive on S. By the previously established persistence and boundedness results, it follows that (1.1) is point dissipative. It is also easy to see that (1.1) is irreducible in S.

Since (1.1) has a unique equilibrium point $E^* = (x^*, y_1^*, y_2^*)$ and

$$\det J_{(1.1)}(x^*, y_1^*, y_2^*) = -\frac{r - ax^*}{1 + mx^*} d_2(D + d_1) < 0,$$

it follows from Theorem 1.2 in Zhu and Smith [39] that either E^* is stable, or, if it is unstable, there is at least one but no more than finitely many periodic orbits and at least one of these is orbitally asymptotically stable. Also, if E^* is stable but not globally stable, then since (1.1) is a three-dimensional competitive system, it follows from Theorem 4.1 in Smith [34, Chapter 3] that (1.1) has a periodic orbit which is necessarily orbitally unstable. Moreover, if E^* is hyperbolic and unstable with a two-dimensional unstable manifold, it follows from Theorem 4.2 in Smith [34, Chapter 3] that the ω-limit of any orbit of (1.1) which does not start on the stable manifold of E^* is a nontrivial periodic orbit. Summarizing the above discussion, one obtains the following result.

THEOREM 4.1. *Suppose that $T < kf(x_0)$ and that E^* is not globally asymptotically stable.*

1. *If either* (4.2) *or its reverse is satisfied, then E^* is hyperbolic and there is at least a nontrivial periodic orbit. The ω-limit of any orbit with positive initial data is either E^* or a nontrivial periodic orbit.*
2. *If* (4.2) *is satisfied (which happens in particular when $x^* > r/(2a)$), then the positive equilibrium E^* is locally asymptotically stable and there is at least a periodic orbit which is necessarily orbitally unstable.*
3. *If the reverse of* (4.2) *is satisfied, then the positive equilibrium E^* is unstable with a two-dimensional unstable manifold and there is at least one but no more than finitely many periodic orbits and at least one of these is orbitally asymptotically stable. Any solution which does not start on the one-dimensional stable manifold of E^* converges to a nontrivial periodic orbit.*

Unfortunately, we are not able to study analytically whether or not the periodic solutions mentioned in parts 2 and 3 above are unique.

5. Concluding remarks. First, we discuss the biological significance of our results. From the above results, we know that $T = d_2(D + d_1)/D$ is a threshold parameter for the stability of the system and that the numerical response of the mature predator plays a major role in the long-term behavior of the system (1.1). More precisely, Theorem 3.1 indicates that if the numerical response of the mature predator for the prey at carrying capacity is lower than the threshold value T, i.e., if few mature predators introduced in a predator-free ambient with prey at carrying capacity cannot reproduce fast enough, the predator classes tend to extinction. Moreover, we can define the basic reproduction number of the system by $R_0 = kf(x_0)\frac{D}{D+d_1}\frac{1}{d_2}$, and then the condition $T \geq kf(x_0)$ is equivalent to $R_0 \leq 1$. This basic reproduction number has a clear biological interpretation: the first term in R_0, $kf(x_0)$, gives the mean number of newborn predators per mature predator; the second term, $\frac{D}{D+d_1}$, gives the probability that an immature predator will survive to adulthood; and the third term, $\frac{1}{d_2}$, is simply the average lifespan of a mature predator. Subsequently, the product of these three terms yields the mean number of offspring by every predator, which is precisely the biological meaning of a basic reproduction number. A similar threshold condition for the coexistence of a predator-prey system had previously been formulated and explained by Pielou [30], among others, but had not been termed a "basic reproduction number" to the best of our knowledge.

Furthermore, if the numerical response of the mature predator for the prey at carrying capacity is higher than the threshold value T and also the prey population ultimately remains higher than another value $\underline{x} > r/(2a)$, that is, if the prey is always abundant enough, it is seen from Theorem 3.3 that the system tends to a positive steady state. We also note that if the death rate d_1 of the immature predator is negligible compared to the rate D at which the immature predators become mature predators, then the inequality $T < kf(x_0)$ becomes a very simple comparison between the death rate of the mature predators and their reproductive rate. Moreover, the stage structure affects the capability of the predator species to survive and become persistent, since it is now $(D + d_1)/D$ times easier for the predator species to become extinct, as can be seen from Theorem 3.1. This means that if it takes too much for the immature predators to mature, or the through-stage death rate of the immature predator is high (that is, D is small compared to d_1), then the total number of offspring produced during the adult stage will not be enough to compensate the total loss of immature predators and the predator classes will tend to extinction.

However, the situation where $R_0 > 1$ (or $T < kf(x_0)$) but $\underline{x} \leq r/(2a)$ is more complicated. When $x^* > r/(2a)$, we know that E^* is locally asymptotically stable, but we do not know of its global properties. This is similar for the case $x^* \leq r/(2a)$, and (4.2) holds (see Theorem 4.1). Moreover, the precise conditions for the existence and uniqueness of the periodic orbits, namely when E^* is not globally stable, are unknown under part 3 of Theorem 4.1. Therefore, we proceed to investigate further by using numerical simulations.

We use the following parameter values for all numerical simulations below: $k = 1$, $b = 1$, $m = 1$, $D = 1$, $d_1 = 0.1$, and $d_2 = 0.2$. For case 1 (see Figure 5.1), we let $r = 1$ and $a = 2$, and subsequently $x_0 = \frac{r}{a} = 0.5$, $R_0 = \frac{kbx_0}{1+mx_0}\frac{D}{d_2(D+d_1)} = 1.515 > 1$, and $x^* > \frac{r}{2a}$. Since $x^* > r/(2a)$, the positive steady state E^* is locally asymptotically stable. Numerical simulations of trajectories starting at various initial populations seem to indicate that the stability is also global for the parameter values we used. Note that, in all the figures below, the black dot located on the x-axis is E_0^*, while the other black dot is E^*. For case 2 (see Figure 5.2), we let $r = 1$ and $a = 1$ so

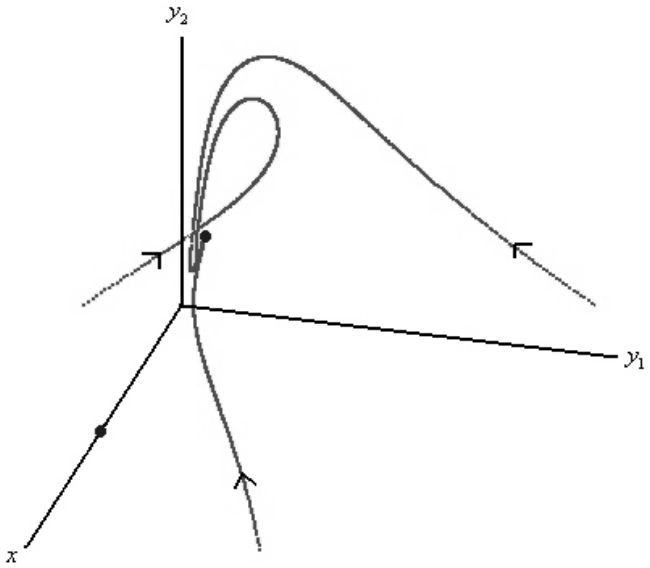


FIG. 5.1. *Simulation for case* 1 *with* $R_0 = 1.515 > 1$ *and* $x^* > \frac{r}{2a}$. *All trajectories approach* E^*.

that $x_0 = \frac{r}{a} = 1$, $R_0 = 2.273 > 1$, $x^* < \frac{r}{2a}$, and (4.2) holds. Since (4.2) holds, we know that the positive equilibrium E^* is locally asymptotically stable. Numerical simulations indicate that its stability is global. It is interesting to note that we are unable to find parameter values under which E^* satisfies (4.2), and hence it is locally asymptotically stable but not globally stable.

We also consider case 3 (see Figure 5.3), where $r = 3$ and $a = 2$, and subsequently $x_0 = \frac{r}{a} = 1.5$, $R_0 = 2.727 > 1$, and $x^* < \frac{r}{2a}$, but (4.2) does not hold. From part 3 of Theorem 4.1, we know the positive equilibrium E^* is unstable and there exists an orbitally asymptotically stable periodic orbit. Our simulation shows that this orbitally stable periodic orbit is unique and its orbital stability appears to be global. We summarize our stability results in Table 5.1. The three cases described by the last three rows of the table are illustrated with Figures 5.1–5.3, respectively. We note that, biologically, when (4.2) fails to hold and E^* becomes unstable, the coexistence of prey and predator populations is still ensured for initial populations not on the one-dimensional stable manifold of E^*, albeit with fluctuating population sizes.

We now continue with a few comments regarding the a priori estimate $\underline{x} > r/(2a)$, which was used to establish the global asymptotic stability of the positive steady state.

Let $0 < l < r/a$. It is seen that

$$x^* > l \Leftrightarrow bkDl < (1 + ml)(D + d_1)d_2,$$

from which it is easy to infer that

$$x^* > l \Leftrightarrow kf(l) < T.$$

Since $\underline{x} > l$ necessarily implies that $x^* > l$ (though this condition is only necessary and is not sufficient), it is seen that in order to have the inequality $\underline{x} > l$ satisfied, it

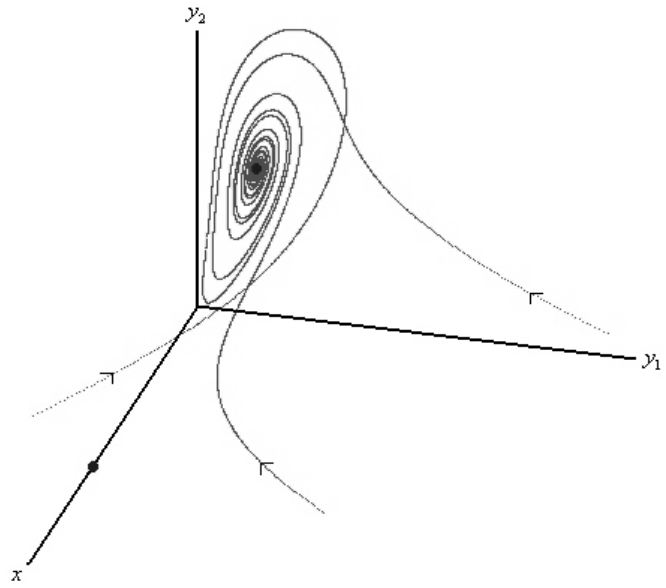


FIG. 5.2. *Simulation for case* 2 *where* $R_0 = 2.273 > 1$, $x^* < \frac{r}{2a}$, *and* (4.2) *holds. All trajectories approach* E^*.

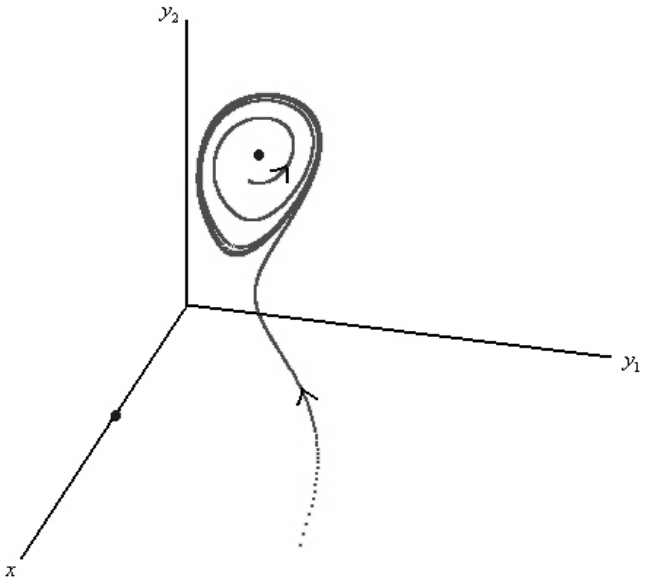


FIG. 5.3. *Simulation for case* 3 *where* $R_0 = 2.727 > 1$ *and* $x^* < \frac{r}{2a}$, *but* (4.2) *does not hold.* E^* *is unstable, and all trajectories approach an orbitally stable periodic orbit.*

TABLE 5.1

Asymptotic states of the system. "NE" denotes nonexistent, "NA" denotes not applicable, "NI" denotes no influence, "OASLC" denotes orbitally asymptotically stable limit cycle, "GAS" and "LAS" denote globally and locally asymptotically stable, respectively, and "(1)" denotes the conclusion from the simulation result.

R_0	E_0	x^*	$\underline{x}$	(4.2)	E^*	$(x, y_1, y_2) \to$
≤ 1	GAS	NE	NI	NA	NE	E_0
> 1	unstable	$> \frac{r}{2a}$	$> \frac{r}{2a}$	YES	GAS	E^*
		$> \frac{r}{2a}$	$\leq \frac{r}{2a}$	YES	LAS	$E^{*(1)}$
		$\leq \frac{r}{2a}$	$\leq \frac{r}{2a}$	YES	LAS	$E^{*(1)}$
		$\leq \frac{r}{2a}$	$\leq \frac{r}{2a}$	NO	unstable	OASLC$^{(1)}$

is necessary that $kf(l) < T$. Note that this inequality alone does not suffice to establish that $\underline{x} > l$. Again, this inequality has a certain biological interpretation. In order to have the prey population ultimately staying above a certain level l, one needs as a prerequisite that the numerical response of the predator for prey at density l be lesser than the threshold value T. Particularizing $l = r/(2a)$, it is seen that in order to obtain that $\underline{x} > r/(2a)$, one needs the inequality $kf(r/(2a)) < T$ satisfied.

Also, it is perhaps fitting to give sufficient conditions here which ensure the validity of our boundedness estimate $\underline{x} > r/(2a)$. From the first equation in (1.1), one obtains

$$(1 + mx)x'(t) = x(t)\left[(r - ax(t))(1 + mx(t)) - by_2(t)\right],$$

which implies

$$(1 + mx)x' \geq x\left[(r - b(\bar{M} + \varepsilon)) + x(rm - a) - amx^2\right]$$

for t large enough, where $\bar{M}$ is an ultimate upper bound for y_2 and $\varepsilon > 0$ is an arbitrary constant. If $r - b(\bar{M} + \varepsilon) > 0$, it follows that $\liminf_{t\to\infty} x(t) \geq x_2$, where x_2 is the positive root of

$$(r - b(\bar{M} + \varepsilon)) + x(rm - a) - amx^2 = 0.$$

From the above relations, one may deduce that $\underline{x} > r/(2a)$ whenever the following conditions are satisfied:

$$r - b(\bar{M} + \varepsilon) > 0, \quad a + \sqrt{(a - rm)^2 + 4(r - b(\bar{M} + \varepsilon))am} > 2mr.$$

Since $\varepsilon > 0$ was arbitrary, a set of conditions which ensures that $\underline{x} > r/(2a)$ is therefore

$$r > b\bar{M}, \quad a + \sqrt{(a - rm)^2 + 4(r - b\bar{M})am} > 2mr. \tag{5.1}$$

However, it is difficult to give a clear biological interpretation of the inequalities (5.1), and we would like to point out that our a priori estimate $\underline{x} > r/(2a)$ is easier to interpret and represents a theoretical device readily adaptable for the study of other systems of a certain structure, in connection with monotonicity properties. For explicit estimations of type (5.1), this sort of adjustment may not be transparent. Note that, from the discussions in section 2 on the boundedness of the solutions of

system (1.1), an ultimate upper bound for y_2 is $\bar{M} = k \max(r/a, x(0))(r+d)/d$, where $d = \min(d_1, d_2)$. See also [38] for a numerical example regarding the feasibility of the condition $\underline{x} > r/(2a)$.

Since the mature predator functional response f depends only on the size of the prey population x, our model (1.1) may be called, following the terminology given in Huisman and DeBoer [13], prey-dependent. By the same terminology, a system in which the mature predator functional response f is a function of the prey-to-predator ratio x/y is called ratio-dependent (or, more generally, predator-dependent). It is also easy to see that our model can be thought as a stage-structured version of the classical predator-prey model given below:

$$\begin{cases} x' = rx\left(1 - \dfrac{x}{K}\right) - \dfrac{bx}{1+mx}y, \\ y' = k\dfrac{bx}{1+mx}y - dy. \end{cases} \tag{5.2}$$

It is therefore not surprising that, as is easily seen from (3.2), our model inherits the structure which generates the so-called paradox of enrichment, formulated by Hairston, Smith, and Slobodkin [10] and by Rosenzweig [32], which states that increasing the carrying capacity of the environment will cause an increase in the sizes of the predator classes at equilibrium but not in that of prey. Also, since the left-hand side of (4.2) is a decreasing function of the carrying capacity r/a while the right-hand side of (4.2) is an increasing function of the same variable, it is seen that an increase in the carrying capacity may destabilize an otherwise stable positive equilibrium.

It has already been noted that all prey equilibria x^* for which $x^* > r/(2a)$ are locally asymptotically stable; that is, high prey equilibrium densities are stable. Moreover, it can also be observed that low prey equilibrium densities are unstable, since the limit of the left-hand side of (4.2) as x^* tends to 0 is also 0, while the same limit of the right-hand side of (4.2) is positive.

Note that, by the Rosenzweig–MacArthur graphical stability criterion, any equilibrium of (5.2) with $x^* > r/(2a) - 1/(2m)$ is stable, while any equilibrium of (5.2) with $x^* < r/(2a) - 1/(2m)$ is unstable. Furthermore, by Theorem 3.2 in Kuang [20], one may prove that if $\underline{x} > r/(2a)$, then (x^*, y^*) is globally asymptotically stable. One may then expect a stability threshold for (1.1) which is sharper than $r/(2a)$. Unfortunately, this result does not carry out nicely for our system (1.1) (see (4.3)). Note also that the equilibria of (1.1) with x^* close to $r/(2a) - 1/(2m)$ are unstable, as the left-hand side of (4.2) becomes arbitrarily small, while the right-hand side remains above a strictly positive lower bound.

It has also been observed in this study that, for the most part of the parameter space, the dynamical outcome does not depend on the initial population sizes and the prey and predator species cannot face extinction simultaneously. These are hallmarks of prey-dependent models, as opposed to ratio-dependent models; as seen, for instance, in Jost, Arino, and Arditi [14] or in Beretta and Kuang [3], mutual extinction may occur for ratio-dependent models, together with other rich dynamics, and the behavior of the system may depend on the initial population sizes (see also Kuang [19]). In this regard, it is believed that prey-dependent predator-prey models are more appropriate for situations in which predation involves a random or no search process, while ratio-dependent predator-prey models are more appropriate for situations in which predation involves a thorough search process. See, for instance, Kuang and Beretta [21].

Our considerations may be easily extended to systems of the form

$$\begin{cases} x' = n(x) - f(x)g(y_2), \\ y_1' = kf(x)g(y_2) - c_1 h(y_1), \\ y_2' = c_2 h(y_1) - c_3 r(y_2), \end{cases} \tag{5.3}$$

to encompass different types of functional responses from the mature predator and possible nonlinearities in the behavior of species, including nonlinearity in the predation process, under appropriate monotonicity assumptions on the functions f, g, h, r. Some examples of f and n which fit into our framework are $f(x) = mx^c$, $0 < c \leq 1$, $f(x) = m(1 - e^{-cx})$, $m, c > 0$, $f(x) = bx^p/(1 + mx^p)$, $0 < p \leq 1$ and $n(x) = x(r - ax)/(1 + \varepsilon x)$, $\varepsilon > 0$, $n(x) = rx(1 - (x/(r/a))^c)$, $0 < c \leq 1$, provided that the threshold value T and the minimal value $r/(2a)$ for $\underline{x}$ are modified accordingly. Another simple extension is to a model in which predators pass through $p > 2$ life stages, as long as the consumption of prey occurs only in the last stage. Note that the last form of $n(x)$ given above is the Richards model, a generalized logistic-type model (which simplifies to the logistic model when $c = 1$) often used to model growth of biological populations [31] or severity of disease outbreak [12].

The function n need not be monotone on its whole domain but only on $[\tilde{\underline{x}}, +\infty)$, $\tilde{\underline{x}}$ being the persistency constant of the prey for the system under consideration. In this situation, condition $\liminf_{t\to\infty} x(t) \geq \tilde{\underline{x}}$ is used to restrict n to its monotonicity domain. See Georgescu and Hsieh [8] for a related argument concerning the global stability of the endemic equilibrium for the propagation of a virus in vivo, with the remark that in [8] there is no need to impose any a priori lower bound condition, since the function which corresponds to n is monotone on the whole feasibility domain. Finally, regarding our construction of a Lyapunov function, we mention that functions of type $V(x_1, x_2, x_3, x_4) = \sum_{i=1}^{4} a_i(x_i - x_i^* \ln x_i)$, to which our function U_3 relates, have also been found useful for the study of SEIR epidemiological models. See Korobeinikov [15] and Korobeinikov and Maini [16] for details. In this regard, global stability results for models which incorporate nonlinear incidence rates of a very general form have recently been obtained by Korobeinikov and Maini in [17].

Acknowledgment. The authors would like to thank two anonymous referees for their constructive comments.

REFERENCES

[1] W. G. Aiello and H. I. Freedman, *A time-delay model of single species growth with stage structure*, Math. Biosci., 101 (1990), pp. 139–153.

[2] R. Arditi and J. Michalski, *Nonlinear food web models and their response to increased basal productivity*, in Food Webs: Integration of Patterns and Dynamics, G. A. Polis and K. O. Winemiller, eds., Chapman and Hall, New York, 1996, pp. 122–133.

[3] E. Beretta and Y. Kuang, *Global analysis in some ratio-dependent predator-prey systems*, Nonlinear Anal., 32 (1998), pp. 381–408.

[4] G. Butler, H. I. Freedman, and P. Waltman, *Uniformly persistent systems*, Proc. Amer. Math. Soc., 96 (1986), pp. 425–430.

[5] A. Fonda, *Uniformly persistent dynamical systems*, Proc. Amer. Math. Soc., 104 (1988), pp. 111–116.

[6] H. I. Freedman, S. Ruan, and M. Tang, *Uniform persistence near a closed positively invariant set*, J. Dynam. Differential Equations, 6 (1994), pp. 583–600.

[7] S. C. Gurney, R. M. Nisbet, and S. P. Blythe, *The systematic formulation of models of stage-structured populations*, in The Dynamics of Physiologically Structured Populations, Lecture Notes in Biomath. 68, Springer, Berlin, 1986, pp. 474–494.

[8] P. Georgescu and Y.-H. Hsieh, *Global stability for a virus dynamics model with nonlinear incidence of infection and removal*, SIAM J. Appl. Math., 67 (2006), pp. 337–353.
[9] S. A. Gourley and Y. Kuang, *A stage structured predator-prey model and its dependence on maturation delay and death rate*, J. Math. Biol., 49 (2004), pp. 188–200.
[10] N. G. Hairston, F. E. Smith, and L. B. Slobodkin, *Community structure, population control, and competition*, Am. Naturalist, 94 (1960), pp. 421–425.
[11] J. Hofbauer and J. W. H. So, *Uniform persistence and repellors for maps*, Proc. Amer. Math. Soc., 107 (1989), pp. 1137–1142.
[12] Y.-H. Hsieh, J. Y. Lee, and H. L. Chang, *SARS epidemiology modeling*, Emerg. Infect. Dis., 10 (2004), pp. 1165–1167.
[13] C. Huisman and R. J. DeBoer, *A formal derivation of the Beddington functional response*, J. Theoret. Biol., 185 (1997), pp. 389–400.
[14] C. Jost, O. Arino, and R. Arditi, *About deterministic extinction in ratio-dependent predator-prey models*, Bull. Math. Biol., 61 (1999), pp. 19–32.
[15] A. Korobeinikov, *Global properties of basic virus dynamics models*, Bull. Math. Biol., 66 (2004), pp. 879–883.
[16] A. Korobeinikov and P. K. Maini, *A Lyapunov function and global properties for SIR and SEIR epidemiological models with nonlinear incidence*, Math. Biosci. Eng., 1 (2004), pp. 57–60.
[17] A. Korobeinikov and P. K. Maini, *Non-linear incidence and stability of infectious disease models*, Math. Med. Biol., 22 (2005), pp. 113–128.
[18] Y. Kuang, *Basic properties of mathematical population models*, J. Biomath., 17 (2002), pp. 129–142.
[19] Y. Kuang, *Rich dynamics of Gause-type ratio-dependent predator-prey system*, Fields Inst. Commun., 21 (1999), pp. 325–337.
[20] Y. Kuang, *Global stability of Gause-type predator-prey systems*, J. Math. Biol., 28 (1990), pp. 463–474.
[21] Y. Kuang and E. Beretta, *Global qualitative analysis of a ratio-dependent predator-prey system*, J. Math. Biol., 36 (1998), pp. 389–406.
[22] J. P. LaSalle, *The Stability of Dynamical Systems*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 25, SIAM, Philadelphia, 1976.
[23] M. Y. Li and J. S. Muldowney, *Global stability for the SEIR model in epidemiology*, Math. Biosci., 125 (2005), pp. 155–164.
[24] S. Liu, L. Chen, and R. Agarwal, *Recent progress on stage-structured population dynamics*, Math. Comput. Modelling, 36 (2002), pp. 1319–1360.
[25] A. Margheri and C. Rebelo, *Some examples of persistence in epidemiological models*, J. Math. Biol., 46 (2003), pp. 564–570.
[26] J. S. Muldowney, *Compound matrices and ordinary differential equations*, Rocky Mountain J. Math., 20 (1990), pp. 857–872.
[27] R. M. Nisbet and W. S. C. Gurney, *"Stage-structure" models of uniform larval competition*, in Mathematical Ecology, Lecture Notes in Biomath. 54, Springer, Berlin, 1984, pp. 97–113.
[28] R. M. Nisbet, W. S. C. Gurney, and J. A. J. Metz, *Stage structure models applied in evolutionary ecology*, in Applied Mathematical Ecology, S. A. Levin, T. G. Hallam, and L. J. Gross, eds., Springer, Berlin, 1989, pp. 428–449.
[29] N. H. Pavel, *Differential Equations, Flow Invariance, and Applications*, Res. Notes in Math. 113, Pitman, London, 1984.
[30] E. C. Pielou, *Introduction to Mathematical Ecology*, Wiley-Interscience, New York, 1969.
[31] F. J. Richards, *A flexible growth function for empirical use*, J. Exp. Bot., 10 (1959), pp. 290–300.
[32] M. R. Rosenzweig, *Paradox of enrichment: Destabilization of exploitation systems in ecological time*, Science, 171 (1969), pp. 385–387.
[33] H. L. Smith, *A classification theorem for three dimensional competitive systems*, J. Differential Equations, 70 (1987), pp. 325–332.
[34] H. L. Smith, *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*, Math. Surveys Monogr. 41, AMS, Providence, RI, 1995.
[35] W. Wang, *Global dynamics of a population model with stage structure for predator*, in Advanced Topics in Biomathematics, L. Chen, S. Ruan, and J. Zhu, eds., World Scientific, River Edge, NJ, 1997, pp. 253–257.
[36] W. Wang and L. Chen, *A predator-prey system with stage structure for predator*, Comput. Math. Appl., 33 (1997), pp. 83–91.

[37] W. Wang, G. Mulone, F. Salemi, and V. Salone, *Permanence and stability of a stage-structured predator-prey model*, J. Math. Anal. Appl., 262 (2001), pp. 499–528.
[38] Y. N. Xiao and L. Chen, *Global stability of a predator-prey system with stage structure for the predator*, Acta Math. Sin. (Engl. Ser.), 20 (2004), pp. 63–70.
[39] H.-R. Zhu and H. L. Smith, *Stable periodic orbits for a class of three-dimensional competitive systems*, J. Differential Equations, 110 (1994), pp. 143–156.

SIAM J. APPL. MATH.
Vol. 67, No. 5, pp. 1396–1417

SPATIOTEMPORAL SYMMETRIES IN THE DISYNAPTIC CANAL-NECK PROJECTION*

MARTIN GOLUBITSKY†, LIEJUNE SHIAU‡, AND IAN STEWART§

Abstract. The vestibular system in almost all vertebrates, and in particular in humans, controls balance by employing a set of six semicircular canals, three in each inner ear, to detect angular accelerations of the head in three mutually orthogonal coordinate planes. Signals from the canals are transmitted to eight (groups of) neck motoneurons, which activate the eight corresponding muscle groups. These signals may be either excitatory or inhibitory, depending on the direction of head acceleration. McCollum and Boyle have observed that in the cat the relevant network of neurons possesses octahedral symmetry, a structure that they deduce from the known innervation patterns (connections) from canals to muscles. We rederive the octahedral symmetry from mathematical features of the probable network architecture, and model the movement of the head in response to the activation patterns of the muscles concerned. We assume that connections between neck muscles can be modeled by a "coupled cell network," a system of coupled ODEs whose variables correspond to the eight muscles, and that this network also has octahedral symmetry. The network and its symmetries imply that these ODEs must be equivariant under a suitable action of the octahedral group. It is observed that muscle motoneurons form natural "push-pull pairs" in which, for given movements of the head, one neuron produces an excitatory signal, whereas the other produces an inhibitory signal. By incorporating this feature into the mathematics in a natural way, we are led to a model in which the octahedral group acts by signed permutations on muscle motoneurons. We show that with the appropriate group actions, there are six possible spatiotemporal patterns of time-periodic states that can arise by Hopf bifurcation from an equilibrium representing an immobile head. Here we use results of Ashwin and Podvigina. Counting conjugate states, whose physiological interpretations can have significantly different features, there are 15 patterns of periodic oscillation, not counting left-right reflections or time-reversals as being different. We interpret these patterns as motions of the head, and note that all six types of pattern appear to correspond to natural head motions.

Key words. vestibular system, Hopf bifurcation, spatiotemporal symmetries, coupled cell systems

AMS subject classifications. 92C20, 37G40, 34C25

DOI. 10.1137/060667773

1. Introduction.

The human vestibular system is a system of tubes that contain sensors for motion and orientation in space, yielding the sense of balance. There are two main components: the otolith organs, which sense linear acceleration of the head (translation), and the semicircular canals, which sense angular acceleration of the head (rotation). Each ear contains three semicircular canals (henceforth "canals") arranged in three approximately mutually orthogonal planes; see Figure 1 below. A similar arrangement occurs in most vertebrates. We do not discuss the otolith system or other physiological features of the sense of balance.

In this paper we focus on two points. First, we rederive the symmetry group Γ of the network of neurons that conveys signals from the six canals to eight principal

*Received by the editors August 18, 2006; accepted for publication (in revised form) March 14, 2007; published electronically July 20, 2007. This work was supported in part by NSF grants DMS-0244529 and DMS-0604429.

http://www.siam.org/journals/siap/67-5/66777.html

†Department of Mathematics, University of Houston, Houston, TX 77204-3008 (mg@uh.edu).

‡Department of Mathematics, University of Houston-Clear Lake, Houston, TX 77058 (shiau@cl.uh.edu).

§Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK (ins@maths.warwick.ac.uk). The work of this author was supported in part by a grant from EPSRC.

muscle groups that control the position of the neck. McCollum and Boyle [12] analyzed experimental work of Shinoda et al. [13, 14, 15] and Wilson and Maeda [16] to discover these symmetries. Our derivation makes transparent the fact that Γ is the 48-element symmetry group of the cube, which is called the *octahedral* group. This network of connections is known as the *canal-neck projection*.

Second, we assume that the octahedral group Γ is also the symmetry group of the internal dynamics of the muscles and associated neural connections, and we use these symmetries to discuss natural rhythmic head motions. We look only for small amplitude periodic head motions that can be sustained by the neck muscles alone. In particular, we assume that the sensory inputs from the canals are not relevant, except to prescribe the symmetries of the system. A similar approach has been applied previously to spatiotemporal patterns in animal locomotion; see Buono and Golubitsky [3], Collins and Stewart [4, 5], and Golubitsky et al. [10, 11]. However, in those papers the patterns of locomotion were used to infer the symmetry of the network of neurons (central pattern generator) that produced them, whereas here we infer the patterns of movement from the known symmetries of the canal-neck projection.

Our approach is straightforward but not completely standard. The work of McCollum and Boyle [12] suggests a simplest network for the motoneurons of the eight muscle groups. Although we do not know (and perhaps cannot know) an accurate differential equation model for the (abstracted) muscle motoneurons, we can presume the form that such a model will take. We use the symmetries and the network structure to answer the question: What are the spatiotemporal symmetries of small amplitude periodic solutions that can be obtained by Hopf bifurcation from a group invariant equilibrium in this class of possible models? (These periodic solutions are the ones that can most naturally exist in models near a position where the head is held fixed and upright. A more general classification of the spatiotemporal symmetries of periodic solutions, whose amplitudes are not necessarily small, can be made using the H/K Theorem [3, 9]. However, we choose to begin our classification with the more restricted problem of small amplitude periodic solutions near an upright head.)

Using a caricature of the physical actions of the muscle groups, we observe that a group invariant equilibrium corresponds to one in which the head is held fixed. Using this caricature, we can also interpret the form that the head motions will take based only on the spatiotemporal symmetries of the associated periodic solutions. In this sense our approach is model-independent; it does not depend on the particular system of ODEs. Our results provide a list, or menu, of the possible head motion types; specific models and specific parameters in the models choose from this menu and determine which solution types exist and which are stable. We do not discuss such model-dependent issues here.

In order to relate these spatiotemporal symmetries to characteristic head motions, we need to make assumptions about how the eight muscle groups move the head. For physiological and mathematical reasons we are led to classify the eight muscle groups into four opposing pairs. When both muscles in a pair are equally activated the head will not move. Indeed, to move the head, one muscle group must pull harder than the opposing one; we classify only those periodic states that satisfy this constraint.

To analyze the possible dynamics we employ the theory of dynamical systems with symmetry, which has implications for the dynamics of such a network. We restrict ourselves to classifying those types of head motions that can be described by small amplitude periodic states near a group invariant equilibrium. The mathematical tool for performing this classification is the equivariant Hopf bifurcation theorem [8, 9]. In particular, we use the results of Ashwin and Podvigina [1] on Hopf bifurcation with

octahedral symmetry.

This classification is "model-independent" in the sense that it does not depend upon the detailed structure of the network of neurons concerned, or on the precise equations used to model neurons, provided that the symmetry constraints are respected. Since all model equations in current use are primarily phenomenological, and the precise architecture of the muscle group network is unknown (even in the cat), model-independent results have a potential advantage: they depend only on the known symmetries of the network. Any specific choice of network architecture and model neuron dynamics (associated, for example, with particular vertebrate species) will generate a list of spatiotemporal patterns taken from the general classification, but with extra model-dependent restrictions on existence and stability. The model-independent features of the problem can also help to structure existence and stability calculations in specific models; see [9].

In order to create this menu and to make predictions about head motions, we must determine the appropriate "phase space" variables upon which the group Γ acts, and also specify the appropriate group action. Our approach, as in the gaits work, is to use the network structure. We assume that each of the eight motoneurons (or more precisely, sets of motoneurons) is identified with variables in $\mathbf{R}^\ell$ so that the phase space of the muscle motoneurons is $Y = (\mathbf{R}^\ell)^8$. We also assume that the octahedral group acts on Y by permuting the coordinates, just as that group permutes the vertices of the cube. Next we assume that the differential equations that describe the time evolution of this coupled system of neck motoneurons have octahedral symmetry. Using this symmetry, we can then classify the types of spatiotemporal symmetries that periodic states of such systems may have.

Specifically we find that there are six types of spatiotemporal symmetries for small amplitude periodic solutions that can bifurcate from a Γ-invariant equilibrium. Each of these symmetry types includes a reasonable pattern of periodic head motion. They are: shaking the head (saying "no" in many cultures), which occurs in two different ways; nodding the head (saying "yes" in those same cultures); a rotating wave in which the head rolls in an approximate horizontal circle; a combination of "yes" and "no," in which the head nods alternately to left and right; and a side-to-side motion with the head rotating to move the nose in the opposite direction (so that the nose always points at a fixed point in the distance).

Organization of the paper. In section 2 we give a brief description of salient features of the physiology of the vestibular system and rederive the octahedral symmetry of the canal-neck projection. We relate the associated network architecture to a graph drawn on a cube and describe a simple caricature of the effects of the eight muscle groups. Section 3 describes the octahedral group in more detail and motivates the choice of action of this group on muscle space. This section also provides an explicit description of the permutation action of the octahedral group on muscle space, lists the relevant subgroups, and classifies the isotropy subgroups—basic data for the application of symmetric dynamics.

The equivariant Hopf theorem is described in section 4, and a discussion of the irreducible representations of the octahedral group, the basic information needed for application of the Hopf theorem, is given. (Proofs, which use character theory, are postponed to the appendix.) Section 5 presents a classification of the possible small amplitude spatiotemporal symmetry patterns for time-periodic motions of the head, determined by the canal-neck projection. We find six distinct (conjugacy classes of) patterns, or 15 distinct patterns (not distinguishing time-reversals or left-right

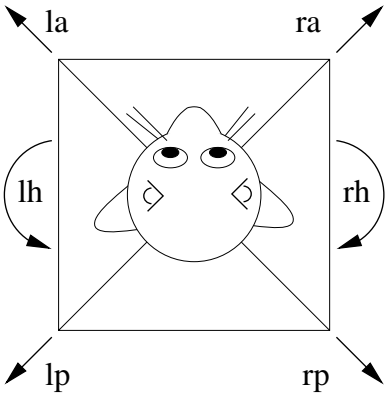


FIG. 1. *Location of the three planes relative to the head, and direction of rotational motion to which canals respond. Canals are drawn schematically near the ears.*

reflections in physical space). These patterns are interpreted as motions of the head in section 6, assuming that the muscle groups act according to our caricature. We pay attention to distinctions arising from conjugate states.

We end with a short conclusions section.

2. Symmetries in the disynaptic canal-neck projection. In this section we rederive the symmetries in the disynaptic canal-neck projection discussed by McCollum and Boyle [12], stating the results in terms of a group of permutations acting on the associated network of neurons. In this aspect of the vestibular system there are six canals (three in each ear) that are connected to eight muscle groups in the neck.

The three canals located in each ear are called *horizontal* h, *anterior* a, and *posterior* p. We denote the six canals by lh, la, lp, rh, ra, rp, where l stands for *left* and r for *right*. Neurons associated with canal hairs have a base firing rate. These hairs are arranged so that fluid flow in one direction in the canal increases the firing rate, and fluid flow in the opposite direction decreases that firing rate. Moreover, the canals are paired ({lh,rh}, {la,rp}, {lp,ra}), so that when one member of a pair transmits an elevated signal, then the other member of that pair transmits a reduced one. These pairs are called *polarity pairs*.

The spatial arrangement of the canals is shown in Figure 1. There are three (approximately) mutually orthogonal planes. One of these planes is horizontal; the other two are vertical, at an angle of 45° to the plane of left-right symmetry of the head. Each polarity pair consists of two canals that are parallel to one of these planes: one canal in the left ear, one in the right. These two canals are oriented in opposite directions in that plane and detect rotations (actually angular accelerations) of the head about an axis perpendicular to that plane. One member of the polarity pair detects acceleration in one orientation (clockwise or counterclockwise), and the other member detects the opposite orientation, as illustrated by the arrows in Figure 1. The four arrows at the corners represent rotations in the direction "along the arrow and down." For example, ra responds to motion in which the nose and right ear move forward to the left and down, while lp responds to motion in which the nose and right ear move backward to the right and up.

Connections from canals to muscles. Experiments show that each of the six canals can transmit signals to each of the eight muscle groups. The muscles also form

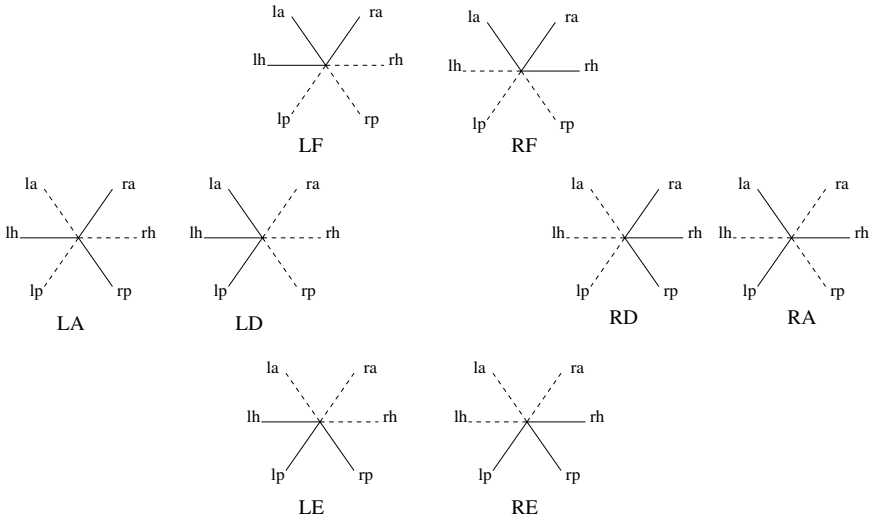


FIG. 2. *Innervation patterns corresponding to eight muscle groups. Dashed lines represent excitatory connections and solid lines inhibitory ones.*

four polarity pairs; if a canal is activated by the motion of the head, then it sends an excitatory signal to one member of each pair and an inhibitory signal to the other member. Physiological investigations (Wilson and Maeda [16], Shinoda *et al.* [13, 14, 15]) suggest that each muscle group is excited by a set of three mutually orthogonal canals (that is, one from each polarity pair) and inhibited by the complementary set of canals (the other members of the polarity pairs).

We describe the details of this arrangement, following McCollum and Boyle, who depict the list of signals transmitted to a given muscle group by an "asterisk," (Figure 2). Each asterisk has three solid lines (inhibitory connections) and three dotted lines (excitatory connections), and diametrically opposite lines have opposite polarity. There are eight possible arrangements of this type. Because the asterisks are drawn in 2-dimensional projection, in a conventional orientation with lh between la and lp, there appear to be two kinds of asterisks: two alternating (with excitation and inhibition alternating) and six nonalternating (with three contiguous excitatory canals). We will shortly see that under a suitable action of the octahedral group, all eight asterisks are equivalent.

The eight neck muscles consist of two flexors in the front (LF, RF), two extensors in the back (LE, RE), and four side (shoulder) muscles. The side muscles are alternating (LA, RA) or directed (LD, RD). McCollum and Boyle [12] discuss the innervation patterns between canal neurons and muscle motoneurons—how the six canal neurons connect to the eight muscle motoneurons, and whether the connection occurs via an excitatory synapse or an inhibitory one. The pattern of connections to each muscle is specified by Figure 2. Each asterisk in Figure 2 is a list of the connections from all six canals to one muscle group, and the type of signal that is transmitted along each connection. Observe that the muscle groups also partition into four polarity pairs:

$$\{LA, RA\}, \quad \{LF, RE\}, \quad \{LE, RF\}, \quad \{LD, RD\}.$$

If one muscle in a polarity pair has an excitatory connection from a canal, then the other muscle in that polarity pair has an inhibitory connection from that canal.

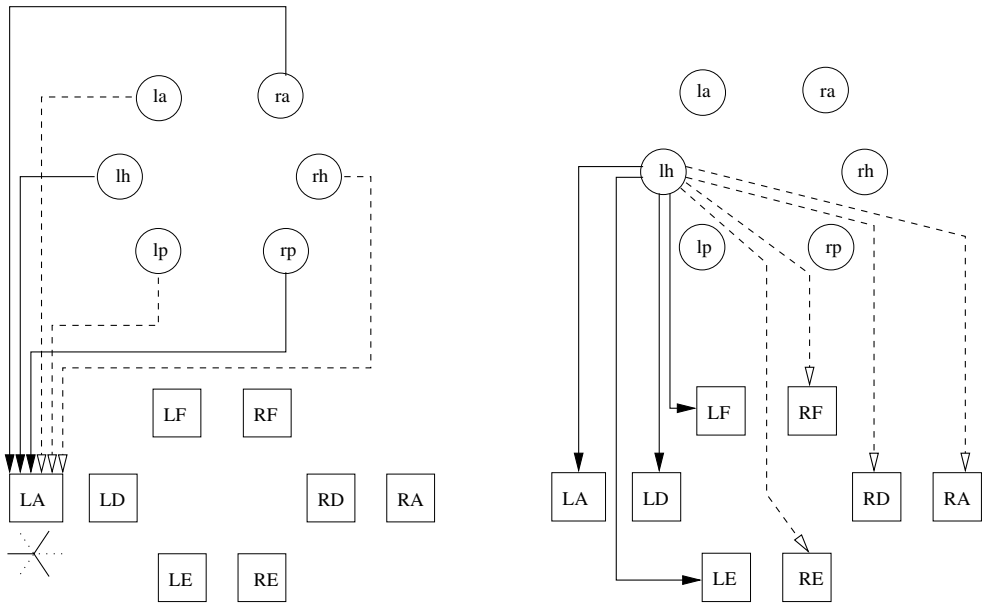


FIG. 3. *Schematic of connections from vestibular nerve afferent to neck motoneuron. Solid line shows inhibitory synapse, dotted line shows excitatory synapse. Left: Connections to a given neck motoneuron, here LA. Right: Connections from a given vestibular nerve afferent, here lh.*

It is useful to display the same information in two other ways. McCollum and Boyle [12] consider only the *disynaptic pathway* from the six vestibular nerve afferents ("canal nerves") to the eight neck motoneurons (by way of the corresponding vestibulospinal neurons). They remark that almost always "the motoneurons of each tested muscle responded to stimulation of all six canal nerves." The responses were classified as either inhibitory or excitatory, as indicated by solid or dotted lines for the relevant arm of the asterisk. This description makes it clear that their Figure 3 (and our Figure 2) is a diagram determining these *connections*.

We make the connection pattern explicit. Figure 3(left) shows connections to a given neck motoneuron, here LA. The associated asterisk is drawn, and the six connections correspond to the six arms. Figure 3(right) shows connections from a given vestibular nerve afferent, here lh. These connections correspond to the eight lh arms in the different asterisks, and connect to the corresponding neck motoneurons.

We do not attempt to draw the complete network since it would contain 48 lines, 24 solid and 24 dotted, and it would be too complicated to convey useful information. However, it is convenient to employ a geometric image in which the canals are identified with the six faces of a cube, and the muscles with the eight vertices. We will describe the network connectivity using the cube.

Octahedral symmetry of canals and muscles. The cube arises naturally from the results of McCollum and Boyle [12], identifying the symmetry group of the canal-neck projection as the 48-element octahedral group. To understand their observation, we identify the canals with faces of a cube, so that polarity pairs of canals are identified with pairs of opposite faces. Up to symmetry there is only one way to make this identification.

To identify the muscles, we observe that every vertex of the cube is in the inter-

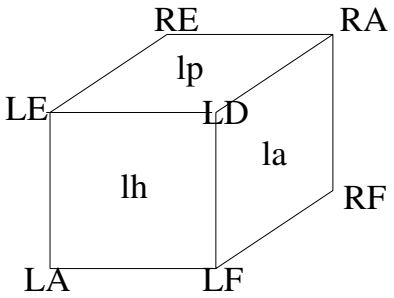


FIG. 4. *Identification of polarity pairs and muscle groups to the cube.*

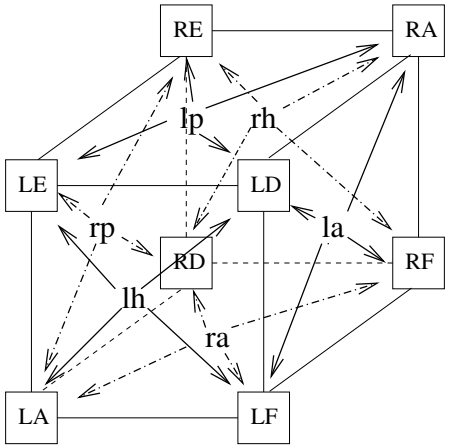


FIG. 5. *Schematic of inhibitory connections from canals to muscles drawn on the cube. Solid lines show connections on "visible" faces, and dot/dashed lines show connections on "hidden" faces. Canals are at centers of faces, muscles at vertices. Connections run to each vertex from the three adjacent faces. The octahedral symmetry of the network is obvious geometrically.*

section of exactly three faces. We identify a given vertex with that muscle that has inhibitory connections from canals corresponding to the three faces adjacent to that vertex. For example, there is a unique vertex that is in the intersection of the three faces corresponding to the left canals lh, lp, la (see Figure 4). We identify this vertex with the left direct muscle LD in Figure 2, since that muscle responds to inhibitory signals from the three left canals.

In Figure 5 we show the 24 inhibitory connections on the cube diagram. The complementary set of connections from canal neurons to muscle motoneurons consists of excitatory connections but is omitted for clarity. The octahedral symmetry of the network is apparent in this figure. The elements of the octahedral group act on the full network by permuting canals, permuting muscles, and permuting the corresponding connections.

Muscle group action: A caricature. What effect do the eight muscle groups have on the head? For purposes of interpretation, we adopt a caricature of the anatomy of the muscle groups, illustrated in Figure 6. Here we assume that the principal effect of a muscle group being activated is to pull the head in the indicated direction. Six muscle groups LF, LD, LE, RF, RD, RE effectively form a "hexagon," and their effect is to tilt the head in various directions. The other two, LA and RA,

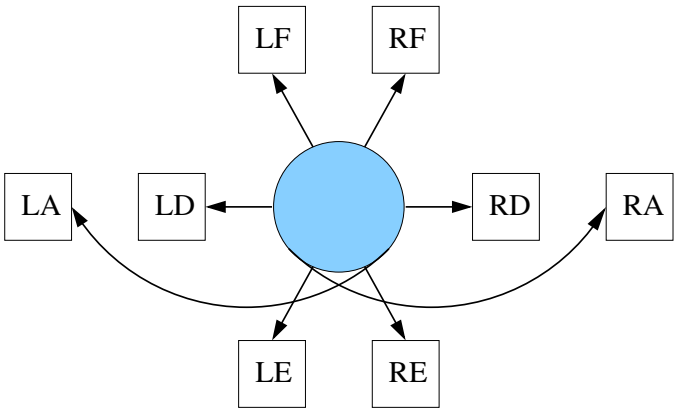


FIG. 6. *Caricature of effect of activation of muscle groups.*

rotate the head about the vertical axis (as sensed by the horizontal canals lh, rh). There is some redundancy here: the hexagon includes three pairs of muscle groups, but the three associated directions are linearly dependent. However, the use of six muscles makes the head position more stable, so there may be physiological reasons for this redundancy. McCollum and Boyle [12] call this hexagon the "central dial." This caricature exhibits the four pairs of opposing muscles (LD, RD), (LE, RF), (LF, RE), (LA, RA), which are just the four polarity pairs.

We stress that this picture of the anatomy is a caricature. At this stage we make no attempt to formulate a more realistic model of the physiology and the mechanics of head movement. However, further detail of this kind could be developed without changing the classification of possible symmetry types of time-periodic motion. What would change would be the fine detail of the corresponding head motions and the precise manner in which each muscle group contributes to that motion.

3. The octahedral group and its actions. We now discuss mathematical features of the octahedral group and various actions of that group that occur in this analysis. In particular, we introduce variables that model the state of the eight muscle groups and discuss how the octahedral group acts on those variables.

The geometry of Figure 5, together with the corresponding figure for excitatory connections (which has the same symmetry), shows that the network of neurons forming the canal-neck projection has octahedral symmetry, where now the octahedral group acts by permuting the eight muscle groups, the six canal neurons, and the connections between them. These permutation actions are distinct from, but induced naturally by, the "standard" action as isometries of $\mathbf{R}^3$ that preserve the cube.

Suppose we fix the cube so that it is centered at the origin. Then the symmetries of the cube have the form R or $-R$, where R is a rotation. It follows that the octahedral group is the direct sum of the group $\mathbb{O}$ of rotation symmetries of the cube and the two-element group $\mathbf{Z}_2^c$ generated by the inversion $-I$. That is, the octahedral group is $\mathbb{O} \oplus \mathbf{Z}_2^c$. The "$c$" in the notation $\mathbf{Z}_2^c$ indicates that this group is the center of the octahedral group.

We are modeling the canal-neck projection by a network of interconnecting neurons, following Figures 3 and 5. This network has symmetry group $\mathbb{O} \oplus \mathbf{Z}_2^c$, which acts on the network by permuting the set of cells and the set of arrows. This permutation action preserves the type of the cell (canal neuron, shown as a circle, or muscle

motoneuron, shown as a square), and it preserves the type of arrow (inhibitory or excitatory).

Phase space for muscles. This permutation action can be transferred to the dynamical variables representing the states of the cells, that is, the phase space of the network. We now describe the effect of this action on the 8 muscle cells. In order to do this we order the eight vertices as in (3.1). The ordering is shown on the cube in Figure 7.

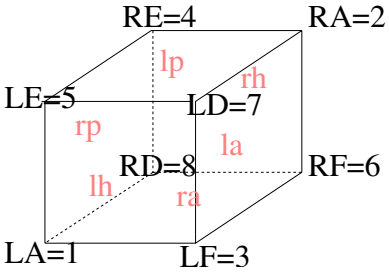


FIG. 7. *Ordering of vertices.*

The simplest model for a state of the eight muscle motoneurons is a point in $\mathbf{R}^8$, with coordinates

$$(y_{\mathrm{LA}}, y_{\mathrm{RA}}, y_{\mathrm{LF}}, y_{\mathrm{RE}}, y_{\mathrm{LE}}, y_{\mathrm{RF}}, y_{\mathrm{LD}}, y_{\mathrm{RD}}). \tag{3.1}$$

Each element of $\mathbb{O} \oplus \mathbf{Z}_2^c$ permutes the eight subscripts LA, RA, LF, RE, LE, RF, LD, RD according to the associated transformation of the vertices of the cube in Figure 7. The overall phase space for any system of ODEs representing the dynamics of the network, consistent with the $\mathbb{O} \oplus \mathbf{Z}_2^c$ symmetry, is therefore equivariant for the permutation action of $\mathbb{O} \oplus \mathbf{Z}_2^c$ on the space $\mathbf{R}^8$.

Our goal is to determine the spatiotemporal symmetries of small amplitude periodic solutions that can be obtained from a synchronous equilibrium by Hopf bifurcation. An important step in this analysis is the computation of the irreducible representations of the symmetry group $\mathbb{O} \oplus \mathbf{Z}_2^c$ on $(\mathbf{R}^\ell)^8$. However, up to isomorphism the answer for general ℓ is identical to the case when $\ell = 1$, for the following reason. Because the group acts by permutations (see the next subsection), the action on $(\mathbf{R}^\ell)^8$ consists of ℓ isomorphic copies of the action on $\mathbf{R}^8$. So the isomorphism types of the irreducible components are the same for all ℓ. However, their multiplicities depend on ℓ. We return to this point in section 4.

The action of $\mathbb{O} \oplus \mathbf{Z}_2^c$ is determined by that of $\mathbb{O}$ and that of $\mathbf{Z}_2^c$. A crucial feature of the "cube" structure is that the action of $-I$ preserves polarity pairs {LA, RA}, {LF, RE}, {LE, RF}, and {LD, RD}, because they label pairs of opposite vertices. Their entries are interchanged by $-I$. That is, $-I$ acts as the permutation

$$(1\ 2)(3\ 4)(5\ 6)(7\ 8).$$

Permutation action of $\mathbb{O}$ on muscle space. It remains to analyze the action of $\mathbb{O}$. We begin the discussion of the action of $\mathbb{O}$ on $\mathbf{R}^8$ by describing how it acts by rotations on $\mathbf{R}^3$ in the "cube" picture. There are three types of rotation: rotations about axes connecting centers of opposite faces, rotations about axes connecting midpoints of opposite edges, and rotations about axes containing opposite vertices. There

TABLE 1
Permutation actions on $\mathbf{R}^8$ *of rotations in* $\mathbb{O}$.

I	I
V_1	(3 8 5)(4 7 6)
V_2	(3 5 8)(4 6 7)
V_3	(1 7 6)(2 8 5)
V_4	(1 6 7)(2 5 8)
V_5	(1 4 7)(2 3 8)
V_6	(1 7 4)(2 8 3)
V_7	(1 4 6)(2 3 5)
V_8	(1 6 4)(2 5 3)
A_1	(1 7)(2 8)(3 5)(4 6)
A_2	(1 4)(2 3)(5 8)(6 7)
A_3	(1 6)(2 5)(3 8)(4 7)

F_1	(1 5 7 3)(2 6 8 4)
F_2	(1 3 7 5)(2 4 8 6)
F_3	(1 5 4 8)(2 6 3 7)
F_4	(1 8 4 5)(2 7 3 6)
F_5	(1 8 6 3)(2 7 5 4)
F_6	(1 3 6 8)(2 4 5 7)
E_{13}	(1 2)(3 7)(4 8)(5 6)
E_{14}	(1 5)(2 6)(3 4)(7 8)
E_{15}	(1 2)(3 4)(5 7)(6 8)
E_{16}	(1 3)(2 4)(5 6)(7 8)
E_{35}	(1 8)(2 7)(3 4)(5 6)
E_{45}	(1 2)(3 6)(4 5)(7 8)

TABLE 2
The 10 *nonidentity subgroups of* $\mathbb{O}$ *up to conjugacy, with generators.*

Subgroup	Order	Generators	Normalizer
$\mathbf{Z}_2^A$	2	A_3	$\mathbf{D}_4$
$\mathbf{Z}_2^E$	2	E_{16}	$\mathbf{D}_2^E$
$\mathbf{Z}_3$	3	V_1	$\mathbf{S}_3$
$\mathbf{D}_2^A$	4	A_1, A_3	$\mathbb{O}$
$\mathbf{D}_2^E$	4	E_{15}, E_{16}	$\mathbf{D}_4$
Subgroup	Order	Generators	Normalizer
$\mathbf{Z}_4$	4	F_3	$\mathbf{D}_4$
$\mathbf{S}_3$	6	E_{15}, V_1	$\mathbf{S}_3$
$\mathbf{D}_4$	8	A_3, F_3	$\mathbf{D}_4$
$\mathbb{T}$	12	A_3, V_4	$\mathbb{O}$
$\mathbb{O}$	24	V_4, F_5	$\mathbb{O}$

are nine rotations corresponding to faces, since there are three pairs of faces and each pair determines three nonidentity rotations. There are six rotations corresponding to edges, since there are six pairs of edges and each pair determines just one nonidentity rotation. There are eight rotations corresponding to vertices, since there are four pairs of vertices and each pair determines two nonidentity rotations.

Denote by V_j the clockwise rotation of 120° about the axis through vertex j for $j = 1, \ldots, 8$ ("clockwise" when viewed with vertex j nearest to the eye). Note that $V_1^2 = V_2$, $V_3^2 = V_4$, $V_5^2 = V_6$, $V_1^7 = V_8$. Denote by F_j the clockwise rotation of 90° about the axis perpendicular to face j for $j = 1, \ldots, 6$. Note that $F_1^3 = F_2$, $F_3^3 = F_4$, $F_5^3 = F_6$. Let $A_i = F_{2i-1}^2$ for $i = 1, 2, 3$. Then the F_j and the A_i are the nine rotations about axes connecting midpoints of opposite faces. Finally, note that each edge is uniquely the intersection of two faces. Denote by E_{ij} the rotation by 180° about the edge in the intersection of faces i and j, where $i < j$. There are six possibilities.

In Table 1 we list the 24 rotations and their permutation actions on $\mathbf{R}^8$. The entries can be read off easily from Figure 7.

Subgroups of $\mathbb{O}$. We use the following notation for groups: $\mathbf{Z}_k$ is the cyclic group of order k, $\mathbf{D}_k$ is the dihedral group of order k, $\mathbf{S}_k$ is the symmetric group of degree k, and $\mathbb{T}$ is the tetrahedral group. This is the unique subgroup of $\mathbb{O}$ that has order 12, and it fixes a tetrahedron inscribed in the cube. Table 2 lists the 11 conjugacy classes of subgroups of $\mathbb{O}$. This calculation was done using the algebra program GAP.

4. Types of Hopf bifurcation. Hopf bifurcation is the tool for finding small amplitude periodic states near an equilibrium. Equivariant Hopf theory [8, 9] states that there is a different type of Hopf bifurcation from a group-invariant equilibrium for each irreducible representation of the group. The equivariant Hopf theorem helps classify the types of spatiotemporal symmetries of periodic states that emanate from a given Hopf bifurcation. We apply this theory in the case of $\mathbb{O}\oplus\mathbf{Z}_2^c$ acting on muscle space $(\mathbf{R}^\ell)^8$, where the equilibrium is $\mathbb{O}\oplus\mathbf{Z}_2^c$-invariant. At such an equilibrium opposing muscles act with equal strength, so that the head is fixed and upright.

As noted previously, the types of irreducible representation of $\mathbb{O}\oplus\mathbf{Z}_2^c$ acting on $(\mathbf{R}^\ell)^8$ are identical with those of $\mathbb{O}\oplus\mathbf{Z}_2^c$ acting on $\mathbf{R}^8$. So the first step is to find the irreducible representations of $\mathbb{O}\oplus\mathbf{Z}_2^c$ acting on $\mathbf{R}^8$. We will show that there are four different irreducible representations, only two of which can lead to periodic states corresponding to nontrivial head motions. One of the associated Hopf bifurcations is simple to analyze, and the other was analyzed previously by Ashwin and Podvigina [1].

Decomposition of $\mathbf{R^8}$ into "push-pull" and "pull-pull" subspaces. The irreducible representations are intimately associated with the action of the inversion $-I$, which plays a key role because it swaps the members of each pair of opposing muscle motoneurons.

We can decompose $\mathbf{R}^8 = Y^+ \oplus Y^-$ into two 4-dimensional $\mathbb{O}\oplus\mathbf{Z}_2^c$-invariant subspaces, so that $-I$ acts trivially on one subspace and changes sign on the other. To do so, define

$$Y^\pm = \{y \in \mathbf{R}^8 : y_{\mathrm{LA}} = \pm y_{\mathrm{RA}}, y_{\mathrm{LF}} = \pm y_{\mathrm{RE}}, y_{\mathrm{LE}} = \pm y_{\mathrm{RF}}, y_{\mathrm{LD}} = \pm y_{\mathrm{RD}}\}.$$

Note that the coordinates corresponding to opposing muscle pairs in Y^+ are equal, and the coordinates corresponding to opposing muscle pairs in Y^- are equal in magnitude but opposite in sign.

As noted previously, three polarity pairs of muscles (the central dial) pull the head in opposite directions, and the muscles of the fourth pair (the alternating muscles) twist the head in opposite directions. In states in Y^+ polarity pairs of muscles act as pull-pull pairs, whereas in states in Y^- these polarity pairs act as push-pull pairs. In fact, all muscles must be under tension; thus push-pull pairs really operate with one muscle group pulling harder than usual while the other pulls less hard. Phenomenologically, we can identify the difference between the tensions of two muscles in a polarity pair with the deviation of the tension (of either muscle, subject to sign) from the rest tension in which the head remains upright.

Next we observe that Hopf bifurcation corresponding to an irreducible representation in Y^+ can only lead to periodic states in which the head is immobile. The reason is simple: $Y^+ = \mathrm{Fix}(-I)$, which is flow-invariant. Thus, in the nonlinear theory, any periodic state emanating from such a bifurcation must itself be fixed by $-I$; consequently, the opposing muscles in each polarity pair are always pulling with the same strength, creating a net motion of zero. As well as being inefficient, this space of motions has no visible effect on the head. In contrast, on the space Y^-, opposing pairs of muscles cooperate to move the head in exactly the same manner, so the muscle actions reinforce each other.

In fact, neither subspace Y^+ or Y^- is irreducible; each subspace decomposes into a 1-dimensional and a 3-dimensional irreducible representation. The previous remark implies that we need focus only on the subspace $Y^- \cong \mathbf{R}^4$.

Decomposition of $\boldsymbol{Y^-}$ into irreducible subspaces. As we have seen, the inversion $-I$ interchanges the muscles in each polarity pair, and the states in Y^- are

TABLE 3
Action of elements in $\mathbb{O}$ on muscle "push-pull" polarity pair space Y^-.

γ	Action on γ on Y^-
I	$(\ y_1,\ y_3,\ y_5,\ y_7)$
V_1	$(\ y_1,\ y_5, -y_7, -y_3)$
V_2	$(\ y_1, -y_7,\ y_3, -y_5)$
V_3	$(-y_5,\ y_3, -y_7,\ y_1)$
V_4	$(\ y_7,\ y_3, -y_1, -y_5)$
V_5	$(\ y_7, -y_1,\ y_5, -y_3)$
V_6	$(-y_3, -y_7,\ y_5,\ y_1)$
V_7	$(-y_5, -y_1,\ y_3,\ y_7)$
V_8	$(-y_3,\ y_5, -y_1,\ y_7)$
A_1	$(\ y_7,\ y_5,\ y_3,\ y_1)$
A_2	$(-y_3, -y_1, -y_7, -y_5)$
A_3	$(-y_5, -y_7, -y_1, -y_3)$

γ	Action on γ on Y^-
F_1	$(\ y_3,\ y_7,\ y_1,\ y_5)$
F_2	$(\ y_5,\ y_1,\ y_7,\ y_3)$
F_3	$(-y_7, -y_5,\ y_1,\ y_3)$
F_4	$(\ y_5,\ y_7, -y_3, -y_1)$
F_5	$(\ y_3, -y_5,\ y_7, -y_1)$
F_6	$(-y_7,\ y_1, -y_3,\ y_5)$
E_{13}	$(-y_1,\ y_7, -y_5,\ y_3)$
E_{14}	$(\ y_5, -y_3,\ y_1, -y_7)$
E_{15}	$(-y_1, -y_3,\ y_7,\ y_5)$
E_{16}	$(\ y_3,\ y_1, -y_5, -y_7)$
E_{35}	$(-y_7, -y_3, -y_5, -y_1)$
E_{45}	$(-y_1, -y_5, -y_3, -y_7)$

ones of the form

$$(y_{\mathrm{LA}}, -y_{\mathrm{LA}}, y_{\mathrm{LF}}, -y_{\mathrm{LF}}, y_{\mathrm{LE}}, -y_{\mathrm{LE}}, y_{\mathrm{LD}}, -y_{\mathrm{LD}}); \tag{4.1}$$

that is, we can parametrize Y^- by the strengths of the four left muscle groups, which correspond to the muscle groups numbered $1, 3, 5, 7$. Thus we can rewrite (4.1) as

$$(y_1, -y_1, y_3, -y_3, y_5, -y_5, y_7, -y_7),$$

which we parametrize by (y_1, y_3, y_5, y_7).

On Y^- the action of $\mathbb{O} \oplus \mathbf{Z}_2^c$ can now be written using signed permutations, since this action preserves polarity pairs and introduces a minus sign when members of a polarity pair are swapped. In particular, we can identify the action of $-I$ on Y^- with the signed permutation $(-y_1, -y_3, -y_5, -y_7)$; that is, $-I$ acts by multiplication by -1 on Y^-, as expected. The signed permutation action of $\mathbb{O}$ on Y^- is given in Table 3.

The subspace Y^- contains the 1-dimensional (hence irreducible) subspace

$$Y_0^- = \mathbf{R}\{(1, -1, -1, 1, -1, 1, 1, -1)\}$$

upon which the elements A_3 and V_4, the generators of the tetrahedral group, act trivially. In addition, $(\mathbb{O} \setminus \mathbb{T}, -I)$ acts trivially, since both $\mathbb{O} \setminus \mathbb{T}$ and $-I$ act as multiplication by -1.

Let Y_1^- be the 3-dimensional invariant complement of Y_0^- in Y^-; so $Y^- = Y_0^- \oplus Y_1^-$. It can be shown that Y_1^- is irreducible, and the action of $\mathbb{O}$ on Y_1^- is isomorphic to the standard action of the cube on $\mathbf{R}^3$. We do this using character theory in the appendix.

Recall that for modeling purposes we assume that the muscle state space Y^- consists of ℓ variables for each polarity pair of muscles. Thus $Y^- \cong (\mathbf{R}^\ell)^4$. As noted previously, the minimal phase space for any of our models occurs when $\ell = 1$. Although the analysis of possible spatiotemporal patterns reduces to the case $\ell = 1$, when we come to consider Hopf bifurcation, it turns out that we must require $\ell \geq 2$. (Reason: equivariant Hopf bifurcation requires certain representations to appear twice, namely, the absolutely irreducible ones, and that multiplicity occurs only when $\ell \geq 2$. See [8, 9].) Since all neurons, and in particular muscle motoneurons, have high-dimensional internal dynamics, this condition poses no difficulties.

5. Symmetry types of periodic state. At a Γ-invariant equilibrium for a Γ-equivariant system of ODEs, the equivariant Hopf theorem [8, 9] states (under several genericity hypotheses) that there exists a branch of small amplitude periodic states corresponding to every $\mathbf{C}$-axial subgroup of $\Gamma \times \mathbf{S}^1$ acting on the center subspace at that equilibrium. Moreover, these periodic states have spatiotemporal symmetries given by the $\mathbf{C}$-axial subgroup. A subgroup of $\Gamma \times \mathbf{S}^1$ is $\mathbf{C}$-*axial* if it is an isotropy subgroup, and its fixed-point subspace, within the eigenspace corresponding to the purely imaginary eigenvalues, has dimension 2.

A complete discussion of equivariant Hopf theory is beyond the scope of this paper; details can be found in [8, 9]. To simplify the remarks we make here, we assume that all periodic solutions have period 1. Then $\mathbf{S}^1$, the group of phase shift symmetries, is parameterized from 0 to 1. We now recall two general points from Hopf theory. First, the phase shift by $\frac{1}{2}$ acts as multiplication by -1 on the center subspace. Second, the kernel of the action of $\Gamma \times \mathbf{S}^1$ on the center subspace is contained in every $\mathbf{C}$-axial subgroup.

In the case at hand, we saw that $-I$ acts as multiplication by -1 on Y^-. Thus the element $(-I, \frac{1}{2})$ in $\mathbf{Z}_2^c \times \mathbf{S}^1$ acts trivially in any Hopf bifurcation with center subspace in Y^-. It follows that every periodic state emanating from such a bifurcation has the property that interchanging polarity pairs is the same as making a half period phase shift. That is,

$$\begin{aligned} y_2(t+\tfrac{1}{2}) &= y_1(t), \qquad & y_4(t+\tfrac{1}{2}) &= y_3(t), \\ y_6(t+\tfrac{1}{2}) &= y_5(t), \qquad & y_8(t+\tfrac{1}{2}) &= y_7(t). \end{aligned} \tag{5.1}$$

Dividing by the subgroup $\mathbf{Z}_2(-I, \frac{1}{2})$ leads to the standard action of $\mathbb{O} \times \mathbf{S}^1$ on the center subspace. To see this, consider the epimorphism $\varphi : (\mathbb{O} \oplus \mathbf{Z}_2^c) \times \mathbf{S}^1 \to \mathbb{O} \times \mathbf{S}^1$ defined by

$$\varphi(\gamma, I, \theta) = (\gamma, \theta) \quad \text{and} \quad \varphi(\gamma, -I, \theta) = (\gamma, \theta + \tfrac{1}{2}).$$

The kernel of φ is $\mathbf{Z}_2(-I, \frac{1}{2})$, and the quotient group is $\mathbb{O} \times \mathbf{S}^1$ with its standard action on Y^-, since $\varphi(\gamma, I, \theta) = (\gamma, \theta)$. It follows that to classify the relevant types of periodic solutions, we need analyze only those periodic solutions that occur in Hopf bifurcations associated to $\mathbb{O}$ acting on Y^- and then add in the constraints (5.1), if needed.

Using the decomposition of phase space into Y^+ and Y^- components, we can write any periodic state in the form $y(t) = y^+(t) + y^-(t)$. When we come to interpret the motions associated with the periodic states, factoring out the Y^+ component will not change these motions in any important manner since, as discussed previously, $y^+(t)$ by itself leaves the head immobile. Moreover, near these Hopf bifurcations the Y^+ components will be small compared to the Y^- components. More precisely, suppose that a Hopf bifurcation supported in Y^- leads to a periodic state of amplitude ε. Then the theory implies that generically $y^-(t)$ will be of order ε, while $y^+(t)$ will be of order ε^2. Finally, coupling (5.1) with the definitions of Y^- and Y^+ leads to the conclusion that $y^+(t)$ oscillates with twice the frequency of $y^-(t)$ and that

$$\begin{aligned} y_1^-(t+\tfrac{1}{2}) &= -y_1^-(t), \qquad & y_3^-(t+\tfrac{1}{2}) &= -y_3^-(t), \\ y_5^-(t+\tfrac{1}{2}) &= -y_5^-(t), \qquad & y_7^-(t+\tfrac{1}{2}) &= -y_7^-(t). \end{aligned} \tag{5.2}$$

In short, when discussing small amplitude periodic solutions of the nonlinear ODEs on muscle phase space, the system can effectively be reduced to an $\mathbb{O}$-equivariant

system of ODEs on the reduced phase space Y^- whose periodic solutions also satisfy (5.2). It is this reduced system that we study for the remainder of this paper.

Spatiotemporal symmetries defined by H and K. In Γ-equivariant systems we can associate two subgroups H and K of Γ to each periodic state $y(t)$. Elements of the subgroup K fix the periodic trajectory pointwise, whereas elements of the subgroup H fix the periodic trajectory setwise. Uniqueness of solutions with a given initial condition implies that each element of H couples with a phase shift to fix the periodic state.

When $\ell \geq 2$, periodic states can have spatiotemporal symmetry group pairs (H, K) only if H/K is cyclic and K is an isotropy subgroup [9]. We describe the symmetries associated with periodic states obtained by Hopf bifurcation in terms of these (H, K) pairs.

Hopf bifurcation in Y^-. Next we classify the types of periodic state that can arise as a small amplitude motion near the steady state (in which there is no head motion). Such states can be found using the equivariant Hopf theorem [8, 9]. This theorem states that there is a possible Hopf bifurcation corresponding to each irreducible representation of $\mathbb{O}$ acting on phase space. Now, the decomposition of Y^- into irreducibles can be viewed as a decomposition $\mathbf{R}^4 = W_0 \oplus W_1$, where

$$W_0 = \mathbf{R}\{(1,-1,-1,1)\} \quad \text{and} \quad W_1 = \mathbf{R}\{(1,1,1,1),(1,1,-1,-1),(1,-1,1,-1)\}.$$

Here W_0 corresponds to Y_0^-, and W_1 corresponds to Y_1^-. Both are irreducible. The kernel of the action of $\mathbb{O}$ on W_0 is $\mathbb{T}$; the representation on W_1 is the standard 3-dimensional irreducible representation, in which $\mathbb{O}$ acts as isometries that preserve the cube.

Hopf bifurcation via W_0 leads to periodic states with $H = \mathbb{O}$ and $K = \mathbb{T}$. Ashwin and Podvigina [1] classify the periodic states that arise from the standard irreducible representation of $\mathbb{O}$. This is the difficult case for Hopf bifurcation. There are five types of periodic state, whose (H, K) pairs are $(\mathbf{D}_4, \mathbf{Z}_4)$, $(\mathbf{D}_2^E, \mathbf{Z}_2^E)$, $(\mathbf{Z}_4, \mathbf{1})$, $(\mathbf{S}_3, \mathbf{Z}_3)$, and $(\mathbf{Z}_3, \mathbf{1})$. Table 4 lists these pairs, together with associated information.

We sketch the derivation of the "muscle oscillation" column of this table. Consider the pair $(H, K) = (\mathbb{O}, \mathbb{T})$ in the first row. Here $\mathbb{T}$ fixes the state of each muscle group at each time. By Table 2, the group $\mathbb{T}$ is generated by A_3 and V_4. Therefore, by Table 3, any state $y(t)$ with the symmetry pair $(\mathbb{O}, \mathbb{T})$ must satisfy

$$y_1(t) \equiv -y_5(t), \qquad y_3(t) \equiv -y_7(t), \qquad y_1(t) \equiv y_7(t),$$

so that

$$y(t) = (u(t), -u(t), -u(t), u(t))$$

for a time-periodic function u. The quotient H/K is isomorphic to $\mathbf{Z}_2$ and is generated (modulo K) by the element $(F_5, \frac{1}{2}) \in \mathbb{O} \times \mathbf{S}^1$. This imposes the same condition $u(t + \frac{1}{2}) = -u(t)$ that was previously noted using the symmetry $(-I, \frac{1}{2})$.

For a more complicated example, consider the pair $(H, K) = (\mathbf{Z}_3, \mathbf{1})$. Since K is trivial, no components of $y(t)$ are forced to be synchronous. The subgroup $\mathbf{Z}_3$ is generated by V_1, whose action on $\mathbf{R}^4$ fixes y_1 and cycles $(y_3, y_5, -y_7)$. The only possible pattern of phase shifts here is $(0, \frac{1}{3}\delta, \frac{2}{3}\delta)$, where $\delta = \pm 1$. So

$$(y_3(t), y_5(t), y_7(t)) = (u(t), u(t + \tfrac{1}{3}\delta), -u(t + \tfrac{2}{3}\delta)),$$

TABLE 4

Conjugacy classes of Hopf-type states, and the associated patterns of muscle activation, where $\delta = \pm 1$, $u(t+\frac{1}{2}) = -u(t)$, $z(t+\frac{1}{2}) = -z(t)$, $v(t+\frac{1}{6}) = -v(t)$. *Column "#" gives a series of reference numbers used for identification in the text.*

Type	H generators	K generators	Muscle oscillation	#
$(\mathbb{O}, \mathbb{T})$	V_4, F_5	V_4, A_3	$(u(t), -u(t), -u(t), u(t))$	1
$(\mathbf{S}_3, \mathbf{Z}_3)$	V_1, E_{15}	V_1	$(u(t), z(t), z(t), -z(t))$	2
	V_3, E_{14}	V_3	$(z(t), u(t), -z(t), z(t))$	3
	V_5, E_{16}	V_5	$(z(t), -z(t), u(t), z(t))$	
	V_7, E_{45}	V_7	$(-z(t), z(t), z(t), u(t))$	4
$(\mathbf{D}_2^E, \mathbf{Z}_2^E)$	E_{16}, E_{15}	E_{16}	$(u(t), u(t), 0, 0)$	5
	E_{14}, E_{13}	E_{14}	$(u(t), 0, u(t), 0)$	
	E_{45}, E_{35}	E_{45}	$(0, u(t), -u(t), 0)$	6
	E_{13}, E_{14}	E_{13}	$(0, u(t), 0, u(t))$	7
	E_{15}, E_{16}	E_{15}	$(0, 0, u(t), u(t))$	
	E_{35}, E_{45}	E_{35}	$(u(t), 0, 0, -u(t))$	8
$(\mathbf{Z}_4, \mathbf{1})$	F_1	I	$(u(t), u(t+\frac{1}{4}\delta), u(t+\frac{3}{4}\delta), u(t+\frac{1}{2}\delta))$	9
	F_3	I	$(u(t), u(t), u(t+\frac{1}{4}\delta), u(t+\frac{1}{4}\delta))$	10
	F_5	I	$(u(t), u(t+\frac{1}{4}\delta), u(t), u(t+\frac{1}{4}\delta))$	
$(\mathbf{Z}_3, \mathbf{1})$	V_1	I	$(v(t), u(t), u(t+\frac{1}{3}\delta), -u(t+\frac{2}{3}\delta))$	11
	V_3	I	$(u(t), v(t), -u(t+\frac{1}{3}\delta), u(t+\frac{2}{3}\delta))$	12
	V_5	I	$(u(t), -u(t+\frac{2}{3}\delta), v(t), u(t+\frac{1}{3}\delta))$	
	V_7	I	$(u(t), -u(t+\frac{1}{3}\delta), -u(t+\frac{2}{3}\delta), v(t))$	13
$(\mathbf{D}_4, \mathbf{Z}_4)$	F_1, A_2	F_1	$(u(t), u(t), u(t), u(t))$	14
	F_3, A_3	F_3	$(u(t), -u(t), u(t), -u(t))$	15
	F_5, A_1	F_5	$(u(t), u(t), -u(t), -u(t))$	

while $y_1(t) = v(t)$ is independent of these. However, the same phase shifts apply to y_1, so we must have $v(t) \equiv v(t + \frac{1}{3}\delta)$. Moreover, like every periodic state arising by Hopf bifurcation, v also satisfies $v(t + \frac{1}{2}) = -v(t)$. These observations lead to the condition $v(t + \frac{1}{6}) = -v(t)$ in the table.

Note that for phase shifts other than $0, \frac{1}{2}$ the states come in pairs, with plus or minus the stated phase shift. These pairs are identical except for time-reversal. For a given imaginary eigenspace, either one of these states occurs, or the other does, but not both. See [9, pp. 112–114]. (When $H/K \cong Z_m$ with $m = 5$ or $m \geq 7$ the same pair H/K can correspond to several distinct phase shifts, even taking the sign into account. For example, the $\mathbf{Z}_5$ case can have phase shift $\frac{2}{5}$ as well as $\frac{1}{5}$. However, these cases do not occur in the group $\mathbb{O}$.)

Table 4 lists (up to conjugacy) five small amplitude periodic state types that can occur by Hopf bifurcation supported by the standard 3-dimensional irreducible representation of $\mathbb{O}$, plus a sixth supported by the 1-dimensional representation. We interpret these motions in terms of our caricature of the muscle groups. We will see that all six cases lead to repetitive motions that seem quite reasonable.

Conjugate states are determined by $\mathbb{O}/(\boldsymbol{N}(\boldsymbol{H}) \cap \boldsymbol{N}(\boldsymbol{K}))$. We briefly discuss the technical issue: states whose associated subgroups are conjugate.

Suppose that $x(t)$ is a periodic state with spatiotemporal symmetry group pair (H, K). Let $\gamma \in \Gamma$. Then $\gamma x(t)$ is a periodic state with spatiotemporal symmetry group pair (H', K'), where

$$H' = \gamma H \gamma^{-1} \quad \text{and} \quad K' = \gamma K \gamma^{-1}.$$

Thus the symmetry group pairs are identical if and only if $\gamma \in N(H) \cap N(K)$. The number of conjugate periodic states with different spatiotemporal symmetries is

$$\frac{|\Gamma|}{|N(H) \cap N(K)|}. \tag{5.3}$$

When we specialize to $\Gamma = \mathbb{O}$, the number of conjugates can be found by computing the normalizers of the appropriate subgroups. The normalizers are found in Table 2. In particular, $N(\mathbb{O}) \cap N(\mathbb{T}) = \mathbb{O}$, $N(\mathbf{S}_3) \cap N(\mathbf{Z}_3) = \mathbf{S}_3$, $N(\mathbf{D}_2^E) \cap N(\mathbf{Z}_2^E) = \mathbf{D}_2^E$, $N(\mathbf{Z}_4) = \mathbf{D}_4$, $N(\mathbf{Z}_3) = \mathbf{S}_3$, and $N(\mathbf{D}_4) \cap N(\mathbf{Z}_4) = \mathbf{D}_4$. It follows that the number of conjugacies of the six solution types are 1, 4, 6, 3, 4, 3, respectively, yielding 21 possibilities.

6. Head motions. The standard equivariant theory classifies solution types up to conjugacy by a symmetry element. However, conjugate states are important here, because, with one exception, the action of $\mathbb{O}$ on the muscle space network does not relate directly to motions of the head in physical space $\mathbf{R}^3$, and that exception is the bilateral (left-right) symmetry of the body, which is realized in our network by E_{45}. So, in general, conjugate symmetry groups can correspond to head motions that are substantially different. Counting conjugates, as we have in Table 4, leads to 21 motions to describe—28 if we include time-reversals for $\pm$ phase shifts. If we consider solution types up to time-reversibility and bilateral symmetry, then there are 15 types to consider. The final column (#) in Table 4 is a reference number which we will use to identify the various patterns of oscillation.

Description of motions listed in Table 4. To explain the derivation of Table 4, we take each conjugacy class in turn and visualize the corresponding periodic state in the following manner. We assume, for simplicity, a 2-dimensional description in which the head is modeled by a circle, as in Figure 6. The position of the neck is identified with the center of this circle. The orientation of the nose (under rotation about the neck axis) is specified by a vector based at the center of the circle with the appropriate orientation.

We decompose the head motion into two distinct components. The spatial motion of the head is obtained by summing the six vectors representing the muscle groups of the central dial. As time t varies through a cycle, the resultant vector describes a closed curve in the horizontal plane, schematically representing the motion of the center of the circle that represents the head position.

Rotations of the neck (caused by muscle groups LA, RA) are represented as rotations of the circle about its instantaneous center. These rotations are assumed to act independently of the translations of the circle. This assumption is invalid in genuine 3-dimensional motion, but it provides an adequate visualization of small amplitude motions, bearing in mind that Figure 6 is itself a caricature.

Next, we choose specific periodic functions u, z, v with the correct symmetry properties. For the figures drawn here we take

$$\begin{aligned} u(t) &= \sin(2\pi t) + 0.2\sin(10\pi t), \\ z(t) &= 0.75\sin(2\pi t) + 0.03\sin(10\pi t), \\ v(t) &= 0.3\sin(6\pi t). \end{aligned}$$

Then we use Table 4 to compute the six vectors of the central dial and the rotation angle of the nose. We denote the center of the circle (head) as a function of time t by

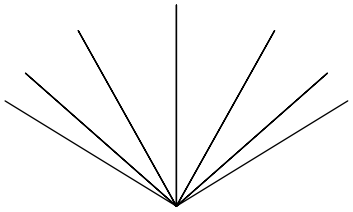


FIG. 8. *Motion for pattern* 1 $(\mathbb{O}, \mathbb{T})$.

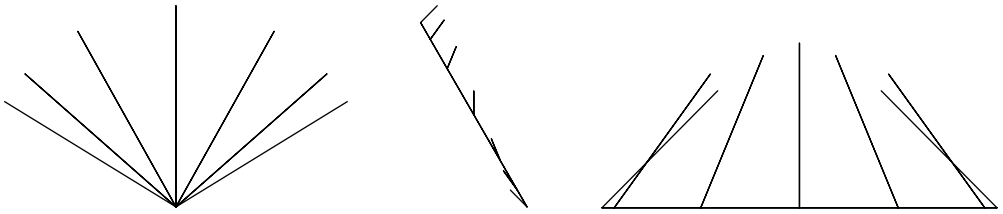


FIG. 9. *Motion for patterns* 2–4 $(\mathbf{S}_3, \mathbf{Z}_3)$.

the curve $\mathcal{C}(t)$. For a periodic state $y(t) = (y_1(t), y_3(t), y_5(t), y_7(t))$ the closed curve $\mathcal{C}(t)$ has the form

$$\begin{aligned} \mathcal{C}_1(t) &= -2y_7(t) - y_3(t) - y_5(t), \\ \mathcal{C}_2(t) &= \sqrt{3}(y_3(t) - y_5(t)). \end{aligned} \tag{6.1}$$

The term $\sin(10\pi t)$ is included to remove some artificial regularities from the pictures, such as motions of the head in a perfect circle. The vector representing the orientation of the nose is drawn at times $\frac{n}{12}$ for $0 \le n \le 11$, as a vector based on the appropriate point of the curve $\mathcal{C}$, of fixed length.

This representation involves some arbitrary choices, but is adequate for our present needs. When interpreting the figures, note that $\mathcal{C}$ may reduce to a line segment (described twice) or even a single point. Also, the segments representing the nose may overlap each other or overlap $\mathcal{C}$. These ambiguities can be resolved by creating a movie.

$(\mathbb{O}, \mathbb{T})$*:*. Here the muscles of the central dial follow the pattern $y_3(t) = y_{\mathrm{LF}}(t) = -u(t)$, $y_5(t) = y_{\mathrm{LE}}(t) = -u(t)$, $y_7(t) = y_{\mathrm{LD}}(t) = u(t)$. From (6.1) we see that the curve $\mathcal{C}$ is a single point, and the center of the head does not move. The nontrivial head motion comes from y_{LA} and y_{RA}, which swivel the head about its vertical axis. The overall invariance under $(-I, \frac{1}{2})$ implies that this swivel motion is the same as its left-right reflection, up to a half-period phase shift. This description corresponds exactly, under the assumptions of the model, to the usual "shake the head" motion indicating the word "no." The schematic visualization of this motion is shown in Figure 8. Here the nose vector oscillates from left to right to form the fan shape illustrated.

$(\mathbf{S}_3, \mathbf{Z}_3)$*:*. If we take $H = \langle V_1, E_{15} \rangle$ and $K = \langle V_1 \rangle$, then this case turns out to be exactly like the previous one, except that the time series of the direct muscled motoneurons $u(t)$ is unequal to the time series of the central dial muscle motoneurons $z(t)$. (Here angle brackets indicate the subgroup generated by their contents.) Since the $z(t)$ motions cancel out, the motion again looks like "no" and is reproduced as the first image in Figure 9.

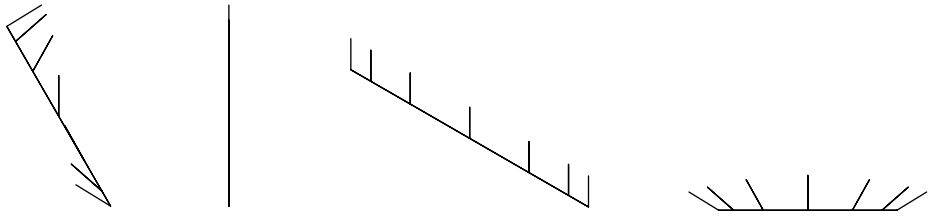


FIG. 10. *Motion for oscillation patterns* 5–8 $(\mathbf{D}_2^E, \mathbf{Z}_2^E)$.

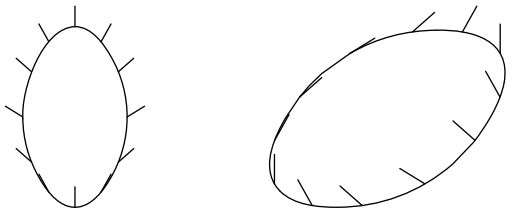


FIG. 11. *Motion for patterns* 9–10 $(\mathbf{Z}_4, \mathbf{1})$.

The patterns for the other two conjugates of this motion can be deduced in a similar manner and are visualized in Figure 9. In pattern 3 the head is inclined alternately down to the left and up to the right, while the nose oscillates from side to side. In pattern 4, the head tilts alternately to left and right while the nose oscillates from side to side.

$(\mathbf{D}_2^E, \mathbf{Z}_2^E)$*:*. First, we consider the conjugate state pattern 5, for which $K = \langle E_{16} \rangle$, $H = \langle E_{16}, E_{15} \rangle$, and $y(t) = (u(t), u(t), 0, 0)$. The phase shift action of H/K implies that $u(t + \frac{1}{2}) = -u(t)$. Now the muscle groups LD, RD, LE, RF are inactive, LA and LF are in phase with each other, and RA and RE are half a period out of phase with LA and LF. The head "nods" down and to the left, then up and to the right, in roughly the direction of the muscle pair LF, RE, with a twist to the right as the head moves down, a twist to the left as it moves up. Another conjugate state has $K = \langle E_{14} \rangle$ and $H = \langle E_{14}, E_{13} \rangle$. This state is just the left/right image of the previous one.

Second, we consider pattern 6, where $K = \langle E_{45} \rangle$ and $H = \langle E_{45}, E_{35} \rangle$. Such a state has $y_{\mathrm{LA}} = y_{\mathrm{RA}} = y_{\mathrm{LD}} = y_{\mathrm{RD}} = 0$. The variables y_{LF} and y_{LE} are half a period out of phase, and the push-pull constraint implies that y_{RF} is in synchrony with y_{LF}, and similarly y_{RE} is in synchrony with y_{LE}. There is thus an overall left-right symmetry, and also a front-back symmetry when combined with a half period phase shift. This is precisely the pattern of movement observed when nodding the head (indicating "yes"). Motions associated with patterns 7 and 8 are found similarly. Note that pattern 8 also corresponds to a standard head motion: one where the head rotates left as it tilts left and then rotates right as it tilts right. See Figure 10 for diagrams of patterns 5–8.

$(\mathbf{Z}_4, \mathbf{1})$*:*. We take $H = \langle F_1 \rangle$. From Table 3, and noting that F_1 induces a phase shift of $\pm\frac{1}{4}$, we obtain the pattern listed in Table 4. (We also use the $(-I, \frac{1}{2})$ symmetry of all periodic states.) The motions are visualized in Figure 11.

In pattern 9, the head moves in an ellipse with long axis pointing towards the front. The nose oscillates from side to side, moving outwards at the front and inwards at the back. There are two conjugates in pattern 10 that are mirror images of each

other. The motion is much as above, but the ellipse is oriented along a different axis.

$(\mathbf{Z}_3, \mathbf{1})$∴ Take $H = \langle V_1 \rangle$. This leads to the pattern stated in Table 4. Bearing in mind the $(-I, \frac{1}{2})$ symmetry, successive phases around the central dial differ by $\frac{1}{6}$. The head rotates in a "circle" (strictly, a closed loop with hexagonal symmetry), combined with a swivel. Choice of plus or minus phase shifts produce clockwise or counterclockwise rotations. Conjugates here replace V_1 by V_3, V_5, V_7, noting that V_3 and V_5 are mirror images. The motions are visualized in Figure 12. In pattern 11 the head rotates in a rounded hexagonal curve, while the nose oscillates slightly. The other two patterns are more complicated and best described using the figure.

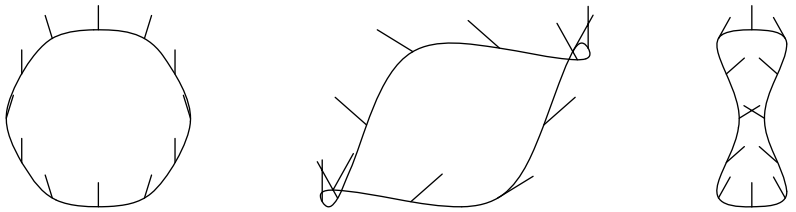


FIG. 12. *Motion for patterns* 11–13 $(\mathbf{Z}_3, \mathbf{1})$.

$(\mathbf{D}_4, \mathbf{Z}_4)$∴ We choose $H = \langle F_3, A_3 \rangle$, $K = \langle F_3 \rangle$, and $y(t) = (u(t), -u(t), u(t), -u(t))$. The conjugates are as shown in Table 4. The motions are visualized in Figure 13. In each case the head moves in one of three planes (so that $\mathcal{C}$ reduces to a line segment), while the nose oscillates from side to side. In pattern 14 the head moves left and right while the nose aims at a fixed central point.

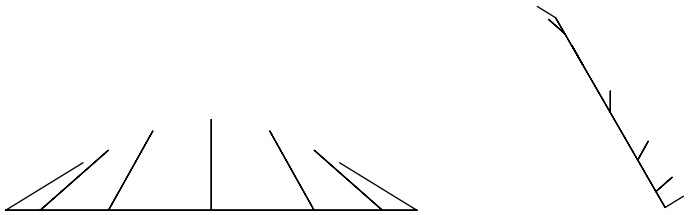


FIG. 13. *Motion for patterns* 14–15 $(\mathbf{D}_4, \mathbf{Z}_4)$.

7. Conclusions. In sections 2–4 we derived the octahedral symmetry of the canal-neck projection first discovered by McCollum and Boyle [12]. After conjecturing that the symmetry of the network of neck muscle motoneurons also has octahedral symmetry, we classified the spatiotemporal symmetry types of small amplitude periodic solutions that can be obtained by Hopf bifurcation. Finally, in section 6, we used the caricature of muscle group actions developed in section 2 to suggest the form that head motions might take.

On a cautionary note, the symmetries of neuronal networks need not reflect symmetries in the physical world. This mismatch in symmetry happens in the network associated with orientation-tuned neurons in the primary visual cortex [2] and is also the case in the vestibular system. Thus, periodic solutions that are symmetrically related in the network need not be (obviously) related in physical space (actual head movements). We believe that the issue of network structure not being directly related to physical world structure will be an important issue in many applications.

The next steps in the program we have described are to include in the model the (symmetry) structure of other projections in the vestibular system, for example the uvula-nodulus [7], to include the semicircular canals, and to make more direct contact with the biology. Two questions arise in this last step: Do the head motions we describe play some special role in the context of general periodic head motions (that is, do these motions appear frequently in animals), and can the classification of spatiotemporal symmetries of small amplitude motions near an upright head give a method for classifying types of head tremor? Our classification, which is based on the symmetries of a network that has been abstracted from the neurobiology of the cat, provides a prediction for likely types of periodic head motions, much like the predictions that were implicit in our previous work on animal gaits [10, 11].

There may exist periodic states that emanate from many types of bifurcation; however, in this study we classify only those types that emanate from a Hopf bifurcation.

Appendix: Characters of the octahedral group. The most efficient way to decompose the space Y^- into irreducible representations of $\mathbb{O}$ or $\mathbb{O} \oplus \mathbf{Z}_2^c$ is to use character theory; see, for example, Curtis and Reiner [6]. Recall that for a given representation of a group Γ, the corresponding *character* χ is the function $\chi : \Gamma \to \mathbf{C}$ for which $\chi(\gamma)$ is the trace of the matrix that represents the action of $\gamma \in \Gamma$. We assume familiarity with character theory.

First, observe that any representation (space) U for $\mathbb{O}$ naturally determines two distinct representations U^+, U^- of $\mathbb{O} \oplus \mathbf{Z}_2^c$ with the same underlying vector space. In both, the elements of $\mathbb{O}$ have the same action as they do on U. The action of $-I$ on U^+ is by the identity, whereas that on U^- is by minus the identity. If U is irreducible for the action of $\mathbb{O}$, then the $U^\pm$ are irreducible for the action of $\mathbb{O} \oplus \mathbf{Z}_2^c$.

It is easy to prove that every irreducible for $\mathbb{O} \oplus \mathbf{Z}_2^c$ arises in this manner, as follows. Every irreducible representation of $\mathbb{O}$ is absolutely irreducible, so by Schur's lemma the only commuting linear maps are scalar multiples of the identity. Since $-I$ commutes with $\mathbb{O}$ and $(-I)^2 = I$, it follows that $-I$ must act as plus or minus the identity. The rest is straightforward.

Therefore we can read off the irreducible representations of $\mathbb{O} \oplus \mathbf{Z}_2^c$ from those of $\mathbb{O}$. It is well known (see, for example, Curtis and Reiner [6, pp. 331–333]) that $\mathbb{O}$ has five distinct irreducible representations: two of dimension 1, one of dimension 2, and two of dimension 3.

We can describe the irreducible representations of $\mathbb{O}$ as follows:

- ρ_0: dimension 1; trivial action.
- ρ_1: dimension 1; $\mathbb{T}$ acts trivially, $\mathbb{O} \setminus \mathbb{T}$ acts by -1.
- ρ_2: dimension 2; kernel is the Klein four-group $\mathbf{D}_2^A$, modulo which $\mathbb{O}$ acts in the standard representation of $\mathbf{D}_3$ on $\mathbf{R}^2$.
- ρ_3: dimension 3; standard action of $\mathbb{O}$ as isometries of $\mathbf{R}^3$ preserving the cube.
- ρ_4: dimension 3; nonstandard action on $\mathbf{R}^3$ in which $\mathbb{T}$ acts as rotations but $\mathbb{O}\setminus\mathbb{T}$ acts as rotations composed with minus the identity. In fact, $\rho_4 = \rho_1 \otimes \rho_3$. This representation is also isomorphic to the standard action of $\mathbf{S}_4$ on the subspace of $\mathbf{R}^4$ consisting of points whose coordinates sum to 0.

The conjugacy classes of $\mathbb{O}$ are also five in number. In the notation of Table 3 they are

$$\{I\}, \quad \{A_j\}, \quad \{V_j\}, \quad \{F_j\}, \quad \{E_j\}.$$

The character table for $\mathbb{O}$ is shown in Table 5, and is derived in Curtis and Reiner [6, pp. 332–333]. It is easy to compute the character χ of the $\mathbb{O}$-action described in Table 3, which is shown in Table 5 in the same format. In particular, we see that $\chi = \rho_1 + \rho_3$. Since characters determine representations uniquely, and direct sums of representations correspond to sums of characters, we see that Y^- decomposes into two irreducible components, the nontrivial 1-dimensional representation and the standard 3-dimensional representation. This is what we claimed in section 4.

TABLE 5
Character table for representations of $\mathbb{O}$.

	$\{I\}$	$\{A_j\}$	$\{V_j\}$	$\{F_j\}$	$\{E_j\}$
ρ_0	1	1	1	1	1
ρ_1	1	1	1	-1	-1
ρ_2	2	2	-1	0	0
ρ_3	3	-1	0	1	-1
ρ_4	3	-1	0	-1	1
χ	4	0	1	0	-2

Acknowledgments. We thank Gin McCollum, Patrick Roberts, Douglas Hanes, David Romano, and Paul Matthews for helpful discussions. We also thank the Newton Institute, University of Cambridge, and the Department of Mathematics, University of Toronto, for their hospitality.

REFERENCES

[1] P. Ashwin and O. Podvigina, *Hopf bifurcation with cubic symmetry and instability of ABC flow*, Proc. R. Soc. Lond. A Math. Phys. Eng. Sci., 459 (2003), pp. 1801–1827.
[2] P. C. Bressloff, J. D. Cowan, M. Golubitsky, P. J. Thomas, and M. C. Wiener, *Geometric visual hallucinations, Euclidean symmetry, and the functional architecture of striate cortex*, Phil. Trans. Royal Soc. London B, 356 (2001), pp. 299–330.
[3] P. L. Buono and M. Golubitsky, *Models of central pattern generators for quadruped locomotion:* I. *Primary gaits*, J. Math. Biol., 42 (2001), pp. 291–326.
[4] J. J. Collins and I. Stewart, *Hexapodal gaits and coupled nonlinear oscillator models*, Biol. Cybern., 68 (1993), pp. 287–298.
[5] J. J. Collins and I. Stewart, *Coupled nonlinear oscillators and the symmetries of animal gaits*, J. Nonlinear Sci., 3 (1993), pp. 349–392.
[6] C. W. Curtis and I. Reiner, *Representation Theory of Finite Groups and Associative Algebras*, Wiley-Interscience, New York, 1962.
[7] I. Z. Foster, D. A. Hanes, N. H. Barmack, and G. McCollum, *Spatial symmetries in vestibular projections to the uvula-nodulus*, Biol. Cybernet., 96 (2007), pp. 439–453.
[8] M. Golubitsky and I. N. Stewart, *Hopf bifurcation in the presence of symmetry*, Arch. Ration. Mech. Anal., 87 (1985), pp. 107–165.
[9] M. Golubitsky and I. Stewart, *The Symmetry Perspective*, Progr. Math. 200, Birkhäuser-Verlag, Basel, 2002.
[10] M. Golubitsky, I. Stewart, P.-L. Buono, and J. J. Collins, *A modular network for legged locomotion*, Phys. D, 115 (1998), pp. 56–72.
[11] M. Golubitsky, I. Stewart, P.-L. Buono, and J. J. Collins, *Symmetry in locomotor central pattern generators and animal gaits*, Nature, 401 (1999), pp. 693–695.
[12] G. McCollum and R. Boyle, *Rotations in a vertebrate setting: Evaluation of the symmetry group of the disynaptic canal-neck projection*, Biol. Cybern., 90 (2004), pp. 203–217.
[13] Y. Shinoda, Y. Sugiuchi, T. Futami, N. Ando, and T. Kawasaki, *Input patterns and pathways from six semicircular canals to motoneurons of neck muscles* I: *The multifidus muscle group*, J. Neurophysiol., 72 (1994), pp. 2691–2702.
[14] Y. Shinoda, Y. Sugiuchi, T. Futami, N. Ando, and J. Yagi, *Input patterns and pathways from six semicircular canals to motoneurons of neck muscles* II: *The longissimus and semispinalis muscle groups*, J. Neurophysiol. 72 (1997), pp. 2691–2702.

[15] Y. Shinoda, Y. Sugiuchi, T. Futami, S. Kakei, Y. Izawa, and J. Na, *Four convergent patterns of input from the six semicircular canals to motoneurons of different neck muscles in the upper cervical cord*, Ann. New York Acad. Sci., 781 (1996), pp. 264–275.
[16] V. J. Wilson and M. Maeda, *Connections between semicircular canals and neck motoneurons in the cat*, J. Neurophysiol., 37 (1974), pp. 346–357.

SIAM J. APPL. MATH.
Vol. 67, No. 5, pp. 1418–1439

A METHOD TO COMPUTE STATISTICS OF LARGE, NOISE-INDUCED PERTURBATIONS OF NONLINEAR SCHRÖDINGER SOLITONS*

R. O. MOORE†, G. BIONDINI‡, AND W. L. KATH§

Abstract. We demonstrate in detail the application of importance sampling to the numerical simulation of large noise-induced perturbations in soliton-based optical transmission systems governed by the nonlinear Schrödinger equation. The method allows one to concentrate numerical Monte Carlo simulations around the noise realizations that are most likely to produce the large pulse deformations connected with errors, and yields computational speedups of several orders of magnitude over standard Monte Carlo simulations. We demonstrate the method by using it to calculate the probability density functions associated with pulse amplitude, frequency, and timing fluctuations in a prototypical soliton-based communication system.

Key words. nonlinear Schrödinger equation, optical fibers, solitons, noise, Monte Carlo simulations, importance sampling

AMS subject classifications. 35Q51m, 35Q55, 65C20, 65C05, 78A40

DOI. 10.1137/060650775

1. Introduction.

The development of high-bit-rate data transmission over optical fibers is one of the major technological achievements of the late 20th century. The information-carrying capacity of such systems has increased by several orders of magnitude over the past quarter-century. There are limits imposed on capacities, however, by various transmission impairments that distort and degrade the signal in a number of ways [1, 2]. One common source of impairments in lightwave communication systems is the amplified spontaneous emission (ASE) noise generated by the erbium-doped fiber amplifiers (EDFAs) used to compensate loss in the fiber [1, 2]. This additive noise perturbs the propagating pulses, producing amplitude, frequency, timing, and phase jitter, which can then lead to bit errors [3, 4, 5].

Since ASE noise is a stochastic phenomenon, Monte Carlo simulations can be used to determine its effects on a system. The direct calculation of bit error rates with standard Monte Carlo simulations is impossible, however. Because data transmission rates are so high (currently, 10 Gb/s or more per channel, with tens of channels usually present per fiber) and errors must be handled by much slower electronic equipment, error rates are required to be very small, typically one error per trillion bits or lower. As a result, an exceedingly large number of Monte Carlo realizations would be needed to observe even a single transmission error, and even more would be required to obtain reliable error estimates.

*Received by the editors January 24, 2006; accepted for publication (in revised form) March 30, 2007; published electronically July 20, 2007.
http://www.siam.org/journals/siap/67-5/65077.html

†Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102 (rmoore@njit.edu). The work of this author was supported in part by the National Science Foundation under grant DMS-0511091.

‡Department of Mathematics, State University of New York, Buffalo, NY 14260 (biondini@buffalo.edu). The work of this author was supported by National Science Foundation grant DMS-0506101.

§Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208-3125 (kath@northwestern.edu). The work of this author was supported by the National Science Foundation (grants DMS-0101476 and DMS-0406513) and by the Air Force Office of Scientific Research (FA9550-04-1-0289).

To overcome this limitation, one approximation is to calculate numerically the variances of pulse amplitude and/or timing and then to extrapolate the results into the tails of the probability density function (pdf) by assuming it to be Gaussian. It is clear, however, that this procedure is inadequate, since nonlinearity and pulse interactions can both contribute to make the resulting distributions non-Gaussian [6, 7, 8]. Nonlinearity arises from several sources, such as self-phase modulation due to the fiber's nonlinear refractive index [9, 10] as well as the nonlinear conversion of optical energy into electrical energy by the photodetector [8].

Various techniques have recently been proposed to address the difficulty in calculating accurate statistics for rare events [8, 11, 12, 13, 14, 15, 16]. One promising approach is a technique known as importance sampling (IS) [17, 18]. In general, IS works by concentrating Monte Carlo samples on those configurations that are most likely to lead to transmission errors, thus significantly speeding up the simulations. Previously, we have successfully applied this technique to the direct simulation of transmission impairments caused by polarization-mode dispersion [14, 15]. More recently, we have presented numerical results demonstrating that IS can also be applied to simulations of ASE-induced transmission impairments [16]. The purpose of this paper is to describe in detail the methods used to produce these numerical simulations. For simplicity, we consider the case of a soliton-based transmission system (where, in the absence of noise, the pulse shape remains fixed), but it is anticipated that the technique can be extended to more realistic systems and more general transmission formats. The advantages of the method are substantial, allowing an increase in efficiency of several orders of magnitude over standard Monte Carlo simulations.

2. Solitons and amplifier noise. The propagation of pulses in an optical fiber with periodically spaced amplification is governed by the nonlinear Schrödinger (NLS) equation [9, 10, 16], which in dimensionless units is

$$i\frac{\partial u}{\partial z} + \frac{1}{2}\frac{\partial^2 u}{\partial t^2} + |u|^2 u = i\sum_{n=1}^{N_a} \delta(z - nz_a) f_n(t). \tag{2.1}$$

Here z and t are distance and retarded time, u is the pulse's electromagnetic field envelope, N_a is the number of amplifiers, and z_a is the amplifier spacing. (The nondimensionalization and our choice of units are described in detail in Appendix A.) In this equation, the periodic power variations due to fiber loss and amplification have been averaged out [9].

The term $f_n(t)$ is the noise added at each amplifier; when the pulse reaches an amplifier at $z = nz_a$ (where z_a is the dimensionless amplifier spacing and $n = 1, 2, \ldots, N_a$, with N_a being the total number of amplifiers in the transmission line), a small amount of noise $f_n(t)$ is added to u: $u(nz_a^+, t) = u(nz_a^-, t) + f_n(t)$, as seen by integrating (2.1) across $z = nz_a$. (Recall that we have averaged out loss and gain; this means that the noise is the only effect remaining at the amplifiers.) The amplifier noise $f_n(t)$ can be modeled as classical zero-mean white noise:

$$\langle f_n(t) \rangle = 0, \tag{2.2a}$$

$$\langle f_m(t) f_n(t') \rangle = 0, \tag{2.2b}$$

$$\langle f_m(t) f_n^\dagger(t') \rangle = \sigma^2 \delta_{mn} \delta(t - t'), \tag{2.2c}$$

where $\langle \cdot \rangle$ denotes ensemble average, the superscript † denotes the complex conjugate, δ_{mn} and $\delta(t-t')$ are the Kronecker and Dirac deltas, respectively, and σ^2 is a parameter describing the noise power. Technically speaking, (2.2c) is not mathematically

correct, since as written it implies an infinite noise bandwidth and thus produces infinite noise power. Any physical system (or any numerical computation) necessarily has a finite noise bandwidth [19]. When calculating amplitude, frequency, and timing fluctuations, however, the specific value of the noise bandwidth is unimportant if it is larger than the soliton bandwidth (this is the case in practice), because only those components of the noise within the same spectral range as the soliton will affect these soliton parameters.

Without the noise term, (2.1) admits the well-known soliton solution

$$u_s(z,t) = u_0(z,t)\, e^{i\Theta(z,t)}, \tag{2.3a}$$

with

$$u_0(z,t) = A\,\mathrm{sech}[A(t-T(z))], \qquad \Theta(z,t) = \Omega\, t + \Phi(z), \tag{2.3b}$$

where $T(z) = T_0 + \Omega z$ and $\Phi(z) = \Phi_0 + \frac{1}{2}(A^2 - \Omega^2)z$ and where the four soliton parameters A, Ω, T_0, and Φ_0 are constant. When noise is added at each amplifier, part of the noise is incorporated into the soliton, where it produces small changes of the soliton parameters [9, 10]. The rest of the noise propagates along with the perturbed soliton. This process is repeated at each amplifier, resulting in a random walk of the four quantities A, Ω, T, and Φ [4, 5]. For typical system configurations, the noise amplitude at each amplifier is small, and thus the noise-induced changes of the soliton parameters at each individual amplifier are also usually small. In rare cases, however, these individual contributions combine to produce large deviations, resulting in potential transmission errors. Because these large pulse deformations are rare, estimating their probability is difficult.

2.1. Soliton perturbation theory. Soliton perturbation theory (SPT) is a standard method that has been used to approximate the effect of noise upon propagating pulses (e.g., see [9, 20, 21, 22, 23]). Rather than using it directly to obtain an analytical approximation to the perturbed pulse, however, here we will use it only as a tool to guide numerical simulations. The key information that is needed to do this comes from the dependence of the soliton solution, equation (2.3), upon the parameters A, Ω, T_0, and Φ_0. Since any value of these parameters is permitted, no resistance is encountered if the noise at an amplifier changes one of them. This lack of resistance allows large variations to build up after many amplifiers. Furthermore, frequency fluctuations change the group velocity of the pulse, and subsequent propagation integrates this velocity shift to produce a large timing shift (as reflected in the dependence of T on Ω).

The small noise-induced changes of the soliton parameters at a single amplifier can be estimated by decomposing $u(z,t)$ into its soliton and radiative (nonsoliton) components,

$$u(z,t) = [u_0(z,t) + v(z,t)]\, e^{i\Theta}, \tag{2.4}$$

and by linearizing the NLS equation (2.1) around the soliton solution (2.3):

$$\frac{\partial v}{\partial z} = L\, v, \qquad L\, v := \frac{i}{2}\frac{\partial^2 v}{\partial t^2} - \frac{i}{2}A^2 v + 2i|u_0|^2 v + i u_0^2\, v^\dagger. \tag{2.5}$$

Importantly, the linearized NLS operator L is non–self-adjoint and nonnormal, and its generalized nullspace admits four modes (localized eigenfunctions) $v_k(z,t)$ ($k = A, \Omega, T, \Phi$) satisfying [9, 10, 24]

$$L\, v_A = A\, v_\Phi, \qquad L\, v_\Omega = v_T, \qquad L\, v_T = 0, \qquad L\, v_\Phi = 0. \tag{2.6}$$

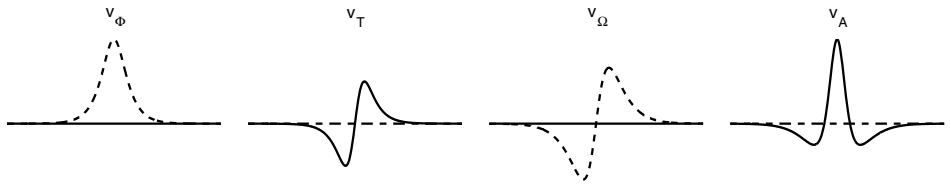


FIG. 1. *The real (solid) and imaginary (dashed) parts of the four neutral modes of the linearized NLS equation associated with the soliton solution* (2.3).

Explicitly, these four modes are

$$v_A(z,t) = \frac{1}{A}\frac{\partial}{\partial t}[(t-T)u_0], \tag{2.7a}$$

$$v_\Omega(z,t) = i(t-T)\,u_0, \tag{2.7b}$$

$$v_T(z,t) = -\frac{\partial u_0}{\partial t}, \tag{2.7c}$$

$$v_\Phi(z,t) = iu_0. \tag{2.7d}$$

Each of these linear modes is associated with a continuous symmetry of the NLS equation [23]: invariance with respect to scaling, Galilean boosts, translations, and phase rotations, respectively. Note also from (2.6) that the timing and phase modes are actual eigenfunctions, whereas the amplitude and frequency modes are generalized eigenfunctions. This is related to two of these symmetries giving rise to modified conservation laws, which are directly related to a change of the pulse frequency producing a subsequent change in its velocity or a change of the pulse amplitude producing a change in its phase upon propagation [23]. These four modes, shown in Figure 1, are also associated with changes of the soliton parameters:[1]

$$\frac{\partial u_s}{\partial A} = v_A\, e^{i\Theta}, \tag{2.8a}$$

$$\frac{\partial u_s}{\partial \Omega} = v_\Omega\, e^{i\Theta} + T v_\Phi\, e^{i\Theta}, \tag{2.8b}$$

$$\frac{\partial u_s}{\partial T} = v_T\, e^{i\Theta}, \tag{2.8c}$$

$$\frac{\partial u_s}{\partial \Phi} = v_\Phi\, e^{i\Theta}. \tag{2.8d}$$

In fact, removing secular terms from the forced linearized NLS equation obtained by adding the right-hand side of (2.1) to (2.5), one finds the local changes of the soliton parameters at the nth amplifier produced by the noise $f_n(t)$ [9, 10]:

$$\Delta A_n = \mathrm{Re}\int \underline{v}_A^\dagger(z,t)e^{-i\Theta} f_n(t)\,dt, \tag{2.9a}$$

$$\Delta \Omega_n = \mathrm{Re}\int \underline{v}_\Omega^\dagger(z,t)e^{-i\Theta} f_n(t)\,dt, \tag{2.9b}$$

$$\Delta T_n = \mathrm{Re}\int \underline{v}_T^\dagger(z,t)e^{-i\Theta} f_n(t)\,dt, \tag{2.9c}$$

$$\Delta \Phi_n = \mathrm{Re}\int (\underline{v}_\Phi(z,t) - T\,\underline{v}_\Omega(z,t))^\dagger e^{-i\Theta} f_n(t)\,dt, \tag{2.9d}$$

[1] Note that the derivatives on the left-hand side of (2.8) are taken with respect to the variables A, Ω, T, and Φ, with the other three variables kept constant. A different choice of parametrization for the soliton solution in (2.3) would lead to different results.

where the integrals are from $-\infty$ to ∞. The functions $\underline{v}_k(z,t)$ are the modes of the *adjoint* linearized NLS operator, defined by $L^{\mathrm{adj}}\underline{v} = -\frac{i}{2}\partial_t^2\underline{v} + \frac{i}{2}A^2\underline{v} - 2i|u_0|^2\underline{v} + iu_0^2\,\underline{v}^\dagger$ and the inner product $\langle \underline{v}, v\rangle = \mathrm{Re}\int \underline{v}^\dagger v\,dt$ [10]. They are

$$\underline{v}_A = -iv_\Phi, \qquad \underline{v}_\Omega = iv_T/A, \qquad \underline{v}_T = iv_\Omega/A, \qquad \underline{v}_\Phi = iv_A. \tag{2.10}$$

(The adjoint modes are easily obtained from those in (2.7), noting that $L^{\mathrm{adj}}(\underline{v}) = iL(i\underline{v})$.) Together, the modes of L and L^{adj} form a biorthonormal basis of the tangent space corresponding to infinitesimal changes in the soliton parameters at a given amplifier,[2] and the source terms in (2.9) represent the projection of the noise onto this basis.

2.2. Noise-induced amplitude, frequency, timing, and phase jitter. Equations (2.9) establish a direct projection from the infinite-dimensional noise which is added at each amplifier to a discrete random walk for the four soliton parameters. In particular, the equations can be easily integrated, including the unperturbed evolution in between amplifiers, to obtain the final values of amplitude A, frequency Ω, timing T, and phase Φ:

$$A_{\mathrm{out}} = A_0 + \sum_{n=1}^{N_a} \Delta A_n, \tag{2.11a}$$

$$\Omega_{\mathrm{out}} = \Omega_0 + \sum_{n=1}^{N_a} \Delta\Omega_n, \tag{2.11b}$$

$$T_{\mathrm{out}} = T_0 + N_a z_a \Omega_0 + \sum_{n=1}^{N_a} \Delta T_n + \sum_{n=1}^{N_a} (N_a + 1 - n)\, z_a \Delta\Omega_n, \tag{2.11c}$$

$$\begin{aligned}\Phi_{\mathrm{out}} = \Phi_0 + \frac{1}{2}N_a z_a (A_0^2 - \Omega_0^2) + \sum_{n=1}^{N_a}\Delta\Phi_n + \sum_{n=1}^{N_a}(N_a+1-n) z_a (A_0\Delta A_n - \Omega_0\Delta\Omega_n) \\ + \frac{1}{2}\sum_{n=1}^{N_a}\sum_{m=1}^{N_a}[N_a - \max(n,m)] z_a (\Delta A_n \Delta A_m - \Delta\Omega_n\Delta\Omega_m).\end{aligned} \tag{2.11d}$$

The fourth term in (2.11c) and the fourth and fifth terms in (2.11d) arise from the above-mentioned Galilean invariance of the NLS equation. Whereas amplitude and timing jitter are the most important failure mechanisms for communication lines using standard on-off keying receivers, phase fluctuations are of critical importance when the receivers are phase-sensitive, as is the case for phase-shift or differential phase-shift keying [25].

Owing to (2.2) and (2.9), ΔA_n, ΔT_n, $\Delta\Omega_n$, and $\Delta\Phi_n$ are Gaussian-distributed random variables at each amplifier, with variances

$$\langle \Delta A_{n+1}^2\rangle = A_n\sigma^2, \tag{2.12a}$$

$$\langle \Delta\Omega_{n+1}^2\rangle = \sigma^2 A_n/3, \tag{2.12b}$$

$$\langle \Delta T_{n+1}^2\rangle = \pi^2\sigma^2/(12A_n^3), \tag{2.12c}$$

$$\langle \Delta\Phi_{n+1}^2\rangle = (1 + \pi^2/12 + T_n^2)\sigma^2/3A_n, \tag{2.12d}$$

[2]Note that, as usual, a true linear mode of the linearized NLS operator corresponds to a generalized mode of the adjoint operator, and vice-versa.

respectively. Note that all of these variances depend on the value of the soliton amplitude immediately prior to arrival at the amplifier, and that the phase variance depends on the soliton position; this is because the time-dependent term in the phase in (2.3b) is not defined to be zero at $t = T(z)$. For small deviations of amplitude and position, one can approximate these variances with their initial values (assuming without loss of generality that the initial position is zero):

(2.13)
$$\sigma_A^2 := A_0\sigma^2, \quad \sigma_\Omega^2 := \sigma^2 A_0/3, \quad \sigma_T^2 := \pi^2\sigma^2/(12A_0^3), \quad \sigma_\Phi^2 := (1+\pi^2/12)\sigma^2/3A_0.$$

The variances of the final soliton amplitude, frequency, and position timing are then easily computed to be

$$\langle A_{\text{out}}^2\rangle \simeq N_a\sigma_A^2, \tag{2.14a}$$

$$\langle \Omega_{\text{out}}^2\rangle \simeq N_a\sigma_\Omega^2, \tag{2.14b}$$

$$\langle T_{\text{out}}^2\rangle \simeq N_a\sigma_T^2 + N_a(N_a+1)(2N_a+1)\,\sigma_\Omega^2\, z_a^2/6, \tag{2.14c}$$

$$\langle \Phi_{\text{out}}^2\rangle \simeq N_a\sigma_\Phi^2 + N_a(N_a+1)(2N_a+1)\,\sigma_A^2\, z_a^2/6, \tag{2.14d}$$

respectively. The cubic dependence on N_a of the growth of $\langle T_{\text{out}}^2\rangle$ and $\langle \Phi_{\text{out}}^2\rangle$ is a discrete analogue to the cubic dependence on distance in a distributed noise approximation, used by Gordon and Haus and by Gordon and Mollenauer, respectively, to derive upper limits for the error-free propagation distance of a soliton transmission system [4, 5].

These calculations, however, are not sufficient to give an accurate estimate of noise-induced transmission penalties, for several reasons. First, the variances in (2.14) are correct only for small deviations of the pulse amplitude. Second, even though the noise is Gaussian-distributed, the full noise-induced changes of the soliton parameters are not necessarily Gaussian. In particular, the variance of each amplitude shift depends on the previous value of the amplitude, which causes the distribution of A to deviate significantly from Gaussian. A Gaussian approximation will therefore be valid only in the limit of exceedingly small amplitude shifts, and quite possibly only in the core region of the pdf and not in the tails. The timing T and frequency Ω also deviate very slightly from Gaussian due to the local dependence of their variances on A (cf. (2.12); see also [13]). Since T, Ω, and Φ have no influence on the random walk of A (or on T, in the case of the phase), however, their statistical behavior is expected to be dominated by the mean value of A. Finally, even if the noise-induced changes of the soliton parameters were approximately Gaussian-distributed, calculating the probability densities in the tails from the (analytically or numerically obtained) variances would require an exponential extrapolation, and any errors or uncertainties would be magnified correspondingly.

3. Importance sampling. The main idea behind importance sampling is to bias the probability distribution functions used to generate the random Monte Carlo samples so that errors occur more frequently than would be the case otherwise [17, 18]. Before we delve into the implementation of importance sampling for amplifier noise, let us briefly present the basic ideas in a general setting.

Let X denote a collection of random variables (RVs) identifying a particular system realization. (In our case, X will be a vector or matrix which determines the noise at all the amplifiers.) Consider a measurable quantity $y(X)$ associated with each system configuration and, therefore, with each value of X. (In our case, $y(X)$ will be the final pulse amplitude or timing.) Suppose that we are interested in calculating

the probability P that $y(X)$ falls in some prescribed range. This probability can be represented as the expectation value of an *indicator function* $I(y(X))$ such that $I(y) = 1$ if the random variable y falls in the prescribed range and $I(y) = 0$ otherwise. That is, the probability P is represented by the multidimensional integral

$$P = \int I(y(x))p_X(x)\,dx = \mathbb{E}[I(y(X))], \tag{3.1}$$

where $p_X(x)$ is the joint pdf of the RVs X, $\mathbb{E}[\,\cdot\,]$ denotes the expectation value with respect to the distribution $p_X(x)$, and the integral is over the entire configuration space. In many cases of interest, a direct evaluation of the integral in (3.1) is impossible. One can then resort to Monte Carlo simulations and write an estimator $\hat{P}$ for P, replacing the integral in (3.1) by

$$\hat{P}_{\rm mc} = \frac{1}{M}\sum_{m=1}^{M} I(y(X_m)), \tag{3.2}$$

where M is the total number of Monte Carlo samples, and where each X_m is a random sample drawn from $p_X(x)$. Equation (3.2) simply expresses the relative number of samples falling in the range of interest. If one is interested in low probability events, however (that is, if $P \ll 1$), an impractically large number of samples is usually necessary in order to see even a single event, and an even larger number is required in order to obtain an accurate estimate.

When the main contribution to P comes from regions of sample space where $p_X(x)$ is small, IS can be used to speed up the simulations. The idea is first to rewrite the the probability P in (3.1) as

$$P = \int I(y(X))\,r(x)p^*(x)\,dx = \mathbb{E}^*[I(y(X))r(X)], \tag{3.3}$$

where $\mathbb{E}^*[\,\cdot\,]$ denotes the expectation value with respect to the *biasing distribution* $p^*(x)$, and where $r(x) = p_X(x)/p^*(x)$ is called the *likelihood ratio* [17]. As before, we then estimate the corresponding integral via Monte Carlo simulations; that is, we write an importance-sampled Monte Carlo estimate for P as

$$\hat{P}_{\rm is} = \frac{1}{M}\sum_{m=1}^{M} I(y(X_m^*))r(X_m^*), \tag{3.4}$$

where now the samples X_m^* are drawn according to the distribution $p^*(x)$. By design, the estimator $\hat{P}_{\rm is}$ is unbiased; i.e., $\mathbb{E}^*[\hat{P}_{\rm is}] = P$. If $p^*(x) \equiv p_X(x)$ (unbiased simulations), $r(x) = 1$ and (3.4) agrees with (3.2) (i.e., one recovers the standard Monte Carlo method). The use of a biasing pdf, however, allows the desired regions of sample space to be visited much more frequently. At the same time, the likelihood ratio automatically adjusts each contribution so that all of the different realizations add properly to give a correct final estimate.

The crucial step when trying to apply IS is to determine a proper choice of the biasing distribution $p^*(x)$ in order to reduce the variance of the estimator $\hat{P}_{\rm is}$. Naturally, not all biasing distributions are appropriate for accomplishing this. One might think that the simplest choice is to increase the overall noise variance, in an attempt to increase the probability of generating errors. It is well-known, however, that this biasing method (which is usually referred to as *variance scaling*) is effective

only in low-dimensional systems [17]. In general, in order for IS to be effective, $p^*(x)$ should concentrate the Monte Carlo samples near the regions that are *most likely* to generate rare events of interest, which in our case means determining the most likely noise instantiations at each amplifier which produce large pulse amplitude or timing variations at the fiber output. The proper choice of biasing distributions when one is interested in amplitude and timing jitter will be discussed in the next section.

If one seeks to reconstruct a broad region of the pdf for the quantity of interest, no single choice of biasing distribution can be expected to capture efficiently all the regions of sample space that give rise to the events of interest. In this case, several different biasing distributions $p_q^*(x)$ can be used and their results combined using a method known as *multiple importance sampling* [26, 27, 28]. With this technique, a weight $w_q(x)$ is associated with each biasing distribution. An importance-sampled estimator for P is then written as

$$\hat{P}_{\text{mis}} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{M_q}\sum_{m=1}^{M_q} w_q(X_{mq}^*)I(y(X_{mq}^*))r_q(X_{mq}^*), \tag{3.5a}$$

where Q is the total number of biasing distributions, M_q is the number of samples drawn from $p_q^*(x)$, X_{mq}^* is the mth such sample, and $r_q(x) = p_X(x)/p_q^*(x)$. Several strategies are possible for selecting the weights; the estimator $\hat{P}$ will be unbiased as long as $\sum_{q=1}^{Q} w_q(x) = 1$ for all x. A particularly simple and effective choice is the *balance heuristic* [26]:

$$w_q(x) = \frac{M_q p_q^*(x)}{\sum_{q'=1}^{Q} M_{q'} p_{q'}^*(x)}. \tag{3.5b}$$

Note that $M_q p_q^*(x)$ is proportional to the expected number of hits from the qth distribution. Thus, the weight of a sample with the balance heuristic is given by the likelihood of realizing that sample with the qth distribution relative to the total likelihood of realizing the same sample with all distributions.

4. IS for amplitude, frequency, timing, and phase jitter. We now turn our attention to the application of IS to Monte Carlo simulations of noise-induced amplitude, frequency, timing, and phase jitter. As explained earlier, in order for IS to be effective we need to bias the simulations towards the events that are most likely to produce the rare events of interest. Therefore, the strategy to bias the simulations toward predetermined target values of each soliton parameter consists of two logically distinct steps: First, we must determine how to bias the noise at each amplifier in order to obtain a specified change of amplitude, frequency, timing, and/or phase. Second, we must determine how to select the individual changes at each amplifier to obtain the desired total change at the end of the transmission line. To accomplish these goals, we need to (i) find the most likely noise configurations that result in a specified soliton parameter change at each amplifier, and (ii) find the most likely combination of individual amplitude, frequency, timing, and phase changes at all of the amplifiers that result in the desired value at the end of the transmission line. The fact that so much is known about NLS solitons greatly simplifies these tasks. The key information comes from the dependence of the soliton solution upon the parameters A, Ω, T_0, and Φ_0. The noise-induced changes in these soliton parameters can be calculated using SPT, as explained in section 2.1. In turn this knowledge can be used to bias the noise at each amplifier to induce larger-than-normal changes of the soliton parameters at the fiber output.

To make these ideas more definite, suppose that we are interested in large deviations of a quantity Y. Later on, this quantity will be identified with the amplitude A, frequency Ω, timing T, or phase Φ of the soliton. For now, let $Y_{\rm in}$ denote the value of Y at the fiber input, and suppose that we want to direct the simulations towards a target value $Y_{\rm out}$. As explained above, we need to (i) find the most likely noise realization at each amplifier to produce a given shift $C_n = Y_n - Y_{n-1}$, and (ii) find the most likely combination of individual contributions $\{C_n\}_{n=1}^{N_a}$ such that the final value of Y is $Y_{\rm out}$. We address these two issues separately.

To solve problem (i), at each amplifier we need to find the noise instantiation that maximizes the probability of realizing a prescribed shift in one of the soliton parameters. In any numerical implementation of (2.1), noise is added to the propagating signal by adding one independent Gaussian random variable to the real part of the optical field, and one to the imaginary part, at each discretized time point (when split-step spectral methods are used to solve (2.1), one can alternatively add an independent Gaussian random variable to the real part and to the imaginary part of every Fourier mode; this is equivalent to the above procedure in the time domain). Maximizing the probability of this Gaussian perturbation amounts to minimizing the sum of the squares of all of the random variables (one for the real part and one for the imaginary part at each time point), and in a continuous-time limit this corresponds to seeking a noise-produced perturbation $f(t)$ that minimizes the L^2-norm

$$\|f\|^2 = \int |f(t)|^2\, dt. \tag{4.1}$$

There is no weighting of the noise because every perturbation of comparable size is equally probable. Of course, we are interested in the noise perturbation that produces not just the most probable change, but rather the most probable change in one of the soliton parameters. This means that the minimization should be performed subject to the constraint

$$\Delta Y_n = \mathrm{Re} \int \underline{v}_Y^\dagger(z_n, t) f(t)\, dt = C_n, \tag{4.2}$$

where $\underline{v}_Y(z_n, t)$ is one of the adjoint linear modes evaluated at $z_n = nz_a$, consistent with (2.9), and C_n is for now an arbitrary constant. This constrained minimization problem can be expressed in Lagrange multiplier form by defining the functional

$$M_n = \int |f(t)|^2\, dt + \lambda \left[\int \underline{v}_Y^\dagger(z_n, t) f(t)\, dt + \int \underline{v}_Y(z_n, t) f^\dagger(t)\, dt - 2C_n \right]. \tag{4.3}$$

The solution to this problem, which is easily obtained using the calculus of variations, is

$$f(t) = C_n \, \frac{\underline{v}_Y(z_n, t)}{\|\underline{v}_Y(z_n, \cdot)\|^2}. \tag{4.4}$$

Here it should be noted that, even though a noise perturbation proportional to one of the linear modes produces a "clean" change in the soliton parameters (that is, a change without additional radiative components), (4.4) implies that the most likely way to realize the same parameter change occurs when the noise is proportional to the corresponding *adjoint* mode, a result which is not evident a priori.

Once the maximum likelihood noise configurations at each amplifier are known, it remains to solve problem (ii), namely to find the coefficients $\{C_1, \ldots, C_{N_a}\}$ that lead

with maximum probability to a prescribed change in the soliton parameter Y over N_a amplifiers. From (2.3), it is immediately apparent that (if one neglects the dependence of the variances in (2.12) on the amplitude) the soliton amplitude A and frequency Ω are not affected by changes to the other parameters, while the soliton timing T and phase Φ are affected both by direct perturbations to T and Φ, respectively, and by integrated changes to the frequency Ω and amplitude A, respectively. We therefore consider these problems separately.

4.1. Amplitude and frequency shifts. In the case of amplitude shifts, the problem is to find the most likely noise realization that produces a prescribed total change in the soliton amplitude, $\Delta A_{\rm tot} = A_{\rm out} - A_{\rm in}$. This amounts to another constrained minimization problem, where we now need to choose the set of individual amplitude shifts at each amplifier, $\{\Delta A_n\}_{n=1}^{N_a}$, in order to minimize the cumulative L^2-norm

$$\sum_{n=1}^{N_a} \|f_n\|^2 = \sum_{n=1}^{N_a} \frac{\Delta A_n^2}{\|\underline{v}_A(z_n,\cdot)\|^2} \tag{4.5}$$

(where (4.4) was used) under the constraint

$$\sum_{n=1}^{N_a} \Delta A_n = \Delta A_{\rm tot}. \tag{4.6}$$

Evaluating the norm of $\underline{v}_A$ using (2.9a), we can also rewrite this optimization problem in Lagrange multiplier form as

$$M = \sum_{n=1}^{N_a} \frac{\Delta A_n^2}{A_n} + \lambda \left[\Delta A_{\rm tot} - \sum_{n=1}^{N_a} \Delta A_n \right], \tag{4.7}$$

where, obviously, at each amplifier $A_n = A_{\rm in} + \sum_{n'=1}^{n} \Delta A_{n'}$. To find the minimum of M we then look for zeros of the gradient of M with respect to all the individual amplitude changes ΔA_n. If the total amplitude change over the N_a amplifiers is not too large, we can write $\Delta A_{\rm tot} = \epsilon\,\Delta$ and employ a perturbation expansion of ΔA_n and λ in powers of ϵ, namely, $\Delta A_n = \epsilon a_n^{(1)} + \epsilon^2 a_n^{(2)} + \cdots$ and of $\lambda = \epsilon\lambda_0 + \epsilon^2\lambda_2 + \cdots$. Minimizing and then matching orders of ϵ gives

$$a_n^{(1)} = \frac{\Delta}{N_a}, \tag{4.8a}$$

$$a_n^{(2)} = -\frac{1}{4}\frac{\Delta^2}{N_a^2}\,(N_a - 2n + 1), \tag{4.8b}$$

and so on. These constants determine the appropriate biasing for amplitude jitter at leading order and up to second order. Note that, at leading order, the desired average change in amplitude over the entire span of amplifiers is simply divided evenly among the individual amplifiers. In practice, this leading order approximation has been found to be adequate for all cases considered, even when the actual amplitude shifts computed were reasonably large.

The optimal biasing problem for the frequency Ω is similar but simpler, in that the L^2-norm of the associated linear mode is independent of Ω. In particular, we seek to minimize

$$\sum_{n=1}^{N_a} \|f_n\|^2 = \sum_{n=1}^{N_a} \frac{\Delta\Omega_n^2}{\|\underline{v}_\Omega(z_n,\cdot)\|^2} \tag{4.9}$$

under the constraint

$$\sum_{n=1}^{N_a} \Delta\Omega_n = \Delta\Omega_{\text{tot}}. \tag{4.10}$$

This leads to

$$M = \sum_{n=1}^{N_a} \frac{\Delta\Omega_n^2}{A_n} + \lambda \left[\Delta\Omega_{\text{tot}} - \sum_{n=1}^{N_a} \Delta\Omega_n \right], \tag{4.11}$$

where, due to the orthogonality of $\underline{v}_\Omega$ and $\underline{v}_A$, the amplitude remains unaffected by the biasing applied to Ω. For this reason, we assume $A_n = 1 \ \forall \ n = 1, \ldots, N_a$, which simply gives

$$\Delta\Omega_n = \frac{\Delta\Omega_{\text{tot}}}{N_a}. \tag{4.12}$$

Note that this assumption would need to be modified if one wished to compute the joint distribution of amplitude and frequency (or amplitude and timing), however.

4.2. Timing and phase shifts. We next look at the most likely noise realization resulting in a prescribed timing shift of the soliton at the fiber output. Because of the Galilean invariance of the NLS equation, in this case we need to consider frequency shifts as well as timing shifts. In other words, we seek to find the most likely set of frequency and timing shifts at each amplifier, $\{\Delta\Omega_n, \Delta T_n\}_{n=1}^{N_a}$, that produce a final value $T_{\text{out}} = T_{\text{in}} + \Delta T_{\text{tot}}$ of timing. Because of the orthogonality of $\underline{v}_T$ and $\underline{v}_\Omega$, this amounts to choosing ΔT_n and $\Delta\Omega_n$ in order to minimize the cumulative L^2-norm

$$\sum_{n=1}^{N_a} \frac{\Delta T_n^2}{\|\underline{v}_T(z_n, \cdot)\|^2} + \sum_{n=1}^{N_a} \frac{\Delta\Omega_n^2}{\|\underline{v}_\Omega(z_n, \cdot)\|^2} \tag{4.13}$$

under the constraint

$$\Delta T_{\text{tot}} = N_a z_a \Omega_{\text{in}} + \sum_{n=1}^{N_a} \Delta T_n + \sum_{n=1}^{N_a} (N_a + 1 - n) \Delta\Omega_n z_a \tag{4.14}$$

for a prescribed value of ΔT_{tot}. Again, we can evaluate the norms in (4.13) and rewrite the above problem in Lagrange multiplier form:

$$M = \frac{6}{\pi^2} \sum_{n=1}^{N_a} A_n^3 \Delta T_n^2 + \frac{3}{2} \sum_{n=1}^{N_a} \frac{\Delta\Omega_n^2}{A_n} + \lambda \left[\sum_{n=1}^{N_a} \Delta T_n + \sum_{n=1}^{N_a} (N + 1 - n) \, \Delta\Omega_n z_a - \Delta T_{\text{tot}} \right]. \tag{4.15}$$

If the noise has components only along $\underline{v}_T$ and $\underline{v}_\Omega$, the soliton amplitude again remains unaffected, so that minimizing the action M gives

$$\Delta T_n = \frac{\pi^2}{12\sigma_{T,\text{tot}}^2} \Delta T_{\text{tot}}, \tag{4.16a}$$

$$\Delta\Omega_n = \frac{(N_a + 1 - n) z_a}{3\sigma_{T,\text{tot}}^2} \Delta T_{\text{tot}}, \tag{4.16b}$$

where

$$\sigma_{T,\mathrm{tot}}^2 = N_a \left[\frac{\pi^2}{12} + \frac{z_a^2}{18} (N_a + 1)(2N_a + 1) \right]. \tag{4.16c}$$

Note that the rule for biasing the frequency given by (4.16b) is rather different from the rule given by (4.12). Whereas the former is designed to produce a given total change in *frequency* with highest likelihood, the latter is designed to produce a given total change in *timing* with highest likelihood, so that frequency shifts occurring earlier in the propagation are weighted much more heavily. In fact, comparing (2.10), (4.14), and (4.16), it can easily be seen that the relative weight of each term is proportional to the variance of its term in the final result. In other words, the most probable way of obtaining a given timing shift at the end of the fiber is to perform relatively larger frequency shifts at the beginning of the fiber, since these are the ones that can accumulate over the longest distances and therefore produce larger deviations for the same "effort" (i.e., for the same contribution to the cumulative L^2-norm of (4.1)).

Just as the optimal noise instantiation to obtain a given total timing shift depends on both frequency and timing shifts at each amplifier, the most probable way of obtaining a prescribed total phase shift requires shifting *three* parameters at each amplifier: phase, amplitude, and frequency. Under the conditions that $\Omega_0 = 0$ and that the individual amplitude shifts are kept small, however, the terms in (2.11d) involving Ω_0 and those involving products of shifts can be neglected, leaving as the action

$$\begin{aligned} M = &\sum_{n=1}^{N_a} \frac{A_n \Delta\Phi_n^2}{\pi^2/18 + 2/3} + \sum_{n=1}^{N_a} \frac{\Delta A_n^2}{2A_n} \\ &+ \lambda \left[\sum_{n=1}^{N_a} \Delta\Phi_n + \sum_{n=1}^{N_a} (N + 1 - n) A_0 \Delta A_n z_a - \left(\Phi_{\mathrm{out}} - \frac{1}{2} N_a z_a A_0^2 \right) \right]. \end{aligned} \tag{4.17}$$

This action has markedly similar form to (4.15) and demonstrates again that the effect of amplitude shifts on phase through self-phase modulation is completely analogous to the effect of frequency shifts on position through Galilean invariance. Here, however, the fact that amplitude appears in the summations is problematic, as amplitude is one of the parameters being shifted in our biasing scheme. To resolve this, we take an approach similar to that for direct amplitude shifting; i.e., we use a perturbation expansion in ΔA_{tot}. At leading order, the optimal phase and amplitude shifts take the same form as the above optimal timing and frequency shifts:

$$\Delta\Phi_n = \frac{\pi^2/36 + 1/3}{\sigma_{\Phi,\mathrm{tot}}^2} \left(\Delta\Phi_{\mathrm{tot}} - \frac{1}{2} N_a z_a \right), \tag{4.18a}$$

$$\Delta A_n = \frac{(N_a + 1 - n) z_a}{\sigma_{\Phi,\mathrm{tot}}^2} \left(\Delta\Phi_{\mathrm{tot}} - \frac{1}{2} N_a z_a \right), \tag{4.18b}$$

where

$$\sigma_{\Phi,\mathrm{tot}}^2 = N_a \left[\frac{\pi^2}{36} + \frac{1}{3} + \frac{z_a^2}{6} (N_a + 1)(2N_a + 1) \right]. \tag{4.18c}$$

4.3. Implementation issues. Having found the most probable configurations that produce given values of amplitude, frequency, timing, and phase shifts, we now

discuss how to use them to guide the biasing of the importance-sampled Monte Carlo simulations.

For concreteness, suppose that we are numerically solving a discretized version of the NLS equation (2.1). As explained earlier, at each amplifier we add random noise f_n. If J is the total number of discrete time points in the computational domain (or, equivalently, the total number of complex Fourier modes), the noise is represented by a vector $\mathbf{x}_n = (x_1, \ldots, x_{2J})^T$ giving the real and imaginary noise components at each discretized time location. In the unbiased case, the x_j are independent identically distributed (i.i.d.) normal RVs with mean zero and variance $\sigma_a^2 = \sigma^2/(2\Delta t)$; explicitly, the probability distribution is $p_{\mathbf{x}}(\mathbf{x}) = \exp[-\mathbf{x}^T\mathbf{x}/2\sigma_a^2]/(2\pi\sigma_a^2)^J$. Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ be the $2J \times N$ matrix that denotes all of the noise components at all of the amplifiers, and suppose that we are interested in reconstructing the pdf of a variable $y(X)$. (Here, y will be the amplitude A, the frequency Ω, the timing T, or the phase Φ.)

At each amplifier, we will perform the biasing by selecting a *deterministic* biasing vector $\mathbf{b}_n$ that will be added to the noise vector $\mathbf{x}_n$ drawn from the unbiased distribution. That is, we will form a biased noise realization as $\mathbf{x}_n^* = \mathbf{x}_n + \mathbf{b}_n$. This corresponds to choosing, at each amplifier, the biasing pdf $p_{\mathbf{x}}^*(\mathbf{x}_n^*) = p_{\mathbf{x}}(\mathbf{x}_n^* - \mathbf{b}_n) = p_{\mathbf{x}}(\mathbf{x}_n)$, which therefore gives a likelihood ratio of $r_{\mathbf{x}}(\mathbf{x}_n^*) = p_{\mathbf{x}}(\mathbf{x}_n + \mathbf{b}_n)/p_{\mathbf{x}}(\mathbf{x}_n)$. One can then obtain the overall likelihood ratio of the noise over N_a amplifiers[3] as

$$r(X^*) = \prod_{n=1}^{N_a} \frac{p_{\mathbf{x}}(\mathbf{x}_n^*)}{p_{\mathbf{x}}(\mathbf{x}_n)},$$

where $X^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_{N_a}^*)$ and $\mathbf{x}_n^* = \mathbf{x}_n + \mathbf{b}_n$, as before. Of course, in order for this strategy to be effective, the choice of the biasing vectors $\mathbf{b}_n$ is crucial. The means by which we choose these biasing vectors in the case of amplitude, frequency, timing, and phase jitter is discussed next.

In order to perform the biasing, at each amplifier we first need to find the soliton parameters associated with the solution immediately before amplification (i.e., the addition of noise). We do this either by solving the Zakharov–Shabat eigenvalue problem [9, 29] or by using the moment integrals for the soliton parameters [10]. (A more detailed discussion of this soliton reconstruction process is contained in Appendices B and C.) The soliton parameters uniquely determine the soliton solution, which in turn determines the linear modes. Since the deterministic biasing term is expressed in the form of a linear combination of modes (as determined in the previous subsection), knowing the soliton parameters allows us to select the proper biasing of the Monte Carlo simulations as given by (4.4) and the determination of C_n in subsections 4.1 and 4.2.

In particular, if one wants to bias the amplitude, one chooses a shift in the mean of the noise $f(t)$ equal to

$$\Delta A_n \frac{\underline{v}_A(z_n, t)}{\|\underline{v}_A(z_n, \cdot)\|^2} = \Delta A_n \frac{\underline{v}_A(z_n, t)}{2A_n}, \tag{4.19}$$

[3]Note that the biased noise realizations at each amplifier are not statistically independent, since at each amplifier the choice of the biasing vector $\mathbf{b}_n$ depends on the current state of the soliton and therefore on the accumulated effect of the noise from the previous amplifiers. Nevertheless, it is easy to show that the overall likelihood ratio for such a Markov process can still be written as a product of the individual likelihood ratios (e.g., see [28]).

where ΔA_n is given by (4.8). Similarly, if one wants to bias the pulse frequency, one chooses a shift in the mean of the noise equal to

$$\Delta\Omega_n \frac{\underline{v}_\Omega(z_n,t)}{\|\underline{v}_\Omega(z_n,\cdot)\|^2} = \Delta\Omega_n \frac{3\,\underline{v}_\Omega(z_n,t)}{2A_n}, \tag{4.20}$$

where $\Delta\Omega_n$ is given by (4.12). To bias the soliton position, however, one must also bias the frequency, and in this case one chooses a shift in the mean of the noise equal to

$$\Delta T_n \frac{\underline{v}_T(z_n,t)}{\|\underline{v}_T(z_n,\cdot)\|^2} + \Delta\Omega_n \frac{\underline{v}_\Omega(z_n,t)}{\|\underline{v}_\Omega(z_n,\cdot)\|^2} = \Delta T_n \frac{6A_n^3\,\underline{v}_T(z_n,t)}{\pi^2} + \Delta\Omega_n \frac{3\,\underline{v}_\Omega(z_n,t)}{2A_n}, \tag{4.21}$$

where ΔT_n and $\Delta\Omega_n$ are now given by (4.16). Finally, to bias the phase one must also bias the amplitude, giving a mean noise shift of

$$\Delta\Phi_n \frac{\underline{v}_\Phi(z_n,t)}{\|\underline{v}_\Phi(z_n,\cdot)\|^2} + \Delta A_n \frac{\underline{v}_A(z_n,t)}{\|\underline{v}_A(z_n,\cdot)\|^2} = \Delta\Phi_n \frac{A_n\,\underline{v}_\Phi(z_n,t)}{\pi^2/18+2/3} + \Delta A_n \frac{\underline{v}_A(z_n,t)}{2A_n}. \tag{4.22}$$

In the discretized version of the problem, this biasing term, i.e., the shift of the mean of the noise $f(t)$, can also be represented as a vector, $\mathbf{b}_n$. Once the biasing direction and strength have been chosen, the actual biasing is straightforward: an unbiased noise realization $\mathbf{x}_n$ is generated, and the biased noise realization $\mathbf{x}_n^*$ is obtained by simply adding $\mathbf{b}_n$ to $\mathbf{x}_n$; that is, $\mathbf{x}_n^* = \mathbf{x}_n + \mathbf{b}_n$, as explained above.

5. Numerical results. To demonstrate the effectiveness of applying IS to Monte Carlo simulations of amplitude, frequency, timing, and phase jitter, we have performed simulations using the procedure described above. In dimensionless units, we took an amplifier spacing of $z_a = 0.1$, a total propagation distance of $N_a z_a = 20$, and a dimensionless noise strength of $\sigma^2 = 6.3 \times 10^{-5}$. The physical parameters generating these values are given in Appendix A. In the simulations, we extracted the soliton parameters at the intermediate amplifiers using moments (see Appendix C), but computed the values at the final distance using the more accurate Zakharov–Shabat eigenvalue problem (see Appendix B).

Figure 2 shows the results of 50,000 importance-sampled Monte Carlo simulations, selectively targeting amplitude fluctuations. Five biasing targets with 10,000 samples per target were used. Different choices of biasing generate the different regions of the pdfs shown in Figure 3, and the results from all Monte Carlo realizations are combined using the balance heuristic described in section 3. Even with a relatively small number of Monte Carlo samples, the method produces values of amplitude and timing jitter far down into the tails of the probability distributions. As shown in Figure 2, these results agree with unbiased Monte Carlo simulations in the main portion of the pdf (the only region that can be reconstructed with unbiased simulations). For smaller probability values, however, the deviations from Gaussian are evident.

A simple model of amplitude fluctuations can be obtained via soliton perturbation theory [10]: $A_{n+1} = A_n + \sqrt{A_n}\, s_{n+1}$, where the s_n are i.i.d. normal RVs with mean zero and variance σ^2. (Of course, this model cannot be correct when the noise is not a small perturbation of the soliton; an obvious erroneous result of this is that there is a very slight probability for negative A_n's to occur. Fortunately, it will be seen that such unphysical cases occur with extremely small probability, and therefore can be ignored.) Note that this model reflects a prepoint approximation of the jump conditions in (2.1). While this approach is closer in spirit to the Markov process created

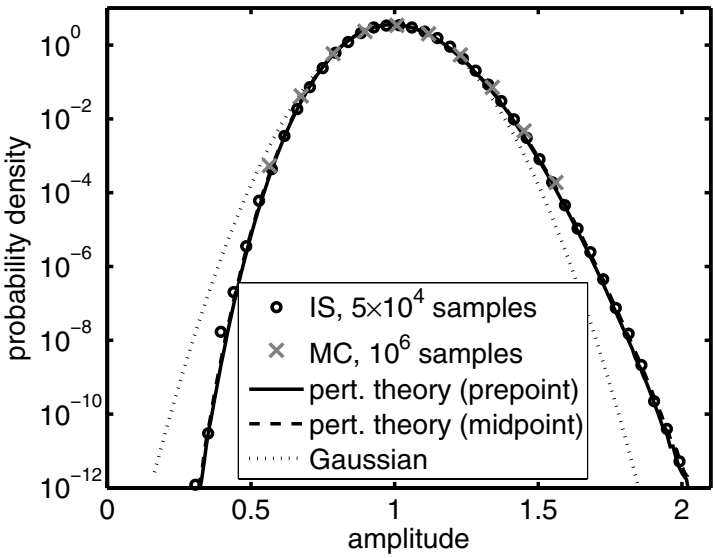


FIG. 2. *The pdf of amplitude jitter in a soliton-based transmission system, obtained from* 50,000 *importance-sampled Monte Carlo simulations. Results from a simple model from perturbation theory and an approximate Gaussian curve are also shown.*

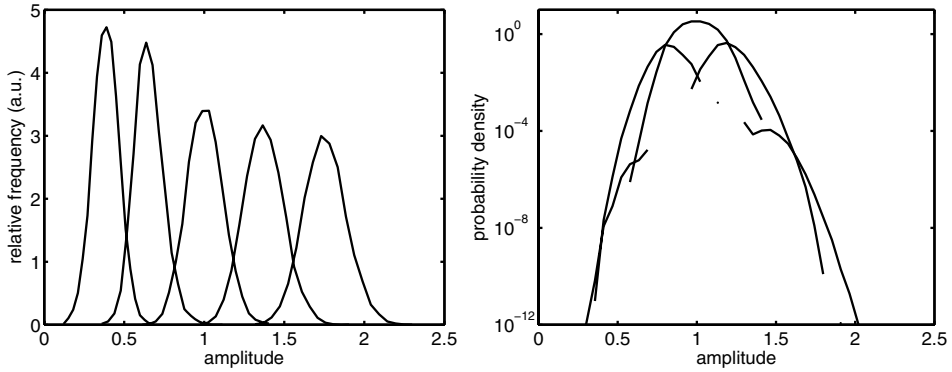


FIG. 3. *(left) Relative frequency plots showing the different ranges of amplitude generated by biasing distributions with five different targets. From left to right, the targets are* $\Delta A_{\rm tot} = \{-0.8, -0.4, 0, 0.4, 0.8\}$. *(right) The relative contribution of each biasing distribution to the final result of Figure* 2 *when weighted by the balance heuristic.*

by the biased Monte Carlo simulations, it is unclear whether the physical process is more accurately represented by this approximation or by a midpoint approximation, given by $A_{n+1} = A_n + (\sqrt{A_{n+1}} + \sqrt{A_n})\, s_{n+1}/2$. (For example, in one interpretation of a quantum-mechanical analysis of noise induced by loss and gain in a periodically amplified system, half of the noise is contributed in a distributed manner by the loss [30].) For comparison, in Figure 2 we show the pdfs obtained from both models, using importance-sampled numerical simulations for the prepoint approximation and simple analysis (see Appendix D) for the midpoint approximation. Although agreement with the importance-sampled simulations of the full NLS equation (2.1) is very

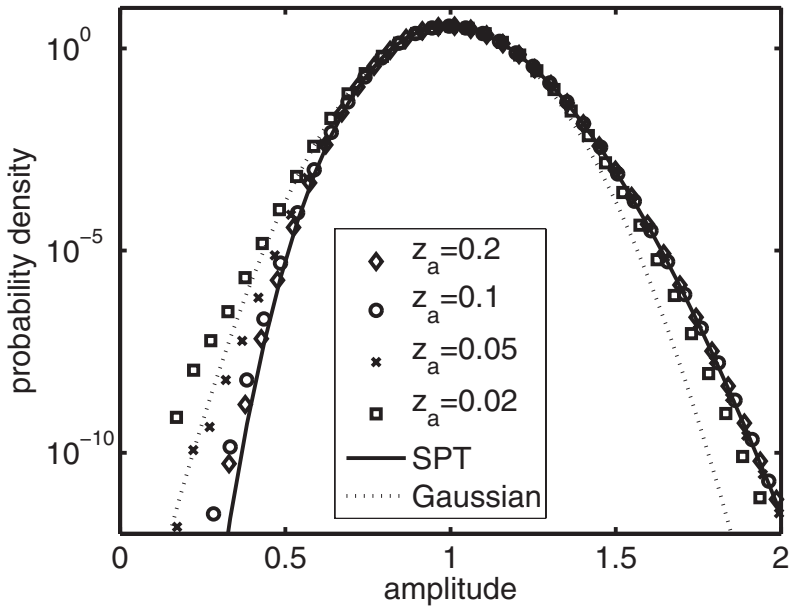


FIG. 4. *The pdf of amplitude jitter when the amplifier separation z_a is varied while the number of amplifiers N_a and the noise strength at each amplifier are held fixed. Even though SPT predicts the solid curve for each of these runs, the agreement with PDE numerics appears to be good only for sufficiently large z_a. Each of the numerical runs was obtained using 5×10^5 simulations.*

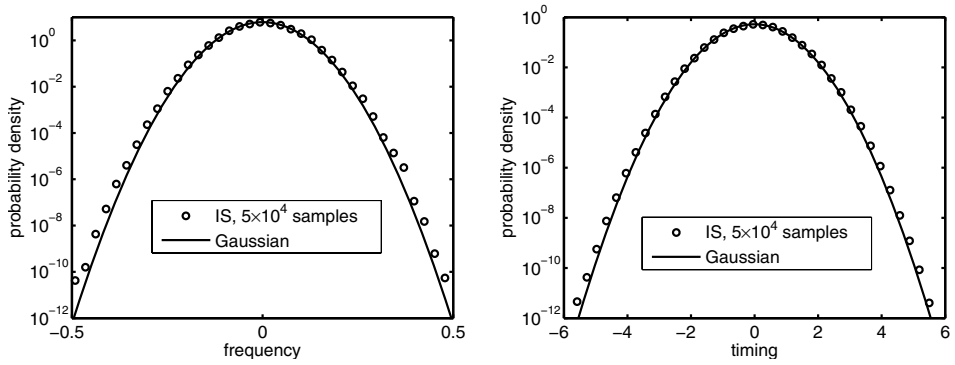


FIG. 5. *The pdfs of frequency (left) and timing jitter (right) in a soliton-based transmission system, each obtained from* 50,000 *importance-sampled Monte Carlo simulations.*

good throughout the range of amplitude values considered, with a slight deviation at small amplitudes, this agreement deteriorates significantly at both small and large amplitudes when the amplifier spacing z_a is decreased, as shown in Figure 4. As z_a is increased, the agreement appears to improve. It is not clear why the numerical results disagree with SPT here; nevertheless, the biasing obtained using SPT is sufficiently close to the correct biasing that the IS simulations accurately capture the pdf.

Figure 5 shows results similar to those of Figure 2, but for importance-sampled Monte Carlo simulations targeting frequency and timing fluctuations. The distributions of frequency and timing jitter agree well over a larger range of probability values with Gaussian curves whose variances are calculated from the theoretical results,

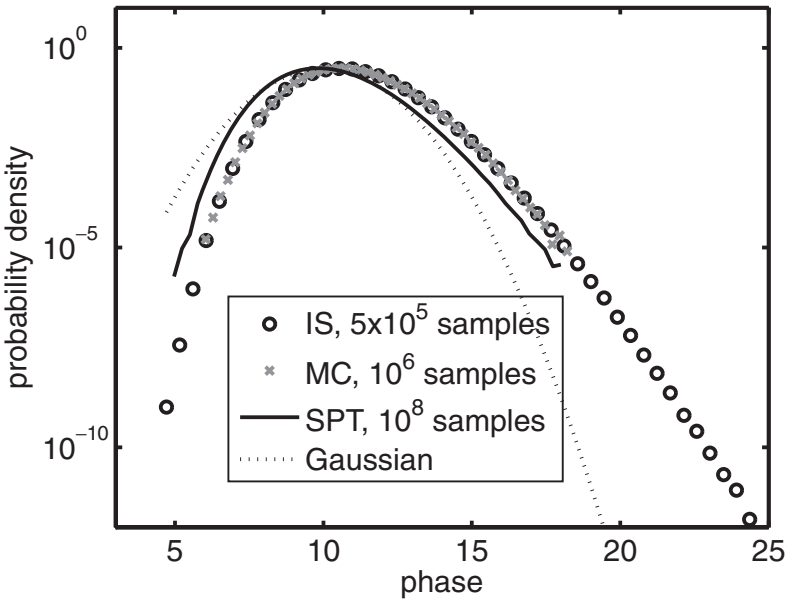


FIG. 6. *Comparison of the pdf of phase jitter with the Gaussian obtained from the linearized equations from SPT and with Monte Carlo simulations of the full PDE and of the nonlinear SPT equations. In this case, the last two results disagree significantly.*

(2.14b) and (2.14c). Note, however, that both for frequency and for timing jitter small deviations from Gaussian behavior are observed in the tails, where the numerically reconstructed pdfs are slightly but systematically larger than the predicted Gaussians. This discrepancy is due to the amplitude dependence of the variances of frequency and timing fluctuations [13]. Such dependence was neglected when the deterministic biasing choices were determined. Nevertheless, because random sampling is used, the numerical simulations access not only the deterministic biasing directions, but also nearby points in sample space around them. If errors in the deterministic biasing directions are not too large, a reasonably large number of random samples will find the correct regions in sample space that contribute most significantly to the pdf.

Finally, the pdf of phase jitter obtained using 5×10^5 IS trials is shown in Figure 6, along with 10^6 Monte Carlo runs to demonstrate the accuracy of the biased runs. We have unwrapped the phase to better illustrate deviations in the pdf tails. As expected, the pdf disagrees with the Gaussian obtained by linearizing the soliton perturbation equations. Somewhat surprising, however, is the fact that it also disagrees with 10^8 Monte Carlo simulations of the full nonlinear SPT equations. This suggests that dispersive radiation plays an important role in the case of phase jitter, rendering the (first-order) SPT equations ineffective in reproducing the correct jitter statistics. These equations are nevertheless still sufficiently accurate to provide effective biasing for the IS runs.

6. Conclusion. In summary, we have presented the application of importance sampling to numerical simulations of large noise-induced perturbations of nonlinear Schrödinger solitons, and we have demonstrated the method by calculating the pdfs of amplitude, frequency, timing, and phase jitter in a soliton-based transmission system. These results show that IS can be an effective tool for assessing the impact of noise in such systems.

Appendix A. NLS nondimensionalization and units. Here we describe the nondimensionalization procedure and the choice of units. The NLS equation is written in dimensional units as

$$i\frac{\partial E}{\partial Z} + \frac{|\beta''|}{2}\frac{\partial^2 E}{\partial T^2} + \gamma|E|^2E = i\sum_{n=1}^{N_a}\delta(Z - nZ_a)F_n(T), \tag{A.1}$$

where $|E|^2$ is optical power in watts, Z and T are dimensional distance in km and retarded time in ps, Z_a is the amplifier spacing, and β'' is the group velocity dispersion parameter in ps^2/km. The nonlinear coefficient is $\gamma = \omega_0 n_2/cA_{\text{eff}}$, where ω_0 is the carrier frequency, n_2 is the Kerr nonlinear-index coefficient, c is the vacuum speed of light, and A_{eff} is the effective area of the fiber core. The periodic cycle of loss and gain introduced by the chain of amplifiers has already been averaged out of (A.1); for details, see [30]. The delta-correlated white noise added at each amplifier then has noise strength

$$\langle F_m(T)F_n^\dagger(T')\rangle = \frac{\hbar\omega_0\eta_{\text{sp}}(G-1)^2}{G\ln G}\delta_{mn}\delta(T-T'), \tag{A.2}$$

where G is the power gain at each amplifier and η_{sp} is the spontaneous emission factor.

We then let $z = Z/L$, $t = T/T_0$, and $u = E/E_0$, where $L = T_0^2/|\beta''|$ is the dispersion length, $T_0 = T_{\text{FWHM}}/1.76$ is the soliton (sech) width, and $E_0 = 1/\sqrt{L\gamma}$ is the characteristic optical power for critical balance between nonlinearity and group velocity dispersion. This reduces (A.1) and (A.2) to (2.1) and (2.2), with

$$\sigma^2 = \frac{\hbar\omega_0\eta_{\text{sp}}\gamma T_0}{|\beta''|}\frac{(G-1)^2}{G\ln G}. \tag{A.3}$$

In the simulations we used a pulse full width at half maximum (FWHM) of 17.6 ps (i.e., a sech width of $T_0 = 10$ ps), an amplifier spacing of $Z_a = 50$ km and a fiber loss of 0.2 dB/km (yielding a power gain of $G = 10$), a spontaneous emission factor of η_{sp} of 1.4, a fiber dispersion $\beta'' = -0.2$ ps^2/km, and a total transmission distance of 10,000 km. The nonlinear coefficient of the fiber was taken to be 2.0 km^{-1}W^{-1}. The dimensionless parameters corresponding to these values are given in section 5.

Appendix B. Soliton extraction via Zakharov–Shabat eigenvalue problem. As discussed in the main text, the first step in implementing importance sampling is to find the soliton part of the solution at each amplifier. One way to do this is to solve the Zakharov–Shabat (Z-S) eigenvalue problem [31, 32]. Given a solution u of the NLS equation at a particular value of z, one can discretize the Z-S eigenvalue problem and solve it numerically [29, 33]. In the case of noisy solutions, which may not be smooth, it may be more robust to use a completely integrable discrete version, such as the Ablowitz–Ladik eigenvalue problem [29].

Unfortunately, an eigenvalue of the Z-S problem (or its discrete equivalent, the Ablowitz–Ladik problem) gives only two of the soliton parameters, the amplitude and the frequency, and there is apparently no way to determine the *exact* values of the soliton's position and phase. One can, however, obtain values that are relatively unaffected by noise, even when this perturbation is large.

To do this, one makes use of the trace formula for the NLS equation [34],

$$u = 2i\sum_{k=1}^{N}\frac{b_k}{a_k'}\psi_1^2 - 2i\sum_{k=1}^{N}\frac{b_k^\dagger}{a_k'^\dagger}\psi_2^{\dagger 2} - \frac{1}{\pi}\int_{-\infty}^{\infty}\left\{\frac{b}{a}(\xi)\psi_1^2(t,\xi) + \frac{b^\dagger}{a^\dagger}(\xi)\psi_2^{\dagger 2}(t,\xi)\right\}d\xi.$$

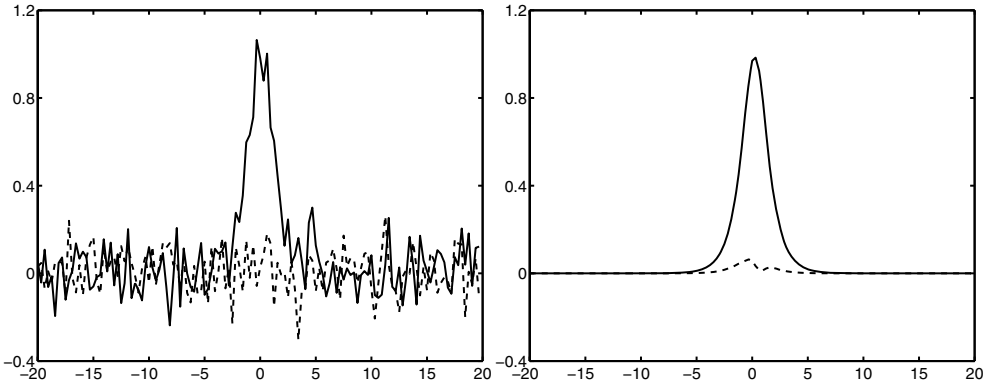


FIG. 7. *A noisy soliton (left) and the "clean" soliton (right) recovered from the Jost functions of the Z-S eigenproblem. Apart from a small ripple in the imaginary component, the Jost functions are seen to produce a reconstructed soliton that is largely free of radiation.*

Essentially, this shows that one can break the solution up into two contributions, one from the eigenfunctions of the Z-S scattering problem and the other from the continuous spectrum. Here, $\psi_1(x,\zeta)$ and $\psi_2(t,\zeta)$ are the components of one set of *Jost functions*, i.e., solutions of the Z-S scattering problem satisfying special boundary conditions, namely,

$$\begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} \sim \begin{pmatrix} 0 \\ 1 \end{pmatrix} e^{i\zeta t} \quad \text{as } t \to +\infty \quad \text{or} \quad \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} \sim \begin{pmatrix} 1 \\ 0 \end{pmatrix} e^{-i\zeta t} \quad \text{as } t \to -\infty.$$

The coefficients b_k, a'_k, $b(\xi)$, and $a(\xi)$ are determined by the connection between these two sets of functions,

$$\begin{pmatrix} \phi_1(t,\zeta) \\ \phi_2(t,\zeta) \end{pmatrix} = a(\zeta) \begin{pmatrix} \psi_2^\dagger(t,\zeta^\dagger) \\ -\psi_1^\dagger(t,\zeta^\dagger) \end{pmatrix} + b(\zeta) \begin{pmatrix} \psi_1(t,\zeta) \\ \psi_2(t,\zeta) \end{pmatrix}.$$

At an eigenvalue ζ_k of the Z-S scattering problem (with $\operatorname{Im}\zeta_k > 0$), one has $a(\zeta_k) = 0$. At such an eigenvalue, $a'_k = a'(\zeta_k)$ and $b_k = b(\zeta_k)$.

In order to extract the position and phase from a noisy soliton in a robust way, the idea is to discard the contribution from the continuous spectrum and use only the discrete part of the trace formula. Because of the exponential decay of the eigenfunctions, this "reconstructed" or "nonlinearly filtered" solution will be smooth, and thus definitions of position and phase using moments will not have any difficulties caused by long noisy tails present in the solution. Of course, one must recognize that the trace formula does not precisely partition the solution into "soliton" and "dispersive radiation" components, as the discrete part of the trace formula does not produce solutions which are exactly hyperbolic-secant shaped. Nevertheless, the solutions appear very much like solitons, as shown in Figure 7.

As written, the trace formula is a little difficult to use, as one still needs the coefficients b_k and a'_k. Fortunately, the ratio of these coefficients can be computed more conveniently. From the orthogonality relation [32, equation A6.6e], one has

$$\int_{-\infty}^{\infty} (\psi_{2k}\phi_{1k} + \psi_{1k}\phi_{2k})\, dt = ia'_k \quad \Rightarrow \quad 2b_k \int_{-\infty}^{\infty} \psi_{1k}\psi_{2k}\, dt = ia'_k,$$

since $\phi_{1k} = b_k\psi_{1k}$ and $\phi_{2k} = b_k\psi_{2k}$. Thus, the discrete part of the trace formula becomes

$$u = -\sum_{k=1}^{N}\left(\frac{\psi_{1k}^2}{\int_{-\infty}^{\infty}\psi_{1k}\psi_{2k}\,dt} + \frac{{\psi_{2k}^{\dagger}}^2}{\int_{-\infty}^{\infty}\psi_{1k}^{\dagger}\psi_{2k}^{\dagger}\,dt}\right).$$

Because the numerator and denominator in this expression are both quadratic in ψ_k, this means that one does not need to normalize the Jost eigenfunctions when computing their contribution to the solution.

Appendix C. Soliton extraction via moments. The method of obtaining the soliton parameters described above is very effective but computationally intensive, requiring the numerical determination of selected eigenvalues and eigenfunctions of a large matrix. The resolution afforded by this method is critical for the final calculation of the soliton parameters in order to determine correct statistics; the formulation of the biasing vectors, however, does not require the same degree of precision. In particular, if the applied biasing vector differs from the optimal biasing vector by a small random amount (caused, for example, by sensitivity of the parameter measurement technique to the presence of radiation), this can be expected to produce only a small reduction in the efficiency of generating an accurate and unbiased estimate through IS. As long as the biased Monte Carlo simulations sample a large enough region around the deterministic biasing direction, the method will remain efficient. Another way to interpret this is that the Monte Carlo sampling corrects for slight inaccuracies in the determination of the optimal biasing direction.

We have therefore used the following filtered moments to generate approximate values for the soliton parameters at each amplifier. We first obtain an estimate for the soliton frequency

$$\Omega_{\text{est}} = \frac{\int \omega\,|\tilde{u}|^2\,d\omega}{\int |\tilde{u}|^2\,d\omega}, \tag{C.1}$$

where $\tilde{u}$ is the Fourier transform of u. We then use this to band-pass filter the soliton to reduce the noise,

$$\tilde{u}_{\text{filt}} = \tilde{u}\,e^{-(\omega-\Omega_{\text{est}})^2/2W_{\text{filt}}^2}. \tag{C.2}$$

This filtered image of the noisy soliton is used to obtain the final parameter estimates,

$$A = \frac{1}{2}\int |u_{\text{filt}}|^2\,dt, \qquad \Omega = \frac{\int \omega|\tilde{u}_{\text{filt}}|^2\,d\omega}{\int |\tilde{u}_{\text{filt}}|^2\,d\omega}, \tag{C.3}$$

$$T = \frac{\int t|u_{\text{filt}}|^2\,dt}{\int |u_{\text{filt}}|^2\,dt}, \qquad \Phi = \frac{\int \arctan(\operatorname{Im} u_{\text{filt}}/\operatorname{Re} u_{\text{filt}})|u_{\text{filt}}|^2\,dt}{\int |u_{\text{filt}}|^2\,dt}. \tag{C.4}$$

Appendix D. Pdf for the midpoint model. The model of the soliton amplitude's random walk obtained by applying first-order SPT and a particular midpoint approximation to (2.1) was given in section 5 as

$$A_{n+1} = A_n + \frac{1}{2}\left(\sqrt{A_{n+1}} + \sqrt{A_n}\right)s_{n+1}, \tag{D.1}$$

where the s_n are i.i.d. normal RVs with mean zero and variance σ^2. To obtain the pdf for this process, it is convenient to introduce $a_n = \sqrt{A_n}$ and to collect terms, noting

that one can then complete the square by adding $s_n^2/16$ to both sides of the resulting equation:

$$\left(a_{n+1} - \frac{1}{4}s_{n+1}\right)^2 = \left(a_n + \frac{1}{4}s_{n+1}\right)^2. \tag{D.2}$$

Taking the positive branch of this square root then gives the much simpler process $a_{n+1} = a_n + s_{n+1}/2$, which is seen immediately to result in a Gaussian distribution for a_n with mean a_0 and variance $\sigma_a^2 = n\sigma^2/4$. Finally, a simple transformation yields the pdf for A_n:

$$p(A_n) = \frac{1}{2}\frac{1}{\sqrt{2\pi A_n \sigma_a^2}} \exp\left(-\frac{(\sqrt{A_n} - \sqrt{A_0})^2}{2\sigma_a^2}\right). \tag{D.3}$$

REFERENCES

[1] E. Iannone, F. Matera, A. Mecozzi, and M. Settembre, *Nonlinear Optical Communication Networks*, Wiley, New York, 1998.

[2] G. P. Agrawal, *Fiber Optics Communication Systems*, 3rd ed., Wiley, New York, 2002.

[3] D. Marcuse, *Calculation of bit-error probability for lightwave system with optical amplifiers and post-detection Gaussian noise*, J. Lightwave Tech., 9 (1991), pp. 505–513.

[4] J. P. Gordon and H. A. Haus, *Random walk of coherently amplified solitons in optical fiber transmission*, Opt. Lett., 11 (1986), pp. 665–667.

[5] J. P. Gordon and L. F. Mollenauer, *Phase noise in photonic communication systems using linear amplifiers*, Opt. Lett., 15 (1990), pp. 1351–1353.

[6] C. R. Menyuk, *Non-Gaussian corrections to the Gordon-Haus distribution resulting from soliton interactions*, Opt. Lett., 20 (1995), pp. 270–272.

[7] T. Georges, *Study of the non-Gaussian timing jitter statistics induced by soliton interaction and filtering*, Opt. Commun., 123 (1996), pp. 617–623.

[8] R. Holzlöhner, V. S. Grigoryan, C. R. Menyuk, and W. L. Kath, *Accurate calculation of eye diagrams and bit error rates in optical transmission systems using linearization*, J. Lightwave Tech., 20 (2002), pp. 389–400.

[9] A. Hasegawa and Y. Kodama, *Solitons in Optical Communications*, Oxford University Press, Oxford, UK, 1995.

[10] E. Iannone, F. Matera, A. Mecozzi, and M. Settembre, *Nonlinear Optical Communication Networks*, Wiley, New York, 1998.

[11] R. Holzlöhner and C. R. Menyuk, *The use of multicanonical Monte Carlo simulations to obtain accurate bit error rates in optical communication systems*, Opt. Lett., 23 (2003), pp. 1894–1896.

[12] G. E. Falkovich, I. Kolokolov, V. Lebedev, and S. K. Turitsyn, *Statistics of soliton-bearing systems with additive noise*, Phys. Rev. E (3), 63 (2001), 025601.

[13] K.-P. Ho, *Non-Gaussian statistics of the soliton timing jitter due to amplifier noise*, Opt. Lett., 28 (2003), pp. 2165–2167.

[14] G. Biondini, W. L. Kath, and C. R. Menyuk, *Importance sampling for polarization-mode dispersion*, Photon. Technol. Lett., 14 (2002), pp. 310–312.

[15] S. L. Fogal, G. Biondini, and W. L. Kath, *Multiple importance sampling for first- and second-order polarization-mode dispersion*, Photon. Technol. Lett., 14 (2002), pp. 1273–1275; *Correction*, 14 (2002), p. 1487.

[16] R. O. Moore, G. Biondini, and W. L. Kath, *Importance sampling for noise-induced amplitude and timing jitter in soliton transmission systems*, Opt. Lett., 28 (2003), pp. 105–107.

[17] P. J. Smith, M. Shafi, and H. Gao, *Quick simulation: A review of importance sampling techniques in communications systems*, IEEE J. Select. Areas Commun., 15 (1997), pp. 597–613.

[18] R. Srinivasan, *Importance Sampling: Applications in Communications and Detection*, Springer-Verlag, New York, 2002.

[19] C. J. McKinstrie and T. I. Lakoba, *Probability-density function for energy perturbations of isolated optical pulses*, Optics Express, 11 (2003), pp. 3628–3648.

[20] D. J. Kaup, *Closure of the squared Zakharov-Shabat eigenstates*, J. Math. Anal. Appl., 54 (1976), pp. 849–864.

[21] D. J. Kaup, *A perturbation expansion for the Zakharov–Shabat inverse scattering transform*, SIAM J. Appl. Math., 31 (1976), pp. 121–133.
[22] Y. S. Kivshar and B. A. Malomed, *Dynamics of solitons in nearly integrable systems*, Rev. Modern Phys., 61 (1989), pp. 763–915.
[23] W. L. Kath, *A modified conservation law for the phase of the nonlinear Schrödinger equation*, Methods Appl. Anal., 4 (1997), pp. 141–155.
[24] J. Yang, *Complete eigenfunctions of linearized integrable equations expanded around a soliton solution*, J. Math. Phys., 41 (2000), pp. 6614–6638.
[25] E. T. Spiller, W. L. Kath, R. O. Moore, and C. J. McKinstrie, *Computing large signal distortions and bit-error ratios in DPSK transmission systems*, Photon. Technol. Lett., 17 (2005), pp. 1022–1024.
[26] E. Veach, *Robust Monte Carlo Methods for Light Transport Simulation*, Ph.D. thesis, Department of Computer Science, Stanford University, Palo Alto, CA, 1997.
[27] A. Owen and Y. Zhou, *Safe and effective importance sampling*, J. Amer. Statist. Assoc., 95 (2000), pp. 135–143.
[28] G. Biondini, W. L. Kath, and C. R. Menyuk, *Importance sampling for polarization mode dispersion: Techniques and applications*, J. Lightwave Technol., 22 (2004), pp. 1210–1215.
[29] J. A. C. Weideman and B. M. Herbst, *Finite difference methods for an AKNS eigenproblem*, Math. Comput. Simulation, 43 (1997), pp. 77–88.
[30] W. L. Kath, A. Mecozzi, P. Kumar, and C. G. Goedde, *Long-term storage of a soliton bit stream using phase-sensitive amplification: Effects of soliton-soliton interactions and quantum noise*, Opt. Commun., 157 (1998), pp. 310–326.
[31] V. E. Zakharov and A. B. Shabat, *Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media*, Soviet Physics JETP, 34 (1972), pp. 62–69.
[32] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, *The inverse scattering transform-Fourier analysis for nonlinear problems*, Studies in Appl. Math., 53 (1974), pp. 249–315.
[33] S. Burtsev, R. Camassa, and I. Timofeyev, *Numerical algorithms for the direct spectral transform with applications to nonlinear Schrödinger type systems*, J. Comput. Phys., 147 (1998), pp. 166–186.
[34] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform*, SIAM Stud. Appl. Math. 4, SIAM, Philadelphia, 1981.

SIAM J. APPL. MATH.
Vol. 67, No. 5, pp. 1440–1452

THE INVERSE CONDUCTIVITY PROBLEM WITH AN IMPERFECTLY KNOWN BOUNDARY IN THREE DIMENSIONS*

VILLE KOLEHMAINEN†, MATTI LASSAS‡, AND PETRI OLA§

Abstract. We consider the inverse conductivity problem in a strictly convex domain whose boundary is not known. Usually the numerical reconstruction from the measured current and voltage data is done assuming that the domain has a known fixed geometry. However, in practical applications the geometry of the domain is usually not known. This introduces an error, and effectively changes the problem into an anisotropic one. The main result of this paper is a uniqueness result characterizing the isotropic conductivities on convex domains in terms of measurements done on a different domain, which we call the model domain, up to an affine isometry. As data for the inverse problem, we assume the Robin-to-Neumann map and the contact impedance function on the boundary of the model domain to be given. Also, we present a minimization algorithm based on the use of Cotton–York tensor, which finds the push forward of the isotropic conductivity to our model domain and also finds the boundary of the original domain up to an affine isometry. This algorithm works also in dimensions higher than three, but then the Cotton–York tensor has to replaced with the Weyl tensor.

Key words. inverse conductivity problem, electrical impedance tomography, unknown boundary, Cotton–York tensor

AMS subject classifications. 35J25, 35R30, 58J32

DOI. 10.1137/060666986

1. Introduction.

We consider the electrical impedance tomography problem (EIT for short), i.e. the determination of the unknown isotropic conductivity distribution inside a domain in $\mathbb{R}^3$, for example the human thorax, from voltage and current measurements made on the boundary. Mathematically this is formulated as follows: Let Ω be the measurement domain, and denote by γ the bounded and strictly positive function describing the conductivity in Ω. The voltage potential u satisfies in Ω the equation

$$\nabla \cdot \gamma \nabla u = 0. \tag{1.1}$$

To uniquely fix the solution u it is enough to give its value on the boundary. Let this be f. In the idealized case, when the contact impedance of the measurement device is zero, one measures for all voltage distributions $u|_{\partial M} = f$ on the boundary the corresponding current flux through the boundary, $\gamma \partial y / \partial \nu$, where ν is the exterior unit normal to $\partial\Omega$. Mathematically this amounts to the knowledge of the Dirichlet–Neumann map Λ corresponding to γ, i.e., the map taking the Dirichlet boundary values to the corresponding Neumann boundary values of the solution to (1.1),

$$\Lambda : \quad u|_{\partial M} \mapsto \gamma \frac{\partial u}{\partial \nu}.$$

*Received by the editors August 7, 2006; accepted for publication (in revised form) April 9, 2007; published electronically July 20, 2007. This research was supported by Finnish Centre of Excellence in Inverse Problems Research (Academy of Finland CoE–project 213476).
http://www.siam.org/journals/siap/67-5/66698.html

†Department of Physics, University of Kuopio, P. O. Box 1627, 70211 Kuopio, Finland (Ville.Kolehmainen@uku.fi).

‡Department of Mathematics, Helsinki University of Technology, 02015 Espoo, Finland (Matti.Lassas@tkk.fi).

§Department of Mathematics, University of Helsinki, 00014 Helsinki, Finland (Petri.Ola@rni.helsinki.fi).

The Calderón inverse problem is then to to reconstruct γ from Λ. The problem was originally proposed by Calderón [5] in 1980 and then solved in dimensions three and higher for isotropic conductivities which are C^∞-smooth in [31] and [22]. The smoothness requirements have been since relaxed, and currently the best known result is [25] with unique determination of conductivities in $W^{3/2,\infty}$; see also [10] for a somewhat different approach to the lack of smoothness. In two dimensions the first global result is due to Nachman [23], and later Astala and Päivärinta showed in [4] that uniqueness holds also for general isotropic L^∞-conductivities. For the corresponding anisotropic case, see [3, 17, 18, 19], and for numerical implementations of the methods with simulated and real data, see [13, 28, 21].

Assuming that the measured Dirichlet-to-Neumann map Λ_{meas} is given, an often used method to solve the EIT problem is to minimize

$$\|\Lambda_{\text{meas}} - \Lambda_\sigma\|^2 + \alpha\|\sigma\|_X^2$$

for σ defined in terms of some triangulation of Ω and $\|\cdot\|_X$ some regularization norm; here Λ_σ is the Dirichlet–Neumann map corresponding to the conductivity σ. One then also fixes the geometry of Ω by assuming that it is, for example, a ball or an ellipsoid. Now, if our measurements have no error, a Bayesian interpretation of this problem as a search of a maximum a posteriori (MAP) estimate suggests that $\alpha = 0$. Usually, the given data Λ_{meas} does not correspond to any isotropic conductivity in the model domain. The reason for this is that there is no conformal map deforming the original domain to the model domain. Therefore, in solving the minimization problem we obtain an incorrect solution σ. This means that a systematic error in modeling causes a systematic error in the reconstruction. In particular, if we consider linearization $\gamma = \gamma_0 + \varepsilon\gamma_1$, where γ_0 is a given known background conductivity and ε is small, it seems that a localized perturbation γ_1 gives a reconstruction $\sigma = \gamma_0 + \varepsilon\sigma_1$, where the reconstructed perturbation σ_1 is not localized. This is clearly seen in brain-activity measurements; see [9] and [14].

This work is continuation of [15], where the corresponding question in two dimensions was studied: We proved that on the model domain there is a unique (anisotropic) conductivity with minimal anisotropy. This follows from a result of Strebel saying that among all quasi-conformal self-maps of the unit disk with a fixed boundary value there is a unique one with minimal complex dilation. In higher dimensions there are several new issues. First, the nonuniqueness due to anisotropy is not understood, except in the case when both the domain and the conductivity function are the real analytic [19, 20]. Also, as we already mentioned, in the plane case one could use the theory of quasi-conformal maps to break the nonuniqueness. The higher dimensional analogue of this is unknown. Finally, there is no analogue of the Riemann mapping theorem that we could use.

The structure of this paper is the following. In the first part, consisting of sections 2–4, we present the uniqueness results that we have on the problem. It is worth noting that we choose to work with the Robin-to-Neumann (RN) map instead of the Dirichlet-to-Neumann (DN) map described above. Mathematically they are equivalent, as we will show, but the RN map is a better model for the actual measurement configuration, since it takes into account the contact impedances at $\partial\Omega$ [29]. Also, we assume that the function modeling the contact impedances of the electrodes is known. This means that we have measured the contact impedance, e.g., using a reference body. There are two key ideas to compensate for our lack of understanding of the full anisotropic problem. The first is to note that if an isotropic conductivity is pushed forward by

a diffeomorphism, the resulting conductivity is still conformally flat, and in three dimensions this is equivalent to the vanishing of the Cotton–York tensor. Second, we assume that our original domain is strictly convex, and then the Cohn–Vossen theorem (see [27]) can be used to determine the original boundary $\partial\Omega$ up to rigid motions.

In the second part we develop an algorithm for finding the shape of the domain Ω and the conductivity inside using a minimization technique. An important feature is that we do not have to construct an embedding of the boundary to the Euclidean space. We plan to report on the numerical implementation of our algorithm in a separate article.

2. Measurements. Let $\Omega \subset \mathbb{R}^n$, $n \geq 3$, be a strictly convex domain, and denote by $\gamma = (\gamma^{ij}(x))_{i,j=1}^n$ the symmetric real valued matrix describing the conductivity in Ω. We assume that the matrix is bounded from above and from below; that is, for some $C, c > 0$ we have

$$c\|\xi\|^2 \leq \langle \xi, \gamma(x)\xi \rangle \leq C\|\xi\|^2 \quad \text{for all } x \in \Omega. \tag{2.1}$$

We will state the precise smoothness of γ later. We start by considering the EIT problem with continuous boundary data. Instead of the DN map we will use the RN map defined below, which corresponds better to the measurements done in practice. We discuss later in this section the relation of the continuous model and the electrode measurements made in practice.

For the electrical potential u we write the model

$$\nabla \cdot \gamma \nabla u = 0, \qquad x \in \Omega, \tag{2.2}$$

$$(z\nu \cdot \gamma \nabla u + u)|_{\partial\Omega} = h, \tag{2.3}$$

where h is the Robin-boundary value of the potential and z is a function describing the contact impedance on the boundary. The contact impedance models the impedance that is caused by electro-chemical phenomena at the interface of the skin and the measurement electrodes in practical measurements [6].

In mathematical terms, the perfect boundary measurements are modeled by the RN map $R = R_{z,\gamma}$ given by

$$R : h \mapsto \nu \cdot \gamma \nabla u|_{\partial\Omega},$$

which maps the potential on the boundary to the current across the boundary. Next we relate this continuous model to measurements done in practice.

The physically realistic measurements are usually modeled by the following *complete electrode model* (see [6, 29]): Let $e_j \subset \partial\Omega$, $j = 1, \ldots, J$, be disjoint open sets of the boundary modeling the electrodes that are used for the measurements. Let u solve the equation

$$\nabla \cdot \gamma \nabla v = 0 \qquad \text{in } \Omega, \tag{2.4}$$

$$z_j \nu \cdot \gamma \nabla v + v|_{e_j} = V_j, \tag{2.5}$$

$$\nu \cdot \gamma \nabla v|_{\partial\Omega \setminus \cup_{j=1}^J e_j} = 0, \tag{2.6}$$

where V_j are constants representing electric potentials on electrode e_j. Then, one measures the currents observed on the electrodes, given by

$$I_j = \frac{1}{|e_j|} \int_{e_j} \nu \cdot \gamma \nabla v(x)\, ds(x), \quad j = 1, \ldots, J.$$

Thus the electrode measurements are given by map $E : \mathbb{R}^J \to \mathbb{R}^J$, $E(V_1, \dots, V_J) = (I_1, \dots, I_J)$. We say that E is the electrode measurement matrix for $(\partial\Omega, \gamma, e_1, \dots, e_J, z_1, \dots, z_J)$.

The complete electrode model can alternatively be defined as follows: The RN map R_η is given by $R_\eta f = \nu \cdot \gamma \nabla u|_{\partial\Omega}$, where u is the solution of

$$\begin{aligned} \nabla \cdot \gamma \nabla u &= 0 \qquad \text{in } \Omega, \\ z\nu \cdot \gamma \nabla v + \eta v|_{\partial\Omega} &= h, \end{aligned} \tag{2.7}$$

where $z \in C^\infty(\partial\Omega)$ is such that its restriction to the electrode e_j is equal to the constant z_j and $\eta = \sum_{j=1}^J \chi_{e_j}$, where χ_{e_j} is the characteristic function of electrode e_j.

We associate with the electrode measurement matrix and with the complete electrode model also the corresponding quadratic forms $E : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}$ and $R_\eta : H^{-1/2}(\partial\Omega) \times H^{-1/2}(\partial\Omega) \to \mathbb{R}$ given by

$$E[V, \widetilde{V}] = \sum_{j=1}^J (EV)_j \widetilde{V}_j |e_j|, \quad R_\eta[h, \widetilde{h}] = \int_{\partial\Omega} (R_\eta h)\, \widetilde{h}\, ds. \tag{2.8}$$

These have the following simple relation to each other: Let $S = \operatorname{span}(\chi_{e_j} : j = 1, \dots, J) \subset H^{-1/2}(\partial\Omega)$ and define $M : V = (V_j)_{j=1}^J \mapsto \sum_{j=1}^J V_j \chi_{e_j}$ to be a map $M : \mathbb{R}^J \to S$. Then

$$E[V, \widetilde{V}] = R_\eta[MV, M\widetilde{V}]. \tag{2.9}$$

By (2.9), the electrode measurement matrix can be viewed as the discretization of the form R_η. By increasing the number of the electrodes and making the gaps between them smaller, we can assume that $\eta \to 1$. In this case R_η approximates the RN map $R_{\gamma,z}$. Note that $E(V, V)$ corresponds to the power needed to maintain the voltages V in electrodes.

In practical EIT experiments, one places a set of measurement electrodes on the boundary $\partial\Omega$, e.g., around the chest of the patient. All the traditional approaches to numerical EIT reconstruction assume that the shape of the domain Ω is known and that the only unknown is the conductivity γ. However, in most EIT experiments the boundary of the body Ω is not known accurately, and since there are no practically reliable measurement methods available for the determination of the boundary, the EIT image reconstruction problem is typically solved using an approximate model domain $\widetilde{\Omega}$, which represents our best guess for the shape of the true body Ω. However, it has been noticed that the use of a slightly incorrect model for the body Ω in the numerical reconstruction can lead to serious artifacts in reconstructed images [14, 1, 9]. This situation is our paradigm for the EIT problem when the boundary is unknown. Next we analyze how the deformation of the domain affects measurements.

3. Deformations of the domain. In this section we analyze the behavior of the electrode models under a diffeomorphism. Let's consider first the RN map R. The corresponding quadratic form, which we still denote by R, is given on the diagonal by

$$R[h, h] = \int_{\partial\Omega} (u + z\nu \cdot \gamma \nabla u) \nu \cdot \gamma \nabla u \, dS_E = \int_\Omega \gamma \nabla u \cdot \nabla u \, dx + \int_{\partial\Omega} z \, |\nu \cdot \gamma \nabla u|^2 \, dS_E, \tag{3.1}$$

where $h \in H^{-1/2}(\partial\Omega)$, u solves (2.7), and dS_E is the Euclidean volume form (or area) of $\partial\Omega$. The value $R[h, h]$ corresponds to the power needed to maintain the current

h on the boundary. From the mathematical viewpoint, using the (incorrect) model domain $\widetilde{\Omega}$ instead of the original domain Ω can be viewed as a deformation of the original domain. Thus, let us next consider what happens to the conductivity equation when the domain Ω is deformed to $\widetilde{\Omega}$. Assume that $F : \Omega \to \widetilde{\Omega}$ is a sufficiently smooth orientation-preserving map with sufficiently smooth inverse $F^{-1} : \widetilde{\Omega} \to \Omega$. Let $f : \partial\Omega \to \partial\widetilde{\Omega}$ be the restriction of F on the boundary. When u is a solution of $\nabla\cdot\gamma\nabla u = 0$ in Ω, then $\widetilde{u}(\widetilde{x}) = u(F^{-1}(\widetilde{x}))$ and $\widetilde{h}(x) = h(f^{-1}(x))$ satisfy the conductivity equation

$$\begin{aligned} &\nabla\cdot\widetilde{\gamma}\nabla\widetilde{u} = 0 \quad \text{in } \widetilde{\Omega}, \\ &\widetilde{z}\widetilde{\nu}\cdot\widetilde{\gamma}\nabla\widetilde{u} + \widetilde{u}|_{\partial\widetilde{\Omega}} = \widetilde{h}, \end{aligned} \tag{3.2}$$

where $\widetilde{\nu}$ is the unit normal vector of $\partial\widetilde{\Omega}$, $\widetilde{z}$ is the deformed contact impedance, and $\widetilde{\gamma}$ is the conductivity

$$\widetilde{\gamma}(x) = \left.\frac{F'(y)\,\gamma(y)\,(F'(y))^{\mathrm{T}}}{|\det F'(y)|}\right|_{y=F^{-1}(x)}, \tag{3.3}$$

where $F' = DF$ is the Jacobian of the map F. This transformation formula can be seen from the weak definition of $\nabla\cdot\gamma\nabla u = 0$ in Ω; i.e., for all $\phi \in C_0^\infty(\Omega)$

$$\begin{aligned} 0 &= \int_\Omega \gamma\nabla u\cdot\nabla\phi\,dx = \int_{\widetilde{\Omega}} \gamma(F^{-1}(y))(F'(y)^T\nabla(u(F^{-1}(y))))\cdot F'(y)^T\nabla(\phi(F^{-1}(y)))\,dy \\ &= \int_{\widetilde{\Omega}} \widetilde{\gamma}\nabla\widetilde{u}\cdot\nabla\widetilde{\phi}\,dy, \end{aligned}$$

where $\widetilde{\phi}(y) = \phi(F^{-1}(y))$; see also [30], for more on the transformation rule (3.3) and its relations to inverse problems. Note that even if γ is isotropic, i.e., scalar valued, the deformed conductivity $\widetilde{\gamma}$ can be anisotropic, i.e., matrix valued.

To determine the deformed contact impedance $\widetilde{z}$, we consider the corresponding invariant $(n-1)$-form

$$J := \nu\cdot\gamma\nabla u\,dS_E \in \Omega^{n-1}(\partial\Omega)$$

corresponding to the current flux through the boundary. Next we denote $\widetilde{x} = F(x)$. A straightforward application of the chain rule gives that

$$\widetilde{\nu}\cdot\widetilde{\gamma}\nabla\widetilde{u}|_{\partial\widetilde{\Omega}} = \big((\det DF)^{-1}\nu\cdot\nabla u\big)\circ f^{-1}|_{\partial\widetilde{\Omega}}$$

since F was orientation preserving and DF is the Jacobian of F in boundary normal coordinates associated with the surface $\partial\Omega \subset \mathbb{R}^n$. In these coordinates $\det DF|_{\partial\Omega} = \det Df$, where $\det Df$ is the determinant of the differential of the the boundary map $f : \partial\Omega \to \partial\widetilde{\Omega}$. We note that $(\det Df\circ f^{-1})f_*(dS_E) = d\widetilde{S}_E$, where dS_E and $d\widetilde{S}_E$ are Euclidean volume forms of $\partial\Omega$ and $\partial\widetilde{\Omega}$, respectively. Hence, $z\nu\cdot\nabla u$ transforms as an invariantly defined function when the contact impedance is interpreted as a density, i.e.,

$$\widetilde{z}(\widetilde{x}) = (\det Df(x))z(x), \tag{3.4}$$

where $f(x) = \widetilde{x}$. Now we see that the boundary measurements are invariant: When $f : \partial\Omega \to \partial\widetilde{\Omega}$ is the restriction of $F : \Omega \to \widetilde{\Omega}$, we say that the map $\widetilde{R} = f_*R_{z,\gamma}$, defined by

$$((f_*R_{z,\gamma})h)(x) = (R_{z,\gamma}(h\circ f))(y)|_{y=f^{-1}(x)}, \quad h \in H^{1/2}(\partial\widetilde{\Omega}),$$

has the property that $\widetilde{R} = R_{\widetilde{z},\widetilde{\gamma}}$. We call $\widetilde{R}$ the push forward of $R_{z,\gamma}$ by f.

Is is also worth noting that in formula (3.1) the integral over Ω as well as the integral over the boundary are invariant because of the deformation rule (3.4) for the contact impedance z; that is, we have

$$R[h, h'] = \widetilde{R}[h \circ f^{-1}, h' \circ f^{-1}]$$

for $h, h' \in H^{-1/2}(\partial\Omega)$.

4. Uniqueness results. Now we are ready to give the exact set-up of the problem we consider: We want to recover an image of the unknown conductivity γ in Ω from the measurements of the RN map, and we assume a priori that γ is isotropic. We assume that z, $\partial\Omega$, and R are not known. Instead, let $\widetilde{\Omega}$, called the model domain, be our best guess for the domain, and let $f_m : \partial\Omega \to \partial\Omega_m$ be a diffeomorphism modeling the approximate knowledge of the boundary.

As the data for the inverse problem, we assume that we are given the boundary of the model domain $\partial\widetilde{\Omega}$, the function $z \circ f^{-1}$ corresponding to the contact impedance of electrodes, and the RN map $\widetilde{R} = (f_m)_* R$. Note that the discrete analogue of this data is to know the voltage-to-power form $V \mapsto E(V, V)$ and the contact impedances of the electrodes, but not the location of the electrodes or the boundary of the domain. It is reasonable to assume that the contact impedance $z \circ f_m^{-1}$ on the boundary of the model domain is known since we can observe and set up the contact impedances of the electrodes the way we want. Hence we have on the boundary of our model domain $\partial\widetilde{\Omega}$ a boundary map $\widetilde{R}$ that does not generally correspond to any isotropic conductivity. Furthermore, we saw above that there are many anisotropic conductivities for which the RN map is the given map $\widetilde{R}$. Next we show that the existence of the "underlying" isotropic conductivity in Ω gives the uniqueness in $\widetilde{\Omega}$ up to a diffeomorphism and that the domain Ω and the isotropic conductivity on it can be uniquely determined.

THEOREM 4.1. *Let $\Omega \subset \mathbb{R}^n$, $n \geq 3$, be a bounded strictly convex C^∞-domain. Assume that $\gamma \in C^\infty(\overline{\Omega})$ is an isotropic conductivity, $z \in C^\infty(\partial\Omega)$, $z > 0$ a contact impedance, and $R_{\gamma,z}$ the corresponding RN map. Let $\widetilde{\Omega}$ be a model of the domain satisfying the same regularity assumptions as Ω, and $f_m : \partial\Omega \to \partial\widetilde{\Omega}$ be a C^∞-smooth orientation-preserving diffeomorphism.*

Assume that we are given $\partial\widetilde{\Omega}$, the values of the contact impedance $z(f_m^{-1}(\widetilde{x}))$, $\widetilde{x} \in \partial\widetilde{\Omega}$, and the map $\widetilde{R} = (f_m)_* R_{\gamma,z}$. Then we can determine Ω up to a rigid motion T and the conductivity $\gamma \circ T^{-1}$ on the reconstructed domain $T(\Omega)$.

We recall also that rigid motion is an affine isometry $T : \mathbb{R}^n \to \mathbb{R}^n$.

Proof. Assume that we are given $\widetilde{R}$ and the values of the contact impedance, that is, the function $z(f_m^{-1}(\widetilde{x}))$. Let $F_m : \Omega \to \widetilde{\Omega}$ be an orientation-preserving diffeomorphism satisfying $F_m|_{\partial\Omega} = f_m$. As noted before, $\widetilde{R} = R_{\widetilde{z},\widetilde{\gamma}}$, where $\widetilde{z}(x) = \det(Df_m) z(f_m^{-1}(x))$ is the contact impedance on $\partial\widetilde{\Omega}$ and $\widetilde{\gamma} = (F_m)_* \gamma$ is the push forward of γ in F_m. The RN map is a classical pseudodifferential operator of order zero, with principal symbol $1/\widetilde{z}$, and hence $\widetilde{R}$ determines $\widetilde{z}$. Since we also assume $z \circ f^{-1}$ known, we can determine the determinant $\det Df_m$; note that this gives the change of boundary area under deformation f_m. For the rest of the proof denote $\beta = \det Df_m$. Also, this implies that we can find the DN map $\Lambda_{\widetilde{\gamma}} = (\widetilde{R}^{-1} - \widetilde{z} I)^{-1}$ on $\partial\widetilde{\Omega}$, that is, the map taking the Dirichlet boundary values to Neumann boundary values. In terms of electrostatics, this means that the impedance given by Λ_γ^{-1} and the contact impedance z are connected in series; i.e., the total impedance is the sum of Λ_γ^{-1} and z.

The Riemannian metric corresponding to the isotropic conductivity $\gamma = \gamma(x)I$ in Ω is given by

$$g_{ij}(x) = \det(\gamma(x)I)^{1/(2-n)}\,(\gamma(x)I)^{-1} = \gamma(x)^{2/(n-2)}\delta_{ij}.$$

Then, if Δ_g is the Laplace–Beltrami operator corresponding to the metric g, we have $\Delta_g = |g|^{-1/2}\nabla\cdot\gamma\nabla$, where $|g| = \det(g_{ij})$. This metric is an invariant object, and in the deformation F_m it is transformed to the metric $\widetilde{g} = (F_m)_* g$ in $\widetilde{\Omega}$. By [19], the DN map $\Lambda_{\widetilde{\gamma}}$ determines the restriction of the n-dimensional metric tensor $\widetilde{g}_{jk}$ on the boundary in the boundary normal coordinates, and since in these coordinates

$$[\widetilde{g}_{jk}]_{j,k=1}^{n} = \begin{pmatrix} \widetilde{g}_{\partial\widetilde{\Omega}} & 0 \\ 0 & 1 \end{pmatrix},$$

where $\widetilde{g}_{\partial\widetilde{\Omega}} = \widetilde{h}$ is the induced metric on $\partial\widetilde{\Omega}$, we can also recover $\widetilde{h}$.

In particular, if we consider $\partial\widetilde{\Omega}$ as a submanifold of $\mathbb{R}^n$ with the metric $\widetilde{h} = \widetilde{i}^*(\widetilde{g})$ inherited from $(\widetilde{\Omega}, \widetilde{g})$ where $\widetilde{i}: \partial\Omega \to \Omega$ is the identity map, we see that our boundary data determines the metric $\widetilde{h}$ on $\partial\widetilde{\Omega}$. Now let metric $h = i^*(g)$ be the corresponding metric on $\partial\Omega$, where $i: \partial\Omega \to \Omega$ is the identical embedding. Then we have

$$\widetilde{h} = (f_m)_* h, \quad h = \gamma^{2/(n-2)} h^E, \tag{4.1}$$

where h^E is the Euclidean metric of $\partial\Omega$. Denote by $\widetilde{h}^E = (f_m)_* h^E$ the metric tensor on $\partial\widetilde{\Omega}$, i.e., the push forward of the Euclidean metric of $\partial\Omega$ by f_m. Recall that dS_E and $d\widetilde{S}_E$ are the Euclidean volume forms of $\partial\Omega$ and $\partial\widetilde{\Omega}$, respectively. Then the Riemannian volume forms $dS_{\widetilde{h}}$ and dS_h of the metrics $\widetilde{h}$ and h, respectively, satisfy

$$dS_{\widetilde{h}} = (f_m)_*(dS_h) = \gamma(f_m^{-1}(\widetilde{x}))(f_m)_*(dS_E) = (\gamma\beta)\circ f_m^{-1}(\widetilde{x})\, d\widetilde{S}_E$$

on $\widetilde{\partial}\Omega$. As β was already determined, this shows that we can find $\gamma(f_m^{-1}(\widetilde{x}))$, $\widetilde{x} \in \partial\widetilde{\Omega}$, and hence by (4.1) we can determine the metric

$$\widetilde{h}^E = \gamma(f_m^{-1}(\widetilde{x}))^{-2/(n-2)}\widetilde{h}.$$

In other words, if we consider $\partial\widetilde{\Omega}$ as an abstract manifold that can be embedded to $\partial\Omega \subset \mathbb{R}^n$, we have found the metric tensor on $\partial\widetilde{\Omega}$ corresponding to the Euclidean metric of $\partial\Omega$. By the Cohn–Vossen rigidity theorem, intrinsically isometric C^2-smooth surfaces that are boundaries of a strictly convex body are congruent in a rigid motion. For uniqueness, see, e.g., [27, Theorems V and VI] and also [11, 12]. This means that the boundary data uniquely determines the map $T \circ f_m^{-1}$, where T is a rigid motion. Hence we can find the surface $T(\partial\Omega)$ and on it the map $T_*\Lambda_{\widetilde{\gamma}} = T_*\Lambda_\gamma$. Using the uniqueness of of the isotropic inverse problem [31, 22], we see that the boundary data determines $\gamma \circ T^{-1}$. ∎

Note that the construction of the surface $\partial\Omega \subset \mathbb{R}^3$ from the intrinsic metric h^E is a more delicate issue (see [24, 26]); hence we take care to avoid it.

5. A reconstruction algorithm and the use of conformal flatness. In this section we consider the case $n = 3$, even though the considerations could be generalized for $n \geq 4$ by changing the Cotton–York tensor to a Weyl tensor in our considerations (see the appendix). As noted before, an actual construction of the

isometric embedding of an abstract manifold to Euclidean space is complicated, and thus we try to avoid it.

We want to find an anisotropic conductivity η such that $R_{\widetilde{z},\eta} = \widetilde{R}$ assuming that $\widetilde{R} = (f_m)_* R_{z,\gamma}$, where γ is an isotropic conductivity. Clearly, when $F_m : \Omega \to \widetilde{\Omega}$ is diffeomorphism satisfying $F_m|_{\partial\Omega} = f_m$, the anisotropic conductivity $(F_m)_*\gamma$ is a solution of the inverse problem, but it is not unique. However, we also know that $(F_m)_*\gamma$ has a conformally flat structure, and this fact will help in solving the inverse problem as we will see. Note that in principle, one could start to solve the inverse problem by minimizing over all pairs (Ω, σ) of smooth domains $\Omega \subset \mathbb{R}^n$ and all isotropic conductivities σ in Ω. However, the minimization over domains is complicated, and our objective is to find a reasonably simple minimization algorithm where we minimize over conductivities in the fixed model domain $\widetilde{\Omega}$ with an appropriately chosen cost function.

Let $\eta = (F_m)_*\gamma$ be a possibly anisotropic conductivity in $\widetilde{\Omega}$ such that γ is isotropic. As already noted, it defines a Riemannian metric g on $\widetilde{\Omega}$, given by

$$[g_{jk}]_{j,k=1}^n = ([g^{jk}]_{j,k=1}^n)^{-1}, \quad g^{jk} = \det(\eta)^{1/(n-2)}\eta^{jk}.$$

From now on we will use the Einstein summation convention and omit the summation symbols. As $F_m^{-1} : \widetilde{\Omega} \to \Omega$ can be considered as coordinates, we see that in proper coordinates the metric g is a scalar function times a Euclidean metric; that is, g is conformally flat. This means that

$$g_{ij}(x) = e^{-2\sigma(x)}\overline{g}_{ij}(x),$$

where $\overline{g}_{ij}(x)$ is a metric with zero curvature tensor (i.e., flat metric) and $\sigma(x) \in \mathbb{R}$. By [8] (for original work, see [7]), the conformal flatness of the metric g in three dimensions is equivalent to the vanishing of the Cotton–York tensor $\mathrm{C} = \mathrm{C}_{ij}$ corresponding to g (see the appendix). Note that we can choose $\sigma = \frac{1}{2-n}\log\gamma$ and $\overline{g} = (F_m)_*(\delta_{ij})$. By [8, formulae (28.18) and (14.1)], σ satisfies a differential equation (with $n = 3$)

$$\sigma_{ij} = -\frac{1}{n-2}\,\mathrm{Ric}_{ij} + \frac{1}{2(n-1)(n-2)}\,g_{ij}\mathrm{R} - \frac{1}{2}\,g_{ij}g^{lm}\sigma_l\sigma_k, \quad i,j = 1,\dots,n, \tag{5.1}$$

where Ric_{ij} is the Ricci curvature tensor of g, R is the scalar curvature of g, and

$$\sigma_k = \frac{\partial\sigma}{\partial x^k}, \quad \sigma_{ij} = \nabla_{e_i}\sigma_j - \sigma_i\sigma_j, \quad \text{where} \quad e_i = \frac{\partial}{\partial x^i},$$

where ∇_{e_i} is the covariant derivative with respect to metric g. Thus if g is given, (5.1) is a second order nonlinear differential equation for σ. By [8, p. 92], the equations (5.1) satisfy the sufficient integrability conditions to be locally solvable if and only if the Cotton–York tensor vanishes. Note that the existence of the isotropic conductivity γ in Ω gives a solution for these equations.

Consider now the following algorithm.

Data: Assume that we are given $\partial\Omega_m$, $\widetilde{R} = (f_m)_* R_{\gamma,z}$, and $z \circ f_m^{-1}$ on $\partial\Omega_m$.

Aim: We look for a metric $\widetilde{g}$ corresponding to the conductivity $\widetilde{\gamma}$ and $\widetilde{z}$ such that on $\partial\Omega_m$, $\widetilde{R} = R_{\widetilde{\gamma},\widetilde{z}}$ and $\widetilde{z} = (f_m)_* z$.

Algorithm.

1. Determine the two leading terms in the symbolic expansion of $\widetilde{R}$. They determine a contact impedance $\widehat{z}$ and a metric $\widehat{g}$ on $\partial\widetilde{\Omega}$ such that if $\widetilde{R} = R_{\widetilde{\gamma},\widetilde{z}}$, then $\widetilde{z} = \widehat{z}$ and $\widetilde{i}^*(\widetilde{g}) = \widehat{g}$.

2. Form the ratio of the given contact impedance, $\widetilde{z}$, and the reconstructed contact impedance, $\widehat{z}$, that is,

$$\widehat{r}(\widetilde{x}) := \frac{z(f_m^{-1}(\widetilde{x}))}{\widehat{z}(\widetilde{x})}, \quad \widetilde{x} \in \partial\widetilde{\Omega}.$$

Note that then

$$\widehat{r}(\widetilde{x})(f_m)_*(dS_E) = d\widetilde{S}_E$$

since the contact impedances transformed as densities.

3. Let $dS_{\widehat{g}}$ be the volume form of $\widehat{g}$ on $\partial\widetilde{\Omega}$. Then

$$dS_{\widehat{g}} = (\det \widehat{g})^{1/2}\, d\widetilde{S}_E.$$

Define

$$\widehat{\gamma} = (\det \widehat{g})^{1/2}\, \widehat{r}.$$

With this choice $\widehat{\gamma}$ will satisfy $\widehat{\gamma}(\widetilde{x}) = \gamma(f_m^{-1}(\widetilde{x}))$ for $\widetilde{x} \in \partial\widetilde{\Omega}$.

4. Define the boundary value $\widehat{\sigma}$ for the function σ by

$$\widehat{\sigma}(\widetilde{x}) = \frac{1}{2-n} \log\left(\widehat{\gamma}(\widetilde{x})\right), \quad \widetilde{x} \in \partial\widetilde{\Omega}.$$

5. Solve the minimization problem

$$\min F_\tau(\widetilde{z}, \widetilde{\sigma}, \widetilde{\gamma}) + \alpha\, H(\widetilde{z}, \widetilde{\sigma}, \widetilde{\gamma}),$$

where $H(\widetilde{z}, \widetilde{\gamma})$ is a regularization functional, say,

$$H(\widetilde{z}, \widetilde{\sigma}, \widetilde{\gamma}) = \|\widetilde{z}\|^2_{H^8(\widetilde{\Omega})} + \|\widetilde{\gamma}\|^2_{H^8(\widetilde{\Omega})} + \|\widetilde{\sigma}\|^2_{H^8(\widetilde{\Omega})};$$

$\alpha \geq 0$ is a regularization parameter; and

$$\begin{aligned} &F_\tau(\widetilde{z}, \widetilde{\sigma}, \widetilde{\gamma}) \\ &= \left\| \widetilde{R} - R_{\widetilde{\gamma},\widetilde{z}} \right\|^2_{L(H^{-1/2}(\partial\widetilde{\Omega}))} + \| \frac{\widetilde{z}(\widetilde{x})}{z(f_m^{-1}(\widetilde{x}))} - \widehat{r}(\widetilde{x}) \|^2_{L^2(\partial\widetilde{\Omega})} + \|\widetilde{\sigma}|_{\partial\widetilde{\Omega}} - \widehat{\sigma}\|^2_{L^2(\partial\widetilde{\Omega})} \\ &\quad + \tau \|\mathrm{C}\|^2_{L^2(\widetilde{\Omega})} \\ &\quad + \sum_{i,j=1}^{n} \left\| \widetilde{\sigma}_{ij} - \left(-\frac{1}{n-2}\mathrm{Ric}_{ij} + \frac{1}{2(n-1)(n-2)} \widetilde{g}_{ij}\mathrm{R} - \frac{1}{2}\widetilde{g}_{ij}\widetilde{g}^{lm}\sigma_l\sigma_k \right) \right\|^2_{L^2(\widetilde{\Omega})}, \end{aligned}$$

where $\tau \geq 0$, $\widetilde{g}$ is the metric tensor corresponding to $\widetilde{\gamma}$, $\mathrm{C} = \mathrm{C}_{ij}$ is the Cotton–York tensor of $\widetilde{g}$, and finally Ric and R are the Ricci curvature and scalar curvature tensors, respectively, of $\widetilde{g}$.

Note that step 1 above requires the use of highly oscillating boundary data, and hence very small values of z might cause problems in practice. Also, the value of the Cotton–York tensor at $x \in \Omega$, $\mathrm{C}_{ij}(x)$, the Ricci curvature tensors $R_{ij}(x)$, and the scalar curvature $R(x)$ depend on the values of the conductivity η and its three first derivatives at x.

PROPOSITION 5.1. *Let $\Omega \subset \mathbb{R}^3$ be a bounded strictly convex C^∞-domain. Assume that $\gamma \in C^\infty(\overline{\Omega})$ is an isotropic conductivity, $z \in C^\infty(\partial\Omega)$, $z > 0$ a contact*

impedance, and $R_{\gamma,z}$ the corresponding RN map. Let $\widetilde{\Omega}$ be a model of the domain satisfying the same regularity assumptions as Ω, and $f_m : \partial\Omega \to \partial\widetilde{\Omega}$ be a C^∞-smooth diffeomorphism.

Assume that we are given $\partial\widetilde{\Omega}$, the values of the contact impedance $z(f_m^{-1}(\widetilde{x}))$, $\widetilde{x} \in \partial\widetilde{\Omega}$, and the map $\widetilde{R} = (f_m)_* R_{\gamma,z}$.

Let $\tau \geq 0$. Then the minimum of $F_\tau(\widetilde{z}, \widetilde{\sigma}, \widetilde{\gamma})$ is zero; any minimizers $\widetilde{z}$, $\widetilde{\sigma}$, and $\widetilde{\gamma}$ of $F_\tau(\widetilde{z}, \widetilde{\sigma}, \widetilde{\gamma})$ satisfy $\widetilde{z} = (f_m)_* z$; and there is a diffeomorphism $\widetilde{F} : \Omega \to \widetilde{\Omega}$ such that $\widetilde{F}|_{\partial\Omega} = f_m$, $\widetilde{\gamma} = \widetilde{F}_* \gamma$, and $\widetilde{\sigma} = -\log \widetilde{\gamma}$.

Proof. Assume first that $\tau > 0$. The minimizer exists because of the existence of Ω, γ, z, and σ, and the minimum is zero. Let $\widetilde{z}$, $\widetilde{\sigma}$, and $\widetilde{g}$ be some minimizers of F_τ. As then the Cotton–York tensor is zero and the equations (5.1) are valid, it follows from [8] that the metric $\overline{g}_{ij} = \exp(2\sigma(\widetilde{x}))g_{ij}(\widetilde{x})$, $x \in \widetilde{\Omega}$, is flat. Since $R_{\widetilde{z},\widetilde{\gamma}} = \widetilde{R}$, we have $\widetilde{z} = (f_m)_* z$, and the metric $\widetilde{g}$ corresponding to $\widetilde{\gamma}$ has to satisfy $i^*\widetilde{g} = \widehat{g}$ on the boundary. This and the vanishing of F_τ imply that

$$\begin{aligned} i^*\overline{g} &= \exp(2\widehat{\sigma}) i^*\widetilde{g} = \exp(2\widehat{\sigma})\widehat{g} = \exp(2\widehat{\sigma})(f_m)_*(\gamma h_E) \\ &= \exp(2\widehat{\sigma})\,\widehat{\gamma}\,(f_m)_*(h_E) = (f_m)_*(h_E). \end{aligned}$$

Consider now $(\widetilde{\Omega}, \overline{g})$ as a Riemannian manifold. As $\overline{g}$ is flat, we know that $(\widetilde{\Omega}, \overline{g})$ can be embedded isometrically to domain $\Omega_0 \subset \mathbb{R}^n$. Let $k : \widetilde{\Omega} \to \Omega_0$ be this embedding. Since $i^*\overline{g} = (f_m)_*(h_E)$, it follows from the Cohn–Vossen rigidity theorem that the boundaries $\partial\Omega_0$ and $\partial\Omega$ are congruent in a rigid motion T and $k \circ f_m = T|_{\partial\Omega}$. Then $(T^{-1} \circ k)_* \widetilde{\gamma}$ is an isotropic conductivity, the contact impedances of $(T^{-1} \circ k)_* \widetilde{z}$ and z coincide, and the RN maps of $(T^{-1} \circ k)_* \widetilde{\sigma}$ and σ coincide. By the uniqueness of the isotropic inverse conductivity problem [31], $(T^{-1} \circ k)_* \widetilde{\gamma} = \gamma$. This proves the claim in the case $\tau > 0$.

Next, consider the case $\tau = 0$. Again, a minimizer exists because of the existence of Ω, γ, z, and σ, and the minimum is zero. Let $\widetilde{z}$, $\widetilde{\sigma}$, and $\widetilde{g}$ be some minimizers. As the minimum of F_τ is zero, the equations (5.1) are valid. By [8, p. 92], the solutions σ satisfy the integrability conditions

$$\nabla_k \sigma_{ij} - \nabla_j \sigma_{ik} = \sigma_l \mathrm{R}^l_{ijk}, \quad i, j, k = 1, \ldots, n, \tag{5.2}$$

which imply that the conformal covariant satisfying R_{ijk} vanishes. Thus the Cotton–York tensor C_{ij} is zero. This means that the minimizers $\widetilde{z}$, $\widetilde{\sigma}$, and $\widetilde{g}$ of F_τ with $\tau = 0$ are also minimizers of F_τ with any $\tau > 0$. □

One can think of τ as a regularization parameter: In general the solvability properties of (5.1) are sensitive to the compatibility conditions, i.e., the vanishing of the Cotton–York (or the Weyl tensor in higher dimensions).

To find the domain Ω, we can continue the above algorithm by applying the fact that conformally Euclidean manifold of dimension n can be a conformally embedded to $\mathbb{R}^n$ in a constructive way (cf. [16]).

6. In steps 1–5 we have found metric tensors $\widetilde{g}$ and $\overline{g} = e^{2\widetilde{\sigma}}\widetilde{g}$ on $\widetilde{\Omega}$ such that $\widetilde{g} = F_*(g)$ and $\overline{g} = F_*(g^E)$, where g is the metric corresponding to the metric γ on Ω, g^E is the Euclidean metric on Ω, and $F : \Omega \to \widetilde{\Omega}$ is some diffeomorphism.

   Let $y \in \widetilde{\Omega}$ and find geodesics $\overline{\mu}_{y,\xi}(s)$ starting from y with respect to the metric $\overline{g}$. We parametrize these geodesics in such a way that $\overline{\mu}_{y,\xi}(0) = y$ and $\partial_s \overline{\mu}_{y,\xi}(0) = \xi$ is a unit tangent vector of the tangent space $(T_y\widetilde{\Omega}, \overline{g})$. These

geodesics correspond to the half-lines in $\mathbb{R}^3$ starting from some point $y_0 \in \Omega$. Let $J : (T_y\widetilde{\Omega}, \overline{g}) \to \mathbb{R}^3$ be a linear isometry, and define a map $\kappa : \widetilde{\Omega} \to \mathbb{R}^3$ by setting

$$\kappa(\overline{\mu}_{y,\xi}(s)) \,=\, s\, J\xi, \quad s \geq 0.$$

Then $\kappa \circ F : \Omega \to \mathbb{R}^3$ is an affine isometry that extends to a rigid motion $T : \mathbb{R}^3 \to \mathbb{R}^3$ with $T(y) = 0$. Thus we can find $\kappa(\widetilde{\Omega}) = T(\Omega)$, $\kappa_*(\widetilde{\sigma}) = T_*\gamma$, and $\kappa_*(\widetilde{z}) = T_*z$

Thus we have shown the following reconstruction result.

COROLLARY 5.2. *Let Ω, γ, z, $\widetilde{\Omega}$, and f_m be as in Proposition* 5.1. *Assume that we are given $\partial\widetilde{\Omega}$, the contact impedance $z(f_m^{-1}(\widetilde{x}))$, $\widetilde{x} \in \partial\widetilde{\Omega}$, and the RN map $\widetilde{R} = (f_m)_* R_{\gamma,z}$. Then the algorithm* 1–6 *determines Ω, γ, and z up to a rigid motion $T : \mathbb{R}^3 \to \mathbb{R}^3$.*

We intend to investigate the numerical implementation of the method and give numerical test results in part II of this paper.

Appendix. Here we define the conformal curvature tensors. We say that a metric g_{ij} in a domain $\Omega \subset \mathbb{R}^n$ is conformally flat if there is a scalar function $a(x) > 0$ such that the curvature of tensor of $a(x)g_{ij}(x)$ is identically zero.

First, let γ be an isotropic conductivity, i.e., a smooth positive function in Ω, and let $F : \Omega \to \widetilde{\Omega}$ be a diffeomorphism. Let $\eta = F_*\gamma$ be a possibly anisotropic conductivity in $\widetilde{\Omega}$. It defines a Riemannian metric $\widetilde{g}$ on $\widetilde{\Omega}$, given by

$$[\widetilde{g}_{jk}]_{j,k=1}^n = ([\widetilde{g}^{jk}]_{j,k=1}^n)^{-1}, \quad \widetilde{g}^{jk} = \det(\eta)^{1/(n-2)}\eta^{jk}.$$

As $F^{-1} : \widetilde{\Omega} \to \Omega$ can be considered as coordinates, we see that in proper coordinates the metric $\widetilde{g}$ is a scalar function times a Euclidean metric; that is, $\widetilde{g}$ is conformally flat.

Next we consider a general metric tensor g_{ij} and recall facts concerning its conformal flatness. Note that below we use the Einstein summation convention and omit the summation symbols when possible. The following tensors are related to conformal flatness:

(a) Assume that $n = 3$. Then the conformal covariant, given in terms of curvature tensors (see the explanation on notation below) is

$$R_{ijk} = \nabla_k R_{ij} - \nabla_j R_{ik} + \frac{1}{2(n-1)}(g_{ik}\nabla_j R - g_{ij}\nabla_k R).$$

In the three dimensional case, R_{ijk} defines a tensor that can be considered as a vector valued 2-form $R_{ijk}dx^j \wedge dx^k$. Operating with the Hodge operator $*$ on this 2-form, we obtain the Cotton–York tensor,

$$\mathrm{C}_{ij} = g^{kp}g^{lq}\nabla_k \left(R_{li} - \frac{1}{4} R\, g_{li} \right) \epsilon_{pqj},$$

where ϵ_{pqj} is the Levi–Civita permutation symbol.

(b) Assume that $n \geq 4$. Then the Weyl tensor is

$$W_{ijkl} = R_{ijkl} + \frac{1}{n-2}(g_{il}R_{kj} + g_{jk}R_{kj} - g_{ik}R_{lj} - g_{jl}R_{ki}) + \frac{1}{(n-1)(n-2)}(g_{ik}g_{lj} - g_{il}g_{kj})R.$$

The crucial fact related to our considerations is that the metric g is conformally flat if and only if in the dimension $n = 3$ the Cotton–York tensor vanishes and in the dimension $n = 4$ the Weyl tensor vanishes; see [8, p. 92] or [7, 32, 2].

Above, R_{ijkl} is the Riemannian curvature tensor,

$$R_{ijkl} = \frac{\partial}{\partial x^k}\Gamma^i_{jl} - \frac{\partial}{\partial x^l}\Gamma^i_{jk} + \Gamma^p_{jl}\Gamma^i_{pk} - \Gamma^p_{jk}\Gamma^i_{pl}, \quad R^p_{jkl} = g^{pi}R_{ijkl},$$

where Γ^i_{jk} are Christoffel symbols,

$$\Gamma^i_{jk} = \frac{1}{2}g^{pi}\left(\frac{\partial g_{jp}}{\partial x^k} + \frac{\partial g_{kp}}{\partial x^j} - \frac{\partial g_{jk}}{\partial x^p}\right),$$

R_{ij} is the Ricci curvature tensor, $R_{ij} = R^k_{ijk}$, and R is the scalar curvature $R = g^{ij}R_{ij}$. Finally, ∇_k is the covariant derivative that is defined for a (0,2)-tensor A_{il} and a (0,1)-tensor B_l by

$$\nabla_k A_{li} = \frac{\partial}{\partial x^k}A_{li} - \Gamma^p_{kl}A_{pi} - \Gamma^p_{ki}A_{lp}, \qquad \nabla_k B_l = \frac{\partial}{\partial x^k}B_l - \Gamma^p_{kl}B_p.$$

Acknowledgment. The authors wish to thank the anonymous referees for their valuable comments.

REFERENCES

[1] A. Adler, R. Guardo, and Y. Berthiaume, *Impedance imaging of lung ventilation: Do we need to account for chest expansion?*, IEEE Trans. Biomed. Engrg., 43 (1996), pp. 414–420.
[2] S. Aldersley, *Comments on certain divergence-free tensor densities in a 3-space*, J. Math. Phys., 20 (1979), pp. 1905–1907.
[3] K. Astala, M. Lassas, and L. Päivärinta, *Calderón's inverse problem for anisotropic conductivity in plane*, Comm. Partial Differential Equations, 30 (2005), pp. 207–224.
[4] K. Astala and L. Päivärinta, *Calderón's inverse conductivity problem in the plane*, Ann. of Math. (2), 163 (2006), pp. 265–299.
[5] A. Calderón, *On an inverse boundary value problem*, in Seminar on Numerical Analysis and Its Applications to Continuum Physics (Rio de Janeiro, 1980), Soc. Brasil Mat., Rio de Janeiro, 1980, pp. 65–73.
[6] K.-S. Cheng, D. Isaacson, J. C. Newell, and D. G. Gisser, *Electrode models for electric current computed tomography*, IEEE Trans. Biomed. Engrg., 36 (1989), pp. 918–924.
[7] E. Cotton, *Sur les varietes a trois dimensions*, Ann. Fac. Sci. Toulouse, 4 (1899), pp. 385–438.
[8] L. Eisenhart, *Riemannian Geometry*, Princeton University Press, Princeton, NJ, 1977.
[9] E. Gersing, B. Hoffman, and M. Osypka, *Influence of changing peripheral geometry on electrical impedance tomography measurements*, Med. Biolog. Engrg. Computing, 34 (1996), pp. 359–361.
[10] A. Greenleaf, M. Lassas, and G. Uhlmann, *The Calderón problem for conormal potentials,* I: *Global uniqueness and reconstruction*, Comm. Pure Appl. Math., 56 (2003), pp. 328–352.
[11] C. Hsü, *Generalization of Cohn-Vossen's theorem*, Proc. Amer. Math. Soc., 11 (1960), pp. 845–846.
[12] J. Iaia, *Isometric embeddings of surfaces with nonnegative curvature in R^3*, Duke Math. J., 67 (1992), pp. 423–459.
[13] D. Isaacson, J. Mueller, J. Newell, and S. Siltanen, *Reconstructions of chest phantoms by the d-bar method for electrical impedance tomography*, IEEE Trans. Medical Imaging, 23 (2004), pp. 821–828.
[14] V. Kolehmainen, M. Vauhkonen, P. Karjalainen, and J. Kaipio, *Assessment of errors in static electrical impedance tomography with adjacent and trigonometric current patterns*, Physiolog. Measurement, 18 (1997), pp. 289–303.
[15] V. Kolehmainen, M. Lassas, and P. Ola, *The inverse conductivity problem with an imperfectly known boundary*, SIAM J. Appl. Math., 66 (2005), pp. 365–383.
[16] A. Katchalov, Y. Kurylev, and M. Lassas, *Inverse Boundary Spectral Problems*, Chapman Hall/CRC Monogr. Surv. Pure Appl. Math. 123, Chapman & Hall, Boca Raton, FL, 2001.

[17] M. Lassas and G. Uhlmann, *On determining a Riemannian manifold from the Dirichlet-to-Neumann map*, Ann. Sci. Ecole Norm. Sup., 34 (2001), pp. 771–787.
[18] M. Lassas, M. E. Taylor, and G. Uhlmann, *The Dirichlet-to-Neumann map for complete Riemannian manifolds with boundary*, Comm. Anal. Geom., 11 (2003), pp. 207–221.
[19] J. Lee and G. Uhlmann, *Determining anisotropic real-analytic conductivities by boundary measurements*, Comm. Pure Appl. Math., 42 (1989), pp. 1097–1112.
[20] W. Lionheart, *Boundary shape and electrical impedance tomography*, Inverse Problems, 14 (1998), pp. 139–147.
[21] J. L. Mueller and S. Siltanen, *Direct reconstructions of conductivities from boundary measurements*, SIAM J. Sci. Comput., 24 (2003), pp. 1232–1266.
[22] A. Nachman, *Reconstructions from boundary measurements*, Ann. of Math. (2), 128 (1988), pp. 531–576.
[23] A. Nachman, *Global uniqueness for a two–dimensional inverse boundary value problem*, Ann. Math., 143 (1996), pp. 71–96.
[24] L. Nirenberg, *The Weyl and Minkowski problems in differential geometry in the large*, Comm. Pure Appl. Math., 6 (1953), pp. 337–394.
[25] L. Päivärinta, A. Panchenko, and G. Uhlmann, *Complex geometrical optics solutions for Lipschitz conductivities*, Rev. Mat. Iberoamericana, 19 (2003), pp. 57–72.
[26] A. Pogorelov, *The rigidity of general convex surfaces*, Dokl. Akad. Nauk SSSR (N.S.), 79 (1951), pp. 739–742 (in Russian).
[27] R. Sacksteder, *The rigidity of hypersurfaces*, J. Math. Mech., 11 (1962), pp. 929–939.
[28] S. Siltanen, J. Mueller, and D. Isaacson, *An implementation of the reconstruction algorithm of A. Nachman for the 2-D inverse conductivity problem*, Inverse Problems, 16 (2000), pp. 681–699.
[29] E. Somersalo, M. Cheney, and D. Isaacson, *Existence and uniqueness for electrode models for electric current computed tomography*, SIAM J. Appl. Math., 52 (1992), pp. 1023–1040.
[30] J. Sylvester, *An anisotropic inverse boundary value problem*, Comm. Pure Appl. Math., 43 (1990), pp. 201–232.
[31] J. Sylvester and G. Uhlmann, *A global uniqueness theorem for an inverse boundary value problem*, Ann. of Math. (2), 125 (1987), pp. 153–169.
[32] J. Thomas, *Conformal invariants*, Proc. Natl. Acad. Sci. USA, 12 (1926), pp. 389–393.

SIAM J. APPL. MATH.
Vol. 67, No. 5, pp. 1453–1464

HETEROCLINIC BIFURCATION IN THE MICHAELIS–MENTEN-TYPE RATIO-DEPENDENT PREDATOR-PREY SYSTEM*

BINGTUAN LI† AND YANG KUANG‡

Abstract. The existence of a heteroclinic bifurcation for the Michaelis–Menten-type ratio-dependent predator-prey system is rigorously established. Limit cycles related to the heteroclinic bifurcation are also discussed. It is shown that the heteroclinic bifurcation is characterized by the collision of a stable limit cycle with the origin, and the bifurcation triggers a catastrophic shift from the state of large oscillations of predator and prey populations to the state of extinction of both populations. It is also shown that the limit cycles related to the heteroclinic bifurcation originally bifurcate from the Hopf bifurcation.

Key words. ratio-dependent predator-prey model, heteroclinic cycle, bifurcation

AMS subject classifications. 34C05, 34D20, 92D25

DOI. 10.1137/060662460

1. Introduction.

In studying the interaction between predators and their prey, it is crucial to determine what specific form of the functional response that describes the amount of prey consumed per predator per unit of time is biologically plausible and provides a sound basis for theoretical development. Traditionally, dependence on prey density has been the starting point, giving a functional response function of the form $p(x)$. In the simplest case, such a function is a linear function of x, which is incorporated into the classical Lotka–Volterra predator-prey model. The linear functional response is a limiting case of the more general and useful Michaelis–Menten or Holling type II response function of the form $p(x) = \frac{cx}{m+x}$. Because $p(x)$ varies solely with prey density, it is usually labeled as "prey-dependence."

Sole dependence of the functional response on prey density has been questioned by several biologists (e.g., DeAngelis, Goldstein, and O'Neill [10], Arditi and Ginzburg [4], Arditi, Ginzburg, and Akcakaya [5], Akcakaya [1], Gutierrez [12]). It has been recognized that predators might interfere with each other's foraging, requiring the functional response to depend on densities of both predators and prey (DeAngelis, Goldstein, and O'Neill [10], Arditi and Akcakaya [2], Beddington [6]). Arditi and Ginzburg [4] have argued that a functional response depending on the ratio of prey to predator abundance is a suitable representation of some of these phenomena. With the Michaelis–Menten or Holling type II-type ratio-dependence functional response $p(x/y)$ and logistic prey growth, the predator-prey system takes the form of

$$\begin{aligned} x'(t) &= rx\left(1-\frac{x}{K}\right)-\frac{cxy}{x+my}, \\ y'(t) &= y\left(\frac{fx}{x+my}-d\right), \end{aligned} \tag{1.1}$$

*Received by the editors June 8, 2006; accepted for publication (in revised form) April 16, 2007; published electronically July 20, 2007.
http://www.siam.org/journals/siap/67-5/66246.html

†Department of Mathematics, University of Louisville, Louisville, KY 40292 (bing.li@louisville.edu). This author's research was partially supported by NSF grant DMS-0211614.

‡Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287-1804 (kuang@asu.edu). This author's research was partially supported by NSF grants DMS-0077790 and DMS/NIGMS-0342388.

where $x(t)$, $y(t)$ represent population densities of prey and predator, respectively, and r, K, c, m, f, d are positive constants that stand for prey intrinsic growth rate, carrying capacity, capturing rate, half saturation constant, maximal predator growth rate, and predator mortality rate, respectively.

The ratio-dependent predator-prey system (1.1) exhibits original dynamic properties that have never been observed in the early prey-dependent predator-prey systems. Specifically, the ratio-dependent predator-prey system (1.1) does not produce the so-called paradox of enrichment (Hairston, Smith, and Slobodkin [13], Rosenzweig [19]) or the paradox of biological control (Arditi and Berryman [3]). It also allows the predator population or both populations to either become extinct or coexist, depending on the initial population values. These are realistic features of predator-prey models that have been observed experimentally (Huffaker [14], Luckinbill [17]).

The dynamics of the ratio-dependent predator-prey system (1.1) has been systematically studied by Kuang and Beretta [16], Hsu, Hwang, and Kuang [15], Berezovskaya, Karev, and Arditi [7], and Xiao and Ruan [22]. These authors have shown that system (1.1) has very rich dynamics. In particular, the origin is a complicated equilibrium point whose characteristics determine some important properties of the system (see [7, 22]), the limit cycle exists and is unique and stable (see [15]), and the heteroclinic bifurcation plays an important role in understanding the dynamics of the system (see [7, 15]). Berezovskaya, Karev, and Arditi [7] have found numerically the heteroclinic cycle in (1.1) that corresponds to the disappearance of the limit cycle. It is thus interesting to rigorously establish the existence of heteroclinic bifurcation and to study the properties associated with the bifurcation. In a recent paper [20], Tang and Zhang reduced the system to a perturbed Hamiltonian system with a Delta-shape heteroclinic loop and computed Melnikov's function by eliminating some complicated terms in establishing the heteroclinic bifurcation. This is a valid and novel approach, yet its implementation is subtle since it involves intensive steps of variable manipulations and computations. The analysis presented in [20] contains a flaw that failed to ensure a proper application of Melnikov's method.

The objective of this paper is to rigorously establish the existence of heteroclinic bifurcation and determine the associated dynamics in system (1.1). This paper is organized as follows. The main results of the paper are provided in section 2. In this section, we use Melnikov's method to determine the existence of heteroclinic bifurcation. It is shown that the heteroclinic bifurcation is characterized by the collision of a stable limit cycle with the origin, and the bifurcation triggers a catastrophic shift from the state of large oscillations of predator and prey populations to the state of extinction of both populations. We also employ Melnikov's method to study limit cycles related to the heteroclinic bifurcation. It is shown that the limit cycles related the heteroclinic bifurcation originally bifurcate from the Hopf bifurcation. The biological interpretations of the theoretical results are also provided. Some concluding remarks are given in section 3.

2. Bifurcations.

2.1. Heteroclinic bifurcation. For simplicity, we nondimensionalize system (1.1) as in Tang and Zhang [20] with the following scaling:

$$x \to Kx, \quad y \to Ky/m, \quad t \to mt/c.$$

(Throughout this paper the variable on the left-hand side of $\to$ always represents the old variable.) With this scaling, system (1.1) takes the form

$$\begin{aligned} x'(t) &= \alpha x(1-x) - \frac{xy}{x+y}, \\ y'(t) &= -\beta y + \frac{\kappa xy}{x+y}, \end{aligned} \tag{2.1}$$

where

$$\alpha = \frac{rm}{c}, \beta = \frac{dm}{c}, \kappa = \frac{fm}{c}. \tag{2.2}$$

As shown in [7, 20, 22], system (2.1) in the first quadrant is equivalent to the polynomial system

$$\begin{aligned} x'(t) &= \alpha x(1-x)(x+y) - xy \\ y'(t) &= -\beta y(x+y) + \kappa xy \end{aligned} \tag{2.3}$$

obtained from (2.1) by a change of the independent variable

$$t \to (x+y)t.$$

As in [7, 20, 22], one can then use Briot–Bouquet's transformation

$$x \to x, \quad y \to yx, \quad t \to t/x \tag{2.4}$$

to convert (2.3) to

$$\begin{aligned} x'(t) &= x[\alpha - \alpha x - (1-\alpha)y - \alpha xy], \\ y'(t) &= y[(\kappa - \alpha - \beta) + \alpha x + (1-\alpha-\beta)y + \alpha xy]. \end{aligned} \tag{2.5}$$

Transformation (2.4) is a homomorphism in the first quadrant, and its inverse maps the y axis to the point $(0,0)$.

Tang and Zhang [20] used variable changes to transform (2.5) to

$$\begin{aligned} v_1'(t) &= v_1\left[\mu_1 + v_1^2 + \frac{1-\alpha}{1-\alpha-\beta}v_2^2\right] + \delta v_1\left(\mu_2 + \frac{1}{1-\alpha-\beta}v_1^2v_2^2\right), \\ v_2'(t) &= v_2\left[\frac{-2(1-\alpha-\beta)}{2-2\alpha-\beta}\mu_1 - v_1^2 - v_2^2\right] + \delta v_2\left(\mu_2 - \frac{1}{1-\alpha-\beta}v_1^2v_2^2\right), \end{aligned} \tag{2.6}$$

where δ, μ_1, and μ_2 are related to α and $\alpha+\beta-\kappa$, and in particular $\alpha = -(\delta\mu_1 + \delta^2\mu_2)$. In [20] the coefficient term $(1-\alpha)/(1-\alpha-\beta)$ is treated as a constant when Melnikov's method is used to carry out bifurcation analysis with respect to parameters δ, μ_1, μ_2. This is not appropriate. One needs to split this coefficient term into a term independent of bifurcation parameters and a perturbation term in applying Melnikov's method.

Instead of working on (2.6), we study the following simpler system

$$\begin{aligned} x'(t) &= x[\alpha - x - (1-\alpha)y - xy], \\ y'(t) &= y[(\kappa - \alpha - \beta) + x + (1-\alpha-\beta)y + xy], \end{aligned} \tag{2.7}$$

obtained from (2.5) by the change of variable

$$x \to x/\alpha.$$

We simply use α and $\nu = \kappa - \alpha - \beta$ (or equivalently α and κ) as our unfolding parameters while fixing β. In (2.7) there are two second order terms whose coefficients depend on α. We decompose these terms and rewrite (2.7) as

$$\begin{aligned} x'(t) &= x(\alpha - x - y) + x(\alpha y - xy), \\ y'(t) &= y(\nu + x + (1-\beta)y) + y(-\alpha y + xy). \end{aligned} \tag{2.8}$$

This system can then be viewed as a perturbation of the system

$$\begin{aligned} x'(t) &= x(\alpha - x - y), \\ y'(t) &= y(\nu + x + (1-\beta)y), \end{aligned} \tag{2.9}$$

as α, ν, x, and y are all small. Note that the coefficients of second order terms in (2.9) do not depend on α and ν.

We shall assume that

$$\beta < 1.$$

System (2.9) is integrable if

$$\nu = -\frac{2(1-\beta)}{2-\beta}\alpha < 0, \tag{2.10}$$

and in this case the function

$$F_\alpha(x,y) = \frac{1}{b}x^a y^b \left(\alpha - x - \frac{2-\beta}{2}y\right), \tag{2.11}$$

where

$$a = 2\frac{1-\beta}{\beta}, \qquad b = \frac{2-\beta}{\beta}, \tag{2.12}$$

is constant along solution curves. In fact, when (2.10) holds, along any solution curve $(x(t), y(t))$ of (2.9), $\frac{dF_\alpha(x,y)}{dt} = \frac{\partial F_\alpha}{\partial x}x'(t) + \frac{\partial F_\alpha}{\partial y}y'(t) = \frac{1}{b}[\alpha a x^{a-1}y^b - (a+1)x^a y^b - \frac{2-\beta}{2}a x^{a-1}y^{b+1}]x(\alpha - x - y) + \frac{1}{b}[\alpha b x^a y^{b-1} - b x^{a+1}y^{b-1} - \frac{2-\beta}{2}(b+1)x^a y^b]y(\nu + x + (1-\beta)y) = 0$. The level curves of F_α take the form shown in Figure 1. Here we have a family of periodic orbits encircling the center at $(\bar{x}, \bar{y}) = (\frac{(1-\beta)\alpha}{2-\beta}, \frac{\alpha}{2-\beta})$ and limiting on the heteroclinic cycle $F_\alpha(x,y) = 0$, which is a triangle connecting the saddles at $(0,0)$, $(\alpha, 0)$, and $(0, \frac{2\alpha}{2-\beta})$.

Using the transformations

$$x \to \epsilon x, \quad y \to \epsilon y, \quad \alpha = \epsilon\nu_1, \quad \nu = -\frac{2(1-\beta)}{2-\beta}\epsilon\nu_1 + \nu_2\epsilon^2$$

and rescaling time $t \to t/\epsilon$, we convert system (2.8) into

$$\begin{aligned} x'(t) &= x[\nu_1 - x - y] + \epsilon(\nu_1 xy - x^2 y), \\ y'(t) &= y\left[-\frac{2(1-\beta)}{2-\beta}\nu_1 + x + (1-\beta)y\right] + \epsilon(\nu_2 y - \nu_1 y^2 + xy^2). \end{aligned} \tag{2.13}$$

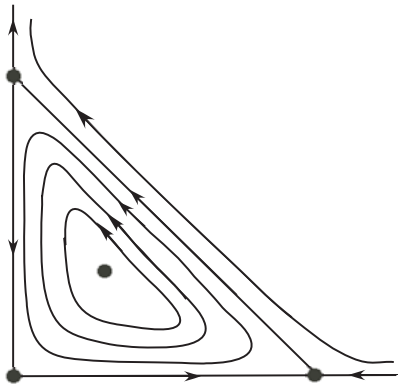


FIG. 1. *The level curves of* $F_\alpha(x,y)$.

Multiplying (2.13) by the integrating factor $x^{a-1}y^{b-1}$, we obtain the "equivalent" perturbed Hamiltonian system:

$$\begin{aligned} x'(t) &= x^a y^{b-1}\{[\nu_1 - x - y] + \epsilon(\nu_1 y - xy)\} \\ y'(t) &= x^{a-1}y^b \left\{ \left[-\frac{2(1-\beta)}{2-\beta}\nu_1 + x + (1-\beta)y \right] + \epsilon(\nu_2 - \nu_1 y + xy) \right\}. \end{aligned} \tag{2.14}$$

One can check that

$$F_{\nu_1}(x,y) = \frac{1}{b}x^a y^b \left(\nu_1 - x - \frac{2-\beta}{2}y \right) \tag{2.15}$$

is the Hamiltonian function for (2.14) when $\epsilon = 0$, where a and b are given in (2.12).

We use the Melnikov theory [9, 11, 21] to locate parameter values that produce a heteroclinic cycle for (2.14) in the case $\epsilon \neq 0$. The analysis that we perform here is similar to what is carried out in section 7.5 of [11] and section 4.7 of [9]. We can set $\nu_1 = 1$ without loss of generality. The heteroclinic cycle for $\epsilon = 0$ lies on the level curve $F_1(x,y) = 0$, denoted by Γ_0, which corresponds to a triangle formed by the three line segments determined by $x = 0$, $y = 0$, and $x + \frac{2-\beta}{2}y = 1$. Let

$$\mathbf{G}(x,y) = (x^a y^{b-1}(y - xy), x^{a-1}y^b(\nu_2 - y + xy)).$$

The Melnikov function is

$$\begin{aligned} M(\nu_2) &= \int\int_{\text{int}\Gamma_0} \text{trace} D\mathbf{G}(x,y)dxdy \\ &= \int\int_{\text{int}\Gamma_0} [(a-b-1)x^{a-1}y^b + (b-a)x^a y^b + bx^{a-1}y^{b-1}\nu_2]dxdy, \end{aligned} \tag{2.16}$$

where $\text{int}\Gamma_0$ denotes the region bounded by Γ_0. $M(\nu_2) = 0$ has a unique solution

$$\nu_2 = -\frac{(a-b-1)I(a-1,b) + (b-a)I(a,b)}{bI(a-1,b-1)}, \tag{2.17}$$

where

$$I(u,v) = \int\int_{\text{int}\Gamma_0} x^u y^v dxdy, \qquad u > -1, v > -1.$$

It is easy to see that

$$I(u,v)=\int_0^1 x^u\int_0^{\frac{(1-x)}{s}} y^v dydx=\frac{1}{(v+1)s^{v+1}}\int_0^1 x^u(1-x)^{v+1}dx,$$

where $s=\frac{2-\beta}{2}$.

We have

$$\begin{aligned} I(u+1,v)&=\frac{1}{(v+1)s^{v+1}}\int_0^1 x^{u+1}(1-x)^{v+1}dx\\ &=\frac{1}{(v+1)s^{v+1}}\int_0^1 x^u(1-x)^{v+1}(x-1+1)dx\\ &=-\frac{1}{(v+1)s^{v+1}}\int_0^1 x^u(1-x)^{v+2}dx+I(u,v)\\ &=-\frac{v+2}{v+1}sI(u,v+1)+I(u,v). \end{aligned} \tag{2.18}$$

Using integration by parts, we obtain

$$\begin{aligned} I(u,v+1)&=\frac{1}{(v+2)s^{v+2}}\int_0^1 x^u(1-x)^{v+2}dx\\ &=\frac{1}{(v+2)s^{v+2}}\left[\frac{x^{u+1}(1-x)^{v+2}}{u+1}\Big|_0^1-\int_0^1\frac{x^{u+1}}{u+1}(-1)(v+2)(1-x)^{v+1}dx\right]\\ &=\frac{v+1}{(u+1)s}I(u+1,v). \end{aligned} \tag{2.19}$$

Using (2.18) and (2.19), we find

$$I(u+1,v)=\frac{u+1}{u+v+3}I(u,v),\qquad I(u,v+1)=\frac{v+1}{(u+v+3)s}I(u,v). \tag{2.20}$$

It follows from (2.12), (2.17), and (2.20) that

$$\begin{aligned} \nu_2&=-\left[\frac{a-b-1}{b}\frac{I(a-1,b)}{I(a-1,b-1)}+\frac{b-a}{b}\frac{I(a,b)}{I(a,b-1)}\frac{I(a,b-1)}{I(a-1,b-1)}\right]\\ &=-\left[\frac{a-b-1}{(a+b+1)s}+\frac{a(b-a)}{(a+b+1)(a+b+2)s}\right]\\ &=\frac{6\beta}{(4-\beta)(2-\beta)^2}. \end{aligned}$$

The Melnikov theory [11] shows that if

$$\nu=-\frac{2(1-\beta)}{2-\beta}\alpha+\frac{6\beta}{(4-\beta)(2-\beta)^2}\alpha^2+O(\alpha^3), \tag{2.21}$$

then system (2.8) has a heteroclinic cycle. It is shown in [7] that the heteroclinic cycle is stable.

Condition (2.21) is equivalent to

$$\kappa=\beta+\frac{\beta}{2-\beta}\alpha+\frac{6\beta}{(4-\beta)(2-\beta)^2}\alpha^2+O(\alpha^3). \tag{2.22}$$

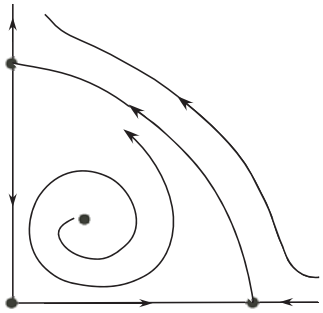


FIG. 2. *Dynamics of system* (2.7) *if the conditions of Lemma* 1 *hold.*

We have obtained the following result.

LEMMA 1. *Assume that* β *is fixed and* $\beta < 1$. *For small* α, *if condition* (2.22) *holds, then system* (2.7) *has a stable heteroclinic cycle connecting saddles at* $(0,0)$, $(\alpha,0)$, *and* $(0, \frac{-(\kappa-\alpha-\beta)}{1-\alpha-\beta})$.

The dynamics of the system in the case that there exists a stable heteroclinic cycle was discussed by Tang and Zhang [20]. For the sake of completeness, we describe the dynamics in this case based on our analytical results. The positive coexistence equilibrium $(\frac{\beta+\alpha\kappa-\kappa}{\kappa}, \frac{\kappa-\beta}{\beta})$ lies inside the heteroclinic cycle. One can check that it is a spiral source. Conditions of Lemma 1 show that for small α

$$\beta < \kappa < \alpha + \beta < 1, \qquad \beta + \alpha\kappa - \kappa > 0, \tag{2.23}$$

which implies that the condition (2.9) in Hsu, Hwang, and Kuang [15] holds. Theorem 2.7 in [15] shows that in this case a limit cycle in the system is always stable and unique once it exists. We therefore conclude that there is no limit cycle inside the heteroclinic cycle since it is attracting. The dynamics of the system in this case is depicted in Figure 2.

Lemma 1 and the transformations used to convert (2.1) to (2.7) imply the following result.

THEOREM 1. *Assume that* β *is fixed and* $\beta < 1$. *If for small* α *condition* (2.22) *holds, then system* (2.1) *has a stable heteroclinic cycle connecting saddles at* $(0,0)$ *and* $(1,0)$.

The conditions of Theorem 1 imply that

$$c - rm - dm > 0, \quad f - r - d < 0, \quad d < f < \frac{cd}{c - rm} \tag{2.24}$$

in the original system (1.1). In view of the properties of the heteroclinic cycle described in Lemma 1, and after Theorem 2.3 and Theorem 2.5 of Xiao and Ruan [22], we see that the heteroclinic cycle in Theorem 1 approaches the origin in the characteristic direction $\theta = \arctan((\kappa - \alpha - \beta)/(\alpha + \beta - 1))$, and the topological structure of the origin consists of a hyperbolic sector and a parabolic sector; see Figure 3.

2.2. Limit cycles near the heteroclinic bifurcation. We have used Melnikov's method to show the persistence of the heteroclinic cycle of the integrable system (2.13) when $\epsilon = 0$ under the perturbation described by (2.22). This method can also be used to study the survival of each periodic cycle in (2.13) under an appropriate perturbation (see section 7.5 of Guckenheimer and Holmes [11] and section 4.7

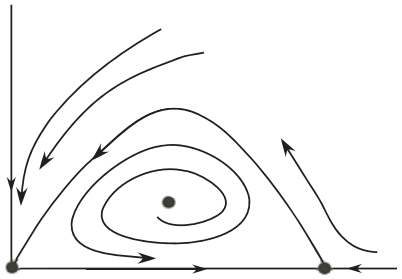


FIG. 3. *Dynamics of system* (2.1) *when a stable heteroclinic cycle exists.*

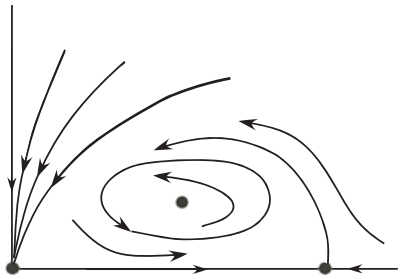


FIG. 4. *Dynamics of system* (2.1) *when* $0 < \nu_2 < \frac{6\beta}{(4-\beta)(2-\beta)^2}$, $\beta < 1$, *and* κ *is given by* (2.25).

of Chow, Li, and Wang [9]). Let Γ_γ be a periodic cycle in Figure 1 that represents the level curves of F_1; then this cycle survives in system (2.1) if

$$\kappa = \beta + \frac{\beta}{2-\beta}\alpha + \nu_2\alpha^2 + O(\alpha^3) \tag{2.25}$$

with

$$\nu_2 = -\frac{\int\int_{\text{int}\Gamma_\gamma} [(a-b-1)x^{a-1}y^b + (b-a)x^a y^b]dxdy}{\int\int_{\text{int}\Gamma_\gamma} bx^{a-1}y^{b-1}dxdy}, \tag{2.26}$$

where $\text{int}\Gamma_\gamma$ denotes the region bounded by Γ_γ. In (2.26), ν_2 represents the solution of $M(\nu_2) = 0$, where $M(\nu_2)$ is given by (2.16) with Γ_0 replaced by Γ_γ. We first study small cycles near the equilibrium. As Γ_γ shrinks to the equilibrium $(\bar{x}, \bar{y}) = (\frac{(1-\beta)\alpha}{2-\beta}, \frac{\alpha}{2-\beta})$, the right-hand side of (2.26) approaches $-\frac{(a-b-1)\bar{x}^{a-1}\bar{y}^b + (b-a)\bar{x}^a\bar{y}^b}{b\bar{x}^{a-1}\bar{y}^{b-1}} = 0 + O(\alpha)$, which shows $\nu_2 = 0$ in (2.25). In this case, standard Hopf bifurcation analysis shows that a supercritical Hopf bifurcation occurs in (2.1). Since (2.25) implies the condition (2.9) in [15], Theorem 2.7 in [15] shows that system (2.1) has at most one limit cycle. Due to this fact and continuity, as Γ_γ moves from a circle near the equilibrium to a circle near the heteroclinic cycle, ν_2 increases from a number near 0 to a number near $\frac{6\beta}{(4-\beta)(2-\beta)^2}$. This shows that as ν_2 increases from 0 to $\frac{6\beta}{(4-\beta)(2-\beta)^2}$, system (2.1) has a unique limit cycle whose size increases from 0 to the size of the heteroclinic cycle.

One can easily check that condition (2.25) and the assumption $\beta < 1$ imply (2.24). Using Theorem 2.3 and Theorem 2.5 of Xiao and Ruan [22] and the properties associated with the heteroclinic cycle discussed above, we depict the dynamics of (2.1) before and after the heteroclinic bifurcation occurs as in Figures 4 and 5.

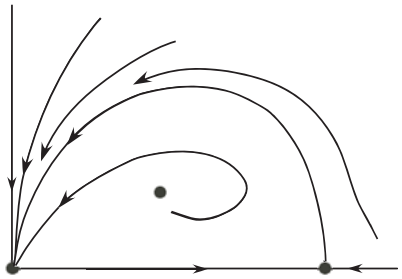


FIG. 5. *Dynamics of system* (2.1) *when* $\nu_2 > \frac{6\beta}{(4-\beta)(2-\beta)^2}$, $\beta < 1$, *and* κ *is given by* (2.25).

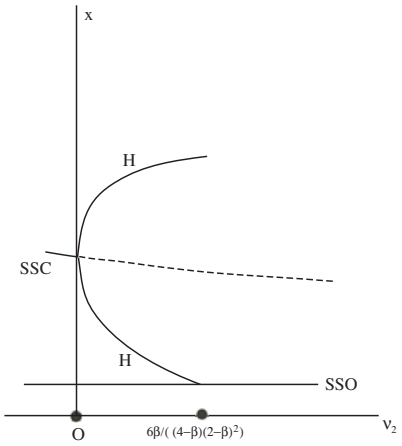
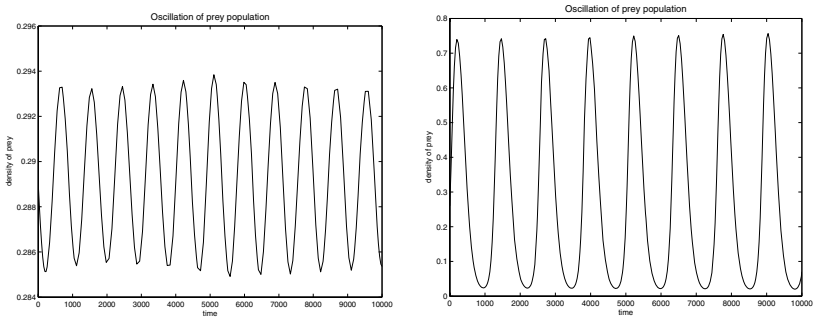


FIG. 6. *Bifurcation diagram of the prey population.*

2.3. Interpretation of the theoretical results. Based on the above results, the bifurcation diagram of the prey population x of system (2.1) is depicted in Figure 6. (The bifurcation diagram of the predator population y is similar.) 0 is the first critical value for ν_2 at which the Hopf bifurcation occurs. If ν_2 is slightly less than 0, both prey and predator populations either tend to the origin (solid line labeled SS0), becoming extinct eventually, or tend toward a stable coexistence equilibrium (solid line labeled SSC), depending on the initial values. If ν_2 is greater than 0, both prey and predator populations either tend toward the origin or tend to a limit cycle (solid line labeled H). In this case, the coexistence equilibrium (dashed line) is unstable. The Hopf bifurcation marks a critical condition at which the coexistence equilibrium becomes unstable and the prey and predator populations near the equilibrium starts oscillating periodically. If ν_2 increases from 0 to $\frac{6\beta}{(4-\beta)(2-\beta)^2}$, the amplitude of the oscillating population becomes larger. $\frac{6\beta}{(4-\beta)(2-\beta)^2}$ is the second critical value for ν_2 at which heteroclinic bifurcation occurs. It represents the collision of a large stable limit cycle with the origin. If ν_2 is slightly greater than $\frac{6\beta}{(4-\beta)(2-\beta)^2}$, the limit cycle attractor does not exist anymore, and both prey and predator populations become "unconditionally extinct"; i.e., the origin attracts all solutions (except the coexistence equilibrium solution). Thus starting in the oscillating state, a small increase in ν_2 may lead to a shift to the attractor—the origin. This heteroclinic bifurcation leading to the collapse of large oscillations of populations and extinction of populations is usually



(a) (b)

FIG. 7. *Small and large periodic oscillations of prey population.* (a) $\nu_2 = 0.2$, (b) $\nu_2 = 0.54$.

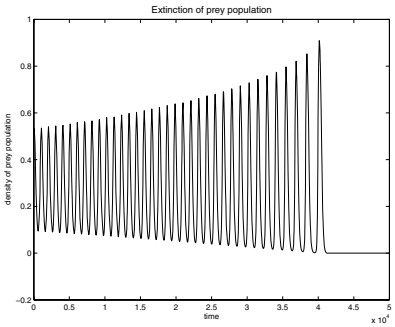


FIG. 8. *Extinction of prey population accompanied with oscillations when* $\nu_2 = 0.56$.

called a "catastrophic bifurcation" (Rinaldi and Scheffer [18]).

We carry out numerical simulations to demonstrate the dynamics of (2.1) described in the bifurcation diagram of Figure 6. We choose $\beta = 0.6$ and $\alpha = 0.02$ (a relatively small number). In this case $\frac{6\beta}{(4-\beta)(2-\beta)^2} = 0.5402$. Figure 7 shows periodic oscillations of the prey population when $\nu_2 = 0.2$ and $\nu_2 = 0.54$. Note that the amplitude of the periodic solution for $\nu_2 = 0.54$ is large, whereas the amplitude of the periodic solution for $\nu_2 = 0.2$ is very small. Figure 8 shows extinction of the population accompanied with oscillations when $\nu_2 = 0.56$, which is slightly greater than the critical value near 0.5402

Recall that α, β, and κ in terms of the parameters in the original system (1.1) are given by

$$\alpha = \frac{rm}{c}, \quad \beta = \frac{dm}{c}, \quad \kappa = \frac{fm}{c}.$$

Our assumption $\beta < 1$ implies $dm < c$. This shows that the capturing rate is relatively large. We fix $\frac{m}{c}$ so that if the prey intrinsic growth rate r is small, then α is small. We rewrite (2.25) in the form that the maximal predator growth rate f is a function of other parameters in (1.1) for small r. The resulting expression can then be used to study the dynamics of (1.1) including the Hopf bifurcation and heteroclinic bifurcation by varying f, based on the dynamics of the equivalent system (2.1). One can see that for relatively large c and f both predator and prey populations become extinct.

3. Concluding remarks. We have rigorously established the existence of a heteroclinic bifurcation and studied the related dynamics for the Michaelis–Menten-type ratio-dependent system by using Melnikov's method. This method is often used to obtain "small" heteroclinic cycles or limit cycles in studying local bifurcations (see Guckenheimer and Holmes [11], Wiggins [21], and Chow, Li, and Wang [9]). The heteroclinic cycle described in Lemma 1 is such a small heteroclinic cycle near the origin. However, the heteroclinic cycle described in Theorem 1 that connects the origin and the equilibrium $(1, 0)$ is not a small one. This is essentially due to the variable change $x \to x/\alpha$ for small α that we have used, which together with other variable changes converts the former heteroclinic cycle into the latter one.

Our results show that near the Hopf bifurcation, depending on the initial conditions, populations of predators and prey either coexist or become extinct. These features have not been described by early prey-dependent predator-prey systems. The heteroclinic bifurcation triggers a shift from the state of periodic coexistence of populations to the state of extinction of both populations, resulting in a "catastrophe" to the predator-prey system.

Rinaldi and Scheffer [18] gave many interesting bifurcation examples in ecological models. They pointed out that a heteroclinic bifurcation is due to the collision of a stable limit cycle and a unstable saddle equilibrium. This is similar to what happens in system (1.1). However, the origin, an unstable point involved in the heteroclinic bifurcation for (1.1), is always an attractor. It is very interesting to note that the heteroclinic bifurcation results in the global attractivity of the origin.

Acknowledgments. We would like to thank the two referees for helpful suggestions for improving the paper. We also want to thank Professor Weinian Zhang for valuable discussions.

REFERENCES

[1] H. R. Akcakaya, *Population cycles of mammals: Evidence for a ratio-dependent predation hypothesis*, Ecol. Monogr., 62 (1992), pp. 119–142.
[2] R. Arditi and H. R. Akcakaya, *Underestimation of mutual interference of predators*, Oecologia (Berlin), 83 (1990), pp. 358–361.
[3] R. Arditi and A. A. Berryman, *The biological control paradox*, Trends Ecol. Evolution, 6 (1991), p. 32.
[4] R. Arditi and L. R. Ginzburg, *Coupling in predator-prey dynamics: Ratio-dependence*, J. Theoret. Biol., 139 (1989), pp. 311–326.
[5] R. Arditi, L. R. Ginzburg, and H. R. Akcakaya, *Variation in plankton densities among lakes: A case of ratio-dependent models*, Amer. Naturalist, 138 (1991), pp. 1287–1296.
[6] J. R. Beddington, *Mutual interference between parasites or predators and its effect on searching efficiency*, J. Animal Ecol., 44 (1975), pp. 331–340.
[7] F. Berezovskaya, G. Karev, and R. Arditi, *Parametric analysis of the ratio-dependent predator-prey model*, J. Math. Biol., 43 (2001), pp. 221–246.
[8] A. A. Berryman, *The origins and evolution of predator-prey theory*, Ecol., 73 (1992), pp. 1530–1535.
[9] S.-N. Chow, C. Li, and D. Wang, *Normal Forms and Bifurcation of Planar Vector Fields*, Cambridge University Press, New York, 1994.
[10] D. L. DeAngelis, R. A. Goldstein, and R. V. O'Neill, *A model for trophic interactions*, Ecol., 56 (1975), pp. 881–892.
[11] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.
[12] A. P. Gutierrez, *The physiological basis of ratio-dependent predator-prey theory: A metabolic pool model of Nicholson's blowflies as an example*, Ecol., 73 (1992), pp. 1552–1563.
[13] N. G. Hairston, F. E. Smith, and L. B. Slobodkin, *Community structure, population control and competition*, Amer. Naturalist, 94 (1960), pp. 421–425.

[14] C. B. Huffaker, *Experimental studies on predation: Dispersion factors and predator-prey oscillations*, Hilgardia, 27 (1958), pp. 343–383.
[15] S.-B. Hsu, T.-W. Hwang, and Y. Kuang, *Global analysis of the Michaelis–Menten type ratio-dependent predator-prey system*, J. Math. Biol., 42 (2001), pp. 489–506.
[16] Y. Kuang and E. Beretta, *Global qualitative analysis of a ratio-dependent predator-prey system*, J. Math. Biol., 36 (1998), pp. 389–406.
[17] L. S. Luckinbill, *Coexistence in laboratory populations of Paramecium aurelia and its predator Didinium nasutum*, Ecol., 54 (1973), pp. 1320–1327.
[18] S. Rinaldi and M. Scheffer, *Geometric analysis of ecological models with slow and fast processes*, Ecosystems, 3 (2000), pp. 507–521.
[19] M. L. Rosenzweig, *Paradox of enrichment: Destabilization of exploitation systems in ecological time*, Science, 171 (1969), pp. 385–387.
[20] Y. Tang and W. Zhang, *Heteroclinic bifurcation in a ratio-dependent predator-prey system*, J. Math. Biol., 50 (2005), pp. 699–712.
[21] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, New York, 2003.
[22] D. Xiao and S. Ruan, *Global dynamics of a ratio-dependent predator-prey system*, J. Math. Biol., 43 (2001), pp. 268–290.

SIAM J. APPL. MATH.
Vol. 67, No. 5, pp. 1465–1478

LIFT ON SLENDER BODIES WITH ELLIPTICAL CROSS SECTION EVALUATED BY USING AN OSEEN FLOW MODEL*

EDMUND CHADWICK† AND NINA FISHWICK†

Abstract. Consider uniform, incompressible flow past a slender body with an elliptical cross section such that the major axis of the body is inclined slightly to the flow direction. Assume that the flow is inviscid everywhere except in a thin boundary layer region and in the vortex core of trailing line vortices that emanate from the body into the vortex wake. Hence, the flow is quasi-inviscid, and so the slip (impermeability) boundary condition is applied. Further assume that outside the boundary layer the velocity is to first order the uniform stream velocity. Then the Oseen approximation can be applied. The resulting solution, up to the slender body approximation, is given, and the lift over the slender body is determined. This solution is then compared with the theoretical and experimental results for flow past a delta wing, the viscous cross-flow method and experimental results for flow past a body with a circular cross section, and Newtonian impact theory and experimental results for flow past a body with an elliptical cross section.

Key words. oseen flow, slender body theory

AMS subject classifications. 76D07, 76D17, 76D09

DOI. 10.1137/060663726

1. Introduction. Slender body theories are used to provide efficient and accurate algorithms to important fluid maneuvering applications, in particular in missile guidance [24] and ship maneuverability [23]. The historical development began with Munk [22] to calculate the moment on an airship. Munk used the inviscid flow model, and for the lift on a slender delta wing Jones [15] demonstrated a high degree of agreement with experiment. However, for bodies that do not have trailing edges or wings, there are significant discrepancies in the lift calculation from the inviscid formulation. Newman [23] notes an almost factor two increase in the experimental lift over the theoretical result for the lift on a slender ship.

Allen and Perkins [1] modify the inviscid result for missiles with a circular cross section by including a viscous cross-flow contribution. The lift is obtained by considering the cross-flow drag along the body, and the results agree well with experiment. However, the moment calculation is not as accurate. Fishwick [10] has determined the center of moment for the viscous lift for a comprehensive set of National Advisory Committee for Aeronautics (NACA) experiments on various slender missile shapes with Reynolds number $Re = 10^3 - 10^6$, aspect ratio, and slenderness parameter of order $O(10^{-1})$. All of the subsequent experimental results referred to in this paper are in this range. For flows less than or equal to 4°, Fishwick demonstrates that the viscous force acts at the end section. This result has also been observed by Clarke [9], who divides a variety of ship hulls into sections along the length and then determines the forces on each section. The results show a large viscous lift (on top of the inviscid lift from the pressure distribution) at the end section only. In contrast, the Allen and Perkins method assumes a viscous cross-flow at each section along the body length, and consequently the viscous force acts close to the midsection. Furthermore, the

*Received by the editors June 26, 2006; accepted for publication (in revised form) June 11, 2007; published electronically August 22, 2007.
http://www.siam.org/journals/siap/67-5/66372.html
†School of Computing, Science, and Engineering, University of Salford, Salford M5 4WT, UK (e.a.chadwick@salford.ac.uk, njfishwick@mail.dstl.gov.uk).

Allen and Perkins method does not work as well for bodies not having a circular cross section, and modifications to the general theory are required.

Jorgensen modifies the viscous cross-flow method to overcome this deficiency by using Newtonian impact theory [16]. For the case of a body with an elliptical cross section, as the semiminor axis of the body is reduced the lift also reduces relative to a body with a circular cross section. Jorgensen's result gives good first order agreement with experiment, and Jorgensen argues that the reduction in lift determined by this theory is justified from experimental observations relating to the change in the position and therefore the lifting effect of trailing line vortices over the slender body.

Trailing vortices emanating from the leading edge of a delta wing are known to generate uplift at a high angle of attack [27] but not at small angles of attack. However, for slender ships the vortex strength and position of the trailing line vortices along the length of the body have been determined experimentally, and these results were fed into an inviscid flow model [13]. This approach gives good agreement with experiment, but the calculation of the vortex strength and position has been determined experimentally rather than from the theoretical model.

Chadwick [6] considers a slender body in Oseen flow. The theory is applicable to large Reynolds number flows (but aerodynamically low-speed) of around $Re = kL = 10^3 - 10^6$, where $k = \rho U/2\mu$ and $d = L\delta$. (Here, the free stream velocity is given by U; ρ and μ are the fluid density and the dynamical coefficient of viscosity, respectively; a typical cross-sectional length is given by d; L is the body length; and δ is the slenderness parameter.) So near the slender body, viscous forces are negligible, and the flow is very nearly inviscid up to a small thickness boundary layer. This means that the slip/impermeability boundary condition can be applied to the inviscid velocity potential. However, in the far-field downstream wake, viscous forces cannot be neglected. Batchelor [2] states that the effect of viscosity, although small, gives the appropriate leading order solution. To obtain the far-field representation of the trailing vortex, Batchelor [3] retains the viscous component and linearizes about the uniform stream, yielding the Oseen equations. Chadwick [8] determines the line vortex in Oseen flow and demonstrates the importance of the viscous term in the calculation of the lift force. Outside the wake, the velocity is given by the Oseen velocity potential. The near-field inviscid potential and far-field Oseen potential are matched. (It is further noted that the near-field inviscid flow region assumes a small perturbation flow [18] such that the perturbed velocity is much smaller than the uniform stream velocity. This is the Oseen approximation, and so to this level of approximation the near-field inviscid potential is identically the Oseen potential.) In the matching a coupled viscous term arises in the Oseen flow field which provides an additional viscous force. In this way, the additional viscous force contribution of Allen and Perkins is obtained without the requirement of employing a semiempirical procedure. For flow past a slender body with a circular cross section, Chadwick [6] demonstrates that the additional viscous lifting force is equal to, and on top of, the inviscid lifting force from the pressure. This agrees with the experimental findings of Newman [23], Clarke [9], and Allen and Perkins [1]. Applied to the problem of flow past a slender delta wing [7], the theory gives a lifting force the same as that given by the inviscid flow theory and experiment detailed by Jones [15].

In the present paper, we shall apply the slender body theory in Oseen flow presented in [6] for the case of a slender body with an elliptical cross section and in particular derive the formula that determines the lift. This is obtained by representing the body by a far-field distribution of Oseen lifting elements over an area bounded by the focii of each elliptical cross section. The potential velocity part of this distribu-

tion in turn approximates to a near-field inviscid flow distribution of normal dipoles. The two velocity potentials can be matched, and by using elliptical coordinates the slip (impermeability) boundary condition is satisfied. This theory then enables us to determine the lift for a delta wing which is compared with Jones's theory and experimental results, bodies of a circular cross section which is compared with Allen's and Perkins's viscous cross-flow method and experimental results, and bodies of an elliptical cross section which is compared with Jorgensen's Newtonian impact theory and experimental results. The results of the comparisons are then discussed.

2. Statement of the problem. We start with the Navier–Stokes equations [20, p. 577]

$$\rho(\mathbf{u}^\dagger.\nabla)\mathbf{u}^\dagger = -\nabla p^\dagger + \mu\nabla^2\mathbf{u}^\dagger\ , \quad \nabla.\mathbf{u}^\dagger = 0. \tag{2.1}$$

$\mathbf{u}^\dagger$ and $p^\dagger$ are the Navier–Stokes velocity and pressure, respectively. ρ and μ are the fluid density and the dynamical coefficient of viscosity, respectively, and are both assumed to be constant. ∇ denotes the gradient operator and ∇^2 the Laplacian operator.

The Navier–Stokes equations are linearized to a uniform stream U by assuming that

$$\mathbf{u}^\dagger = U\hat{\mathbf{x_1}} + \mathbf{u} + O(U\delta_{Oseen}^2)\ , \quad p^\dagger = p_\infty + p + O(\rho U^2\delta_{Oseen}^2), \tag{2.2}$$

where "O" means "of the order of." The ratio of the perturbed velocity to the uniform stream velocity is given by δ_{Oseen} and is much less than 1: $|\mathbf{u}/U| = O(\delta_{Oseen}) << 1$. $\hat{\mathbf{x_1}}$ is the unit vector in the x_1-direction for the Cartesian coordinates (x_1, x_2, x_3). p_∞ is the far-field pressure upstream from the body and so from Bernoulli's equation is of order $\mathrm{O}(\rho U^2)$. As the value is a constant, it is often taken to be zero. $\mathbf{u}$ and p are the Oseen velocity and pressure, respectively. In the linearization about the parameter δ_{Oseen}, $\mathbf{u}$ is of order $\mathrm{O}(U\delta_{Oseen})$, and p is of order $\mathrm{O}(\rho U^2\delta_{Oseen})$. We note that δ_{Oseen} is independent of the Reynolds number Re, so the linearization does not imply any restrictions on Re [2], [3], [4]. For the lifting problem at a small angle of attack, $\alpha \sim V/U$, where V is the uniform stream in the cross-flow direction. The order for the velocity $\mathbf{u}$ is given by the cross-flow order $\mathrm{O}(V)$, which is $\mathrm{O}(U\alpha)$. So, δ_{Oseen} is given by the angle of attack α, unless $\delta > \alpha$, where δ is the slenderness parameter. In this case, like the nonlifting problem, the order of the perturbed velocity is given by the outflow at a cross section, and so δ_{Oseen} is given by δ. The order of the error in the velocity and pressure using this approximation is then given from (2.2). (We note that this error is much larger than the error associated with that due to the boundary layer thickness for this Reynolds number range, slenderness ratio, and angle of attack range.) This yields the Oseen equations [25, pp. 30–38]

$$\rho U\frac{\partial\mathbf{u}}{\partial x_1} = -\nabla p + (\mu\nabla^2)\mathbf{u}\ , \quad \nabla.\mathbf{u} = 0, \tag{2.3}$$

$$\nabla^2 p = 0. \tag{2.4}$$

As $R = \sqrt{x_1^2 + x_2^2 + x_3^2} \to \infty$, then $\mathbf{u}$, $p \to 0$.

The force on the body is represented by a surface integral enclosing the body [5] such that

$$\mathbf{F}^\dagger = \int\int_S -p^\dagger\mathbf{n} + \mu(\mathbf{n}.\nabla)\mathbf{u}^\dagger - \rho\mathbf{u}^\dagger\mathbf{u}^\dagger.\mathbf{n}\mathrm{d}s. \tag{2.5}$$

$\mathbf{n} = (n_1, n_2, n_3)$ denotes the normal vector to the surface, and $\mathbf{F}^\dagger$ is the force integral in terms of the Navier–Stokes velocity and pressure. Therefore

$$\mathbf{F} = \int\int_S -p\mathbf{n} + \mu(\mathbf{n}.\nabla)\mathbf{u} - \rho U\mathbf{u}n_1 \mathrm{d}s, \tag{2.6}$$

where $\mathbf{F}$ is the force integral in terms of the Oseen velocity and pressure. The singular force solutions are then obtained by decomposing the fluid velocity into a potential velocity and a wake velocity [20], [11], [12], [25] such that

$$\mathbf{u} = \nabla\Phi + \mathbf{w}\ , \quad p = -\rho U\frac{\partial\Phi}{\partial x_1}. \tag{2.7}$$

The wake velocity $\mathbf{w}$ is obtained from the wake velocity potential χ defined separately for the drag and lift Oseenlet in the following sections 2.1 and 2.2, and the two potentials satisfy

$$\nabla^2\Phi = 0\ , \quad \left(\nabla^2 - 2k\frac{\partial}{\partial x_1}\right)\chi = 0. \tag{2.8}$$

2.1. Drag Oseenlet. This gives the unit drag [25], [19], where

$$\mathbf{u^d} = \nabla\phi^d + \nabla\chi^d - 2k\chi^d\mathbf{\hat{x}_1}\ , \quad p^d = -\rho U\frac{\partial\phi^d}{\partial x_1}, \tag{2.9}$$

$$\phi^d = \frac{1}{4\pi\rho U}\frac{\partial}{\partial x_1}\log(R - x_1)\ , \quad \chi^d = -\frac{1}{4\pi\rho U}e^{-k(R-x_1)}\frac{\partial}{\partial x_1}\log(R - x_1). \tag{2.10}$$

2.2. Lift Oseenlet. This gives the unit lift [25], [19], where

$$\mathbf{u^l} = \nabla\phi^l + \nabla\chi^l - 2k\chi^{*l}\mathbf{\hat{x}_2}\ , \quad p^l = -\rho U\frac{\partial\phi^l}{\partial x_1}, \tag{2.11}$$

$$\phi^l = \frac{1}{4\pi\rho U}\frac{\partial}{\partial x_2}\log(R - x_1)\ , \quad \chi^l = -\frac{1}{4\pi\rho U}e^{-k(R-x_1)}\frac{\partial}{\partial x_2}\log(R - x_1) \tag{2.12}$$

and where $\frac{\partial\chi^{*l}}{\partial x_2} = \frac{\partial\chi^l}{\partial x_1}$.

Consider the limit as the Reynolds number tends to infinity, such that outside of the boundary layer there is inviscid flow everywhere except in the core of the viscous wake comprising of trailing line vortices in which the viscous wake velocity term is present and such that the fluid velocity is finite there. Such a wake can be constructed from a distribution of singular lifting solutions over an area A. This results in a distribution of bound and free vortex lines, and the inviscid part is given by the velocity potential for quasi-inviscid flow

$$\phi(x_1, x_2, x_3) = \int\int_A \frac{l(y_1, y_3)}{4\pi\rho U}\frac{\partial}{\partial x_2}\ln(R_{13} - x_{11})\mathrm{d}y_1\mathrm{d}y_3. \tag{2.13}$$

A point in space is given by (x_1, x_2, x_3); a point on the area A is given by (y_1, y_2, y_3); $R_{13} = \sqrt{(x_1 - y_1)^2 + x_2^2 + (x_3 - y_3)^2}$; and $x_{11} = x_1 - y_1$.

The total lift is given by

$$L = \int\int_A l(y_1, y_3)\mathrm{d}y_1\mathrm{d}y_3. \tag{2.14}$$

For a slender body such that $0 \leq x_1 \leq x_e$, we can define a measure of the lift contribution from the singularities up to the station x_1 by

$$\begin{aligned} L(x_1) &= \int_0^{x_1} \int_{y_3} l(y_1, y_3) \mathrm{d}y_3 \mathrm{d}y_1 \\ &= \int_{y_3} l_1(x_1, y_3) \mathrm{d}y_3, \end{aligned} \tag{2.15}$$

where the double integration is such that it is over the area A, and so the total lift is

$$L = L(x_e), \tag{2.16}$$

where x_e is the end section of the body. The slip body boundary condition for inviscid flow is

$$\mathbf{u}.\mathbf{n} = \frac{\partial \Phi}{\partial n} = 0 \tag{2.17}$$

on the body surface, where Φ is the total velocity potential given by

$$\Phi = \phi^{(symm)} + \phi + Ux_1 + Vx_2 \tag{2.18}$$

for a uniform stream velocity $(U, V, 0)$; $\phi^{(symm)}$ is related to terms symmetric about the x_2 axis, and ϕ is related to terms antisymmetric about the x_2 axis. The boundary condition is satisfied by letting

$$\begin{aligned} \nabla \phi^{(symm)}.\mathbf{n} &= -U\mathbf{x_1}.\mathbf{n}, \\ \nabla \phi.\mathbf{n} &= -V\mathbf{x_2}.\mathbf{n}, \end{aligned}$$

where the solution for the potential ϕ, which is related to the lift, is of primary concern.

3. Inviscid flow theory. Inviscid flow theory gives the lift L^p determined from the pressure distribution over the body surface. Assume at each two-dimensional (2-D) cross section the 2-D Laplacian holds for the potential ϕ such that

$$\nabla^2 \phi = 0. \tag{3.1}$$

The boundary condition is then given by

$$\frac{\partial \phi}{\partial n} = -V\mathbf{x_2}.\mathbf{n}. \tag{3.2}$$

Consider a body with an elliptical cross section described by Figure 1. The ellipse is described by $\xi = \xi_0$, where (ξ, η) are the elliptic coordinates

$$\begin{aligned} x_2 &= c \cos \eta \sinh \xi, \\ x_3 &= c \sin \eta \cosh \xi. \end{aligned}$$

In elliptic coordinates, the uniform stream incident potential is then

$$\phi^{inc} = Vy = Vc \cos \eta \sinh \xi, \tag{3.3}$$

and the boundary condition is

$$\left. \frac{\partial \phi^{tot}}{\partial \xi} \right|_{\xi = \xi_0} = 0, \tag{3.4}$$

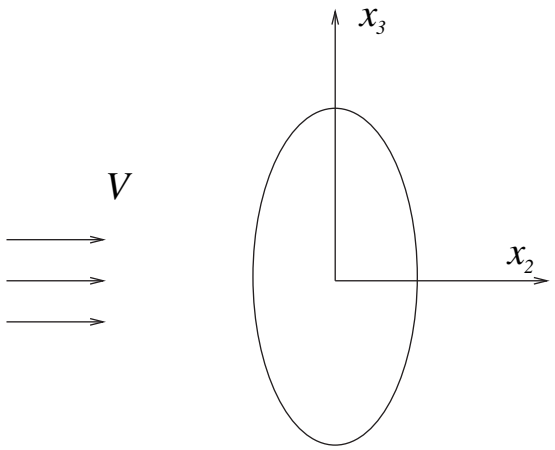


FIG. 1. *Elliptical cross section.*

where $\phi^{tot} = \phi^{inc} + \phi$. The total solution for the velocity potential is the sum of the uniform potential and a perturbation potential that exists due to the presence of the body in the uniform stream. This (perturbation) potential then satisfies

$$\phi = Vc\cos\eta\cosh\xi_0 e^{-(\xi-\xi_0)}. \tag{3.5}$$

3.1. Lift force. The lift force at the station x_1 from the pressure distribution alone, which in general is a different value from the lift measure $L(x_1)$, is given by the integral

$$\begin{aligned} L^p &= -\int\int_{S_B} p\mathbf{n}.\hat{\mathbf{y}}_\mathbf{2}\mathrm{d}S \\ &= \rho U\int_0^{x_e}\int_{ellipse}\frac{\partial\phi}{\partial y_1}\mathbf{n}.\hat{\mathbf{y}}_\mathbf{2}\mathrm{d}q\mathrm{d}y_1, \end{aligned} \tag{3.6}$$

where $\hat{\mathbf{y}}_\mathbf{2}$ is the unit vector in the y_2 direction, $\mathrm{d}q$ is an element of length along the ellipse boundary at this section, $y_1 = x_e$ is the position of the end section, and $\int_{ellipse}$ is the integral along the ellipse closed contour boundary. We then have

$$L^p(y_1) = \rho U\int_{ellipse}\phi\mathbf{n}.\hat{\mathbf{y}}_\mathbf{2}\mathrm{d}q. \tag{3.7}$$

Changing to elliptic coordinates, we have

$$L^p(y_1) = \rho U\int_0^{2\pi} Vc(y_1)\cos\eta\cosh\xi_0 e^{-(\xi-\xi_0)}c\cos\eta\cosh\xi\mathrm{d}\eta\Big|_{\xi=\xi_0}, \tag{3.8}$$

since $\mathbf{n} = (0, \frac{\partial y_2}{\partial\xi}, \frac{\partial y_3}{\partial\xi})/\sqrt{(\frac{\partial y_2}{\partial\xi})^2 + (\frac{\partial y_3}{\partial\xi})^2}$, $\mathrm{d}q = \sqrt{(\frac{\partial y_2}{\partial\eta})^2 + (\frac{\partial y_3}{\partial\eta})^2}\mathrm{d}\eta$, and $(\frac{\partial y_2}{\partial\xi})^2 + (\frac{\partial y_3}{\partial\xi})^2 = (\frac{\partial y_2}{\partial\eta})^2 + (\frac{\partial y_3}{\partial\eta})^2$, where the distance of the two focii of the ellipse from the coordinate origin is c. This gives the standard result

$$L^p(y_1) = \rho UV(\pi s(y_1)^2), \tag{3.9}$$

where $s(y_1)$ is the semispan of the ellipse $s(x_1) = c(x_1)\cosh\xi_0$ and determined using different methods from this one in different contexts by Lighthill [21], Jones [15], and Nielsen [24]. However, this method does not calculate the additional lift due to viscous terms. To determine this, we consider Oseen flow and use matched asymptotics rather than apply the Allen and Perkins method.

4. Lift from Oseen flow slender body theory. By continuing the approximate 2-D near-field flow into the slender body and onto a singular sheet, we can represent the flow by an integral distribution of normal 2-D dipoles; see the appendix. Similarly, by using the slender body approximation, the flow near the same singular sheet can be approximated from the integral representation of lifting elements. This approximation is also given in terms of an integral distribution of normal 2-D dipoles, and in this way the two flows are matched.

First, from Green's integral theorem it can be shown that ϕ can be represented by (see the appendix)

$$\begin{aligned}\phi(x_1,x_2,x_3) &= \int_{-c(x_1)}^{c(x_1)} \frac{f(y_1,y_3)}{2\pi}\frac{\partial}{\partial y_2}\ln r_{23}\mathrm{d}y_3\bigg|_{y_2=0}\\ &= -\int_{-c(x_1)}^{c(x_1)} \frac{f(y_1,y_3)}{2\pi}\frac{x_2}{x_2^2+(x_3-y_3)^2}\mathrm{d}y_3,\end{aligned} \tag{4.1}$$

where $r_{23} = \sqrt{(x_2-y_2)^2+(x_3-y_3)^2}$, $c(x_1)$ is the x_3 position of the focii of the ellipse boundary at section x_1, and $f(x_1,x_3)$ is the dipole strength along the singularity sheet area A. The singular sheet is defined within the above integral limits. Therefore the discontinuity in ϕ across the singular sheet is given by

$$\begin{aligned}\phi(x_1,x_2\to 0_\pm,x_3) &= \lim_{x_2\to 0}\left\{\mp\frac{f(x_1,x_3)}{2\pi}\left[\tan^{-1}\frac{c(x_1)-x_3}{|x_2|}-\tan^{-1}\frac{-c(x_1)-x_3}{|x_2|}\right]\right\}\\ &= \mp\frac{f(x_1,x_3)}{2}\end{aligned} \tag{4.2}$$

for $0\le x_1\le x_e$, $-c(x_1)\le x_3\le c(x_1)$. So $f(x_1,x_3) = \phi(x_1,x_2\to 0_-,x_3)-\phi(x_1,x_2\to 0_+,x_3)$ as shown in (A.9).

4.1. Normal dipole representation of the near-field 2-D flow representation. The near-field 2-D inviscid flow has been obtained in section 3 and is given by the potential (3.5). This potential represents a distribution of singularities along the line between the ellipse focii. Hence, we can continue (analytically in this case) the potential up to this line, replacing the body boundary by a region of fluid. In this way, the type and strength of the singularities that generate this flow can be determined. From the appendix, it is demonstrated that the flow can be represented by a distribution of dipoles normal to the line connecting the ellipse focii. The appendix also enables us to find the strength of the distribution. Continuing the flow into the ellipse such that $\xi\to 0$, then

$$\begin{aligned}\phi &= Vc(x_1)\cos\eta\cosh\xi_0 e^{-(\xi-\xi_0)}\\ &\sim \pm V\sqrt{c(x_1)^2-x_3^2}\cosh\xi_0 e^{\xi_0}\end{aligned} \tag{4.3}$$

since $\xi(x_1,x_2\to 0_\pm)\sim\pm\frac{x_2}{\sqrt{c(x_1)^2-x_3^2}}$. Hence, from (4.2) and (A.9), the strength is given by

$$f(x_1,x_3) = -2V\sqrt{c(x_1)^2-x_3^2}\cosh\xi_0 e^{\xi_0}. \tag{4.4}$$

4.2. Normal dipole representation of the 3-D flow representation. We use the slender body approximation given by Chadwick [6] relating a line distribution of lifting elements to a near-field approximate distribution of normal dipoles. This gives, using (2.13),

$$\begin{aligned}\phi &= \int\int_A \frac{l(y_1,y_3)}{4\pi\rho U}\frac{\partial}{\partial x_2}\ln(R_{13}-x_{11})\mathrm{d}y_1\mathrm{d}y_3 \\ &\sim \int_{-c(x_1)}^{c(x_1)} \frac{l_1(y_1,y_3)}{2\pi\rho U}\frac{x_2}{x_2^2+(x_3-y_3)^2}\mathrm{d}y_3,\end{aligned} \tag{4.5}$$

where $\frac{\partial}{\partial y_1}l_1(y_1,y_3)=l(y_1,y_3)$. Hence

$$f(y_1,y_3) = -\frac{l_1(y_1,y_3)}{\rho U}. \tag{4.6}$$

4.3. Matching the two flows to find the lift force. Matching the two flows gives

$$\frac{l_1(y_1,y_3)}{\rho U} = 2V\sqrt{c(y_1)^2-y_3^2}\cosh\xi_0 e^{\xi_0};$$

therefore,

$$\begin{aligned}L(y_1) &= 2\rho UV\cosh\xi_0 e^{\xi_0}\int_{-c(y_1)}^{c(y_1)}\sqrt{c(y_1)^2-y_3^2}\mathrm{d}y_3 \\ &= \pi\rho UVc(y_1)^2\cosh\xi_0 e^{\xi_0}.\end{aligned} \tag{4.7}$$

Relating the potential ϕ to the Oseen potential then gives $L(y_1)=\int_{-c(y_1)}^{c(y_1)} l_1(y_1,y_3)\mathrm{d}y_3$, which is the lift generated at each section. So,

$$L(y_1) = \pi s(y_1)^2\rho UV\left(\frac{e^{\xi_0}}{\cosh\xi_0}\right), \tag{4.8}$$

where $s(y_1)$ is the semispan of the major axis of the ellipse at station y_1, and so $s(y_1)=c(y_1)\cosh\xi_0$.

Hence, the total lift is given by

$$\begin{aligned}L &= (\pi s_e^2)\rho UV\left(\frac{e^{\xi_0}}{\cosh\xi_0}\right) \\ &= (\pi s_e^2)\rho UV\left(\frac{s_e^{min}+s_e}{s_e}\right),\end{aligned} \tag{4.9}$$

where the end of the slender body is at station y_e, and $s_e = s(y_e)$. s_e^{min} is the semispan of the minor axis of the ellipse at station y_e.

Alternatively, the total lift can be found by considering the end section only, in the following way.

4.4. Alternative calculation for lift at the end section. We follow the method given in [6] to find a general formula for the lift of a slender body with an elliptical cross section.

In the far field, the elliptic coordinates tend to $\eta \sim \theta$ and $e^{-\xi} \sim \frac{c}{2r}$ according to Batchelor [4, p. 465].

Hence, the potential at the end section becomes

$$\begin{aligned}\phi &= Vc(x_1)\cos\eta\cosh\xi_0 e^{-(\xi-\xi_0)}\Big|_{x_1=x_e} \\ &\sim \left(\frac{\cos\theta}{r}\right)\left(\frac{1}{2}c(x_1)^2 V\cosh\xi_0 e^{\xi_0}\right)\Big|_{x_1=x_e}. \end{aligned} \tag{4.10}$$

Using [6, equation (4.3)], the lift is then given by

$$\begin{aligned} L &= \pi\rho UVc(x_1)^2\cosh\xi_0 e^{\xi_0}\big|_{x_1=x_e} \\ &= (\pi s(x_1)^2)\,\rho UV\left(\frac{e^{\xi_0}}{\cosh\xi_0}\right)\Big|_{x_1=x_e} \\ &= (\pi s_e^2)\rho UV\left(\frac{s_e^{min}+s_e}{s_e}\right). \end{aligned} \tag{4.11}$$

Therefore, the lift on the body L is not given by the lift calculation over the slender body surface $L^p(x_e)$ calculated from the surface pressure. This is because, in order to find the total lift, the contribution from the viscous term must also be included in the lift calculation.

5. Lift on a delta wing. For a delta wing, with rounded leading edges such that flow separation is avoided, the lift over its surface between the ends is given by

$$L^p = (\pi s_e^2)\rho UV. \tag{5.1}$$

The Oseen lift on the body is given by

$$L = (\pi s_e^2)\rho UV. \tag{5.2}$$

Hence the two lift evaluations agree with each other and also with the lift expression given by Jones [15], who then verifies it by experiment.

6. Lift on a slender body with a circular cross section. The lift from inviscid flow theory due to pressure over the surface of a body with a circular cross section between the ends is given by

$$L^p = (\pi s_e^2)\rho UV. \tag{6.1}$$

This is in contrast to the lift on the body due to the Oseen theory, which is given by

$$L = 2(\pi s_e^2)\rho UV. \tag{6.2}$$

So there is a doubling of the lift force due to the viscous terms within the Oseen representation. This is similar to the motivation of the viscous cross-flow method of Allen and Perkins [1], who assume an additional contribution to the lift from viscous forces. However, they determine this from empirical data related to cross-flow drag past a circular cylinder applied at each body section. In Figure 2, we compare the Oseen theory with the viscous cross-flow method of Allen and Perkins against the two experiments given in their report. We see that both theories give good first order approximations to the experiments, unlike the inviscid lift calculation from the pressure field. Small angles only have been taken because a linear variation in the angle of attack is assumed. Many other NACA reports give similar results, and these are detailed in [10]. However, Figure 3 shows that, for a less conventional missile profile, results were obtained by Jack [14] which show that the Allen-Perkins method compares far less favorably.

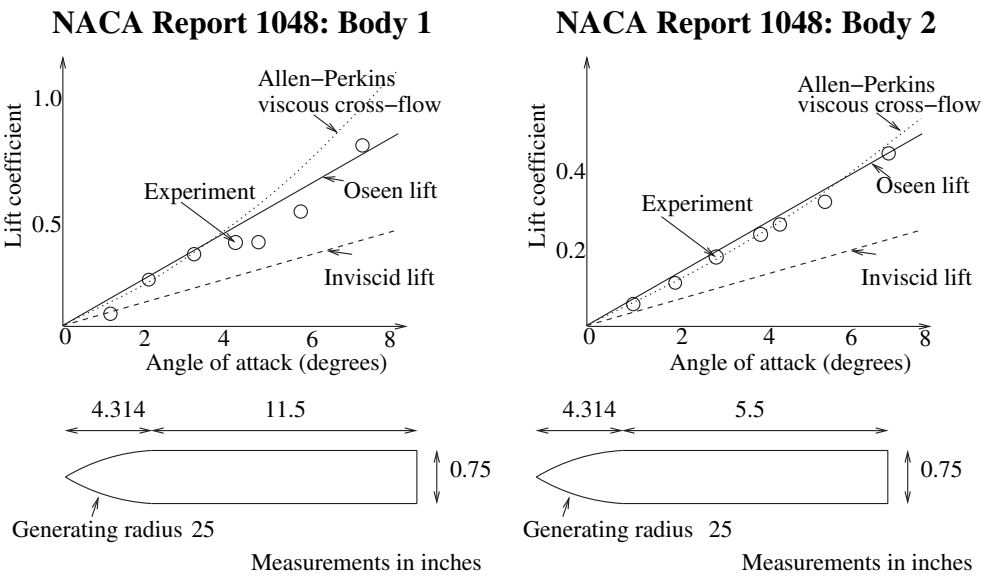


FIG. 2. *Lift of slender bodies with a circular cross section.*

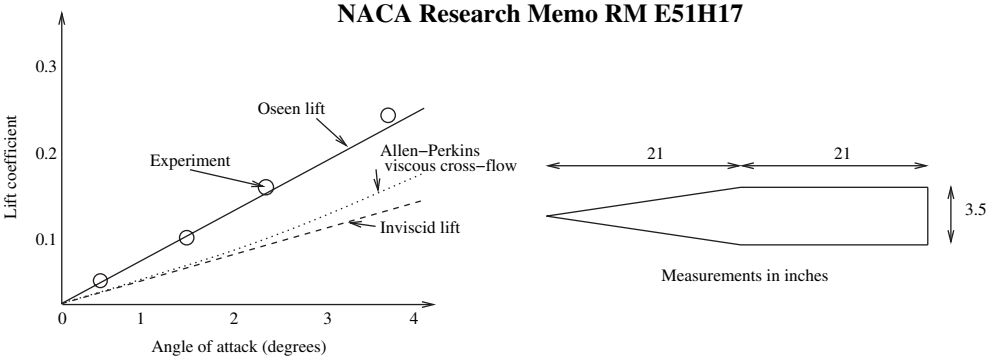


FIG. 3. *Lift of slender bodies with a circular cross section.*

7. Comparison with experiment and other theories for slender bodies with an elliptical cross section. Consider a slender body of an elliptical cross section with a fixed ellipticity. In Figure 4, we plot on the y-axis the ratio of the lift coefficient with the lift coefficent of an equivalent slender body with a circular cross section such that the semimajor axis b equals the circular cylinder radius r. On the x-axis we plot the ellipticity given by the ratio of the semiminor axis to the semimajor axis a/b. A variety of such slender bodies with varying ellipticity from 0 to 1 is considered. For an ellipticity of 0, the slender body is a slender wing, and the result of Jones [15] applies. The horizontal inviscid lift line also goes through this point, since Lighthill's result [21] states that the inviscid lift is dependent upon the maximum span of the slender body only; for slender bodies with an elliptical cross section, this result is also given by (3.9). The Oseen lift line is also plotted in the figure from the result of (4.9). We see that, for a slender body with a circular cross section, the total lift is twice that for a slender wing with the same maximum span. Also plotted in the figure

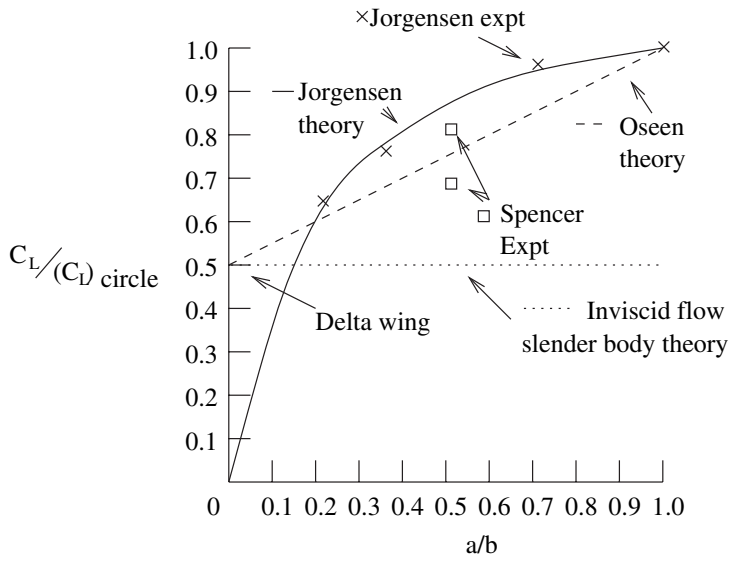


FIG. 4. *Ratio of lift coefficient to lift coefficient of a circular cross section slender body plotted against ellipticity.*

are the expected results from Jorgensen's Newtonian impact theory [16]. In [17], Jorgensen gives the percentage reduction from the viscous cross-flow method of Allen and Perkins (which is applied to a slender body with a circular cross section) for a slender body with a particular ellipticity. Jorgensen's theoretical results match well with his experimental results and better than the Oseen theory presented here. However, consider two experimental results denoted by squares plotted in the figure given by Sigal [26] quoting the results of Spencer and Phillips [28]. Sigal gives the percentage reduction in lift force evaluated from experiment for a slender body with an elliptical cross section compared to a slender body with a circular cross section having the same span. Two results for ellipticity 0.5 but with different fineness ratio are plotted, and the Oseen theory fits these experimental results better than Jorgensen's theory.

8. Discussion. The slender body theory in Oseen flow given by Chadwick [6] is applied to bodies with an elliptical cross section and compared with various other theories and methods and with experimental results. In particular, the lift on a slender body with an elliptical cross section is given in Oseen flow by representing the flow field by a distribution of lift Oseenlets between the focii of the ellipse at each cross section. The potential part of this solution is then matched to the expected two-dimensional near-field solution given in elliptic coordinates such that the slip (impermeability) boundary condition is satisfied. This lift formula is compared against other theories and experiment, such as Jones's results for a delta wing [15], the Allen and Perkins viscous cross-flow method [1], and Jorgensen's Newtonian impact theory [16]. For a delta wing with zero ellipticity, the Oseen theory predicts that the additional viscous force is zero and so reduces to the standard inviscid result given by Jones [15]. In contrast, the viscous cross-flow method of Jorgensen cannot be applied to this problem and gives zero total force (see Figure 4). The moment depends upon the distribution of lift, and the experimental results suggest that the additional viscous lift force (over and above the lift force from inviscid potential flow theory) is distributed towards the

rear of the slender body rather than close to the middle as predicted by the Allen and Perkins viscous cross-flow method. So, for the moment calculation at angles at or below 4°, the experimental results agree better with the Oseen flow method. These results are detailed extensively by Fishwick [10].

Appendix A. We start with Green's integral theorem for 2-D potential flow in the $x_2 - x_3$ plane, which defines the flow as an integral distribution of sources and dipoles over a closed contour C_y such that

$$\phi(x_2, x_3) = -\frac{1}{2\pi} \int_{C_y} \left(\phi(y_2, y_3) \frac{\partial}{\partial n} \ln r_{23} - \frac{\partial}{\partial n} \phi(y_2, y_3) \ln r_{23} \right) \mathrm{d}y_C, \tag{A.1}$$

where (x_2, x_3) is a point in the fluid, (y_2, y_3) is a point on the contour, a length element along the contour C_y is given by $\mathrm{d}y_C$, and $r_{23} = \{(x_2 - y_2)^2 + (x_3 - y_3)^2\}^{1/2}$. Consider ϕ such that it can be continued onto the line $x_3 = 0$, $-c \le x_2 \le c$. Let the closed contour C include the two lines $y_3 = \delta$, $-c \le y_2 \le c$ and $y_3 = -\delta$, $-c \le y_2 \le c$ and also include the two semicircular arcs $r_\delta^- = \delta$, $0 \le \theta^- \le \pi$, $y_2 + c = r_\delta^- \cos\theta^-$, $y_3 = r_\delta^- \sin\theta^-$ and $r_\delta^+ = \delta$, $-\pi \le \theta^+ \le 0$, $y_2 - c = r_\delta^+ \cos\theta^+$, $y_3 = r_\delta^+ \sin\theta^+$.

We go around the contour in the clockwise sense, and the contour is described pictorially in Figure 5. Letting $\delta \to 0$, we then get the expression for ϕ:

$$\begin{aligned} \phi = &-\frac{1}{2\pi} \int_{-c}^{c} \left\{ (\phi^+ - \phi^-) \frac{\partial}{\partial y_2} \ln r_{23} - \left(\frac{\partial}{\partial y_2} \phi^+ - \frac{\partial}{\partial y_2} \phi^- \right) \ln r_{23} \right\} \mathrm{d}y_3 \\ &+ I^+ + I^-, \end{aligned} \tag{A.2}$$

where $\phi^\pm = \lim_{y \to 0^\pm} \phi$ and $\frac{\partial \phi^\pm}{\partial y_2} = \lim_{y \to 0^\pm} \frac{\partial \phi}{\partial y_2}$ are assumed to exist, and

$$\begin{aligned} I^+ &= \lim_{r_\delta^+ \to 0} \left\{ \frac{1}{2\pi} \int_0^{\pi} \left(\phi \frac{\partial}{\partial r_\delta^+} \ln r_{23} - \ln r_{23} \frac{\partial}{\partial r_\delta^+} \phi \right) r_\delta^+ \mathrm{d}\theta^+ \right\}, \\ I^- &= \lim_{r_\delta^- \to 0} \left\{ \frac{1}{2\pi} \int_0^{-\pi} \left(\phi \frac{\partial}{\partial r_\delta^-} \ln r_{23} - \ln r_{23} \frac{\partial}{\partial r_\delta^-} \phi \right) r_\delta^- \mathrm{d}\theta^- \right\}. \end{aligned} \tag{A.3}$$

In particular, consider the solution (3.5) $\phi = A \cos\eta e^{-\xi}$, where $A = Vc \cosh\xi_0 e^{\xi_0}$. Then $\phi^\pm$ and $\frac{\partial \phi^\pm}{\partial y_2}$ exist. Furthermore, $\frac{\partial \phi^+}{\partial y_2} = \frac{\partial \phi^-}{\partial y_2}$.

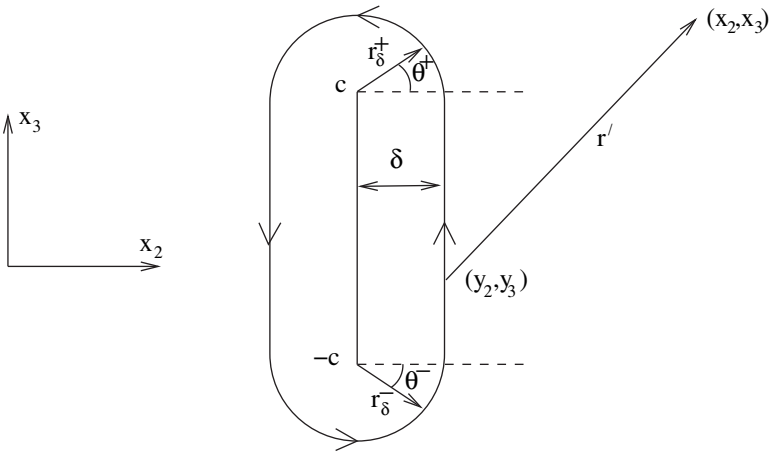


FIG. 5. *The closed contour C.*

We will now show that the integrals $I^{\pm}$ are zero. Consider the first terms in the integral expressions given for $I^{\pm}$. The first factor in the term is ϕ. In the limit as $r_{\delta}^{\pm} \to 0$, $\xi \to 0$ and $\eta \to \pm\pi/2$. So,

$$\lim_{r_{\delta}^{\pm} \to 0} \phi = 0. \tag{A.4}$$

The second factor in the first term is $\frac{\partial \ln r_{23}}{\partial r_{\delta\pm}}$. This is

$$\begin{aligned} \frac{\partial \ln r_{23}}{\partial r_{\delta\pm}} &= \frac{1}{r_{23}} \frac{\partial r_{23}}{\partial r_{\delta\pm}} = \frac{1}{r_{23}} \left\{ \frac{\partial y_2}{\partial r_{\delta\pm}} \frac{\partial r_{23}}{\partial y_2} + \frac{\partial y_3}{\partial r_{\delta\pm}} \frac{\partial r_{23}}{\partial y_3} \right\} \\ &= O(1/r_{23}). \end{aligned} \tag{A.5}$$

So the first term in the integrands of the integral expressions for $I^{\pm}$ give integral contributions $O(|\phi| r_{\delta}^{\pm}/r_{23})$, and in the limit as $r_{\delta}^{\pm} \to 0$, these tend to zero.

The first factor in the second term of the integrand of $I^{\pm}$ is $\frac{\partial \phi}{\partial r_{\delta}^{\pm}}$. We find the order of magnitude of this term by considering

$$\left| \frac{\partial \phi}{\partial r_{\delta\pm}} \right| \leq \left| \frac{\partial \phi}{\partial \xi} \right| \left| \frac{\partial \xi}{\partial r_{\delta\pm}} \right| + \left| \frac{\partial \phi}{\partial \eta} \right| \left| \frac{\partial \eta}{\partial r_{\delta\pm}} \right| \tag{A.6}$$

as $r_{\delta}^{\pm} \to 0$. Similarly,

$$\begin{aligned} \left| \frac{\partial \xi}{\partial r_{\delta\pm}} \right| &\leq \left| \frac{\partial \xi}{\partial y_2} \frac{\partial y_2}{\partial r_{\delta\pm}} \right| + \left| \frac{\partial \xi}{\partial y_3} \frac{\partial y_3}{\partial r_{\delta\pm}} \right| \\ &\leq \frac{1}{c} \end{aligned} \tag{A.7}$$

since $\frac{\partial \xi}{\partial y_2} \sim \frac{1}{c}$, $\frac{\partial \xi}{\partial y_3} \sim 0$, $\frac{\partial \eta}{\partial y_2} \sim 0$, and $\frac{\partial \eta}{\partial y_3} \sim \frac{1}{c}$ as $r_{\delta}^{\pm} \to 0$. Furthermore, $\frac{\partial \phi}{\partial \xi} \sim A$ and $\frac{\partial \phi}{\partial \eta} \sim 0$ as $r_{\delta}^{\pm} \to 0$, since $\phi = A \cos \eta e^{-\xi}$. So the second term in the integrands of $I^{\pm}$ gives the integral contributions of order $O(A|\ln r_{23}| r_{\delta}^{\pm}/c)$. So as $r_{\delta}^{\pm} \to 0$, then the integral contributions tend to zero. Hence in the limit $I^{+} = I^{-} = 0$. So

$$\phi = -\frac{1}{2\pi} \int_{-c}^{c} \left\{ (\phi^{+} - \phi^{-}) \frac{\partial}{\partial y_2} \ln r_{23} - \left(\frac{\partial}{\partial y_2} \phi^{+} - \frac{\partial}{\partial y_2} \phi^{-} \right) \ln r_{23} \right\} \mathrm{d}y_3. \tag{A.8}$$

From the symmetry of the problem, $\frac{\partial \phi^{+}}{\partial y_2} = \frac{\partial \phi^{-}}{\partial y_2}$, and so

$$\phi = -\frac{1}{2\pi} \int_{-c}^{c} \left\{ (\phi^{+} - \phi^{-}) \frac{\partial}{\partial y_2} \ln r_{23} \right\} \mathrm{d}y_3. \tag{A.9}$$

Acknowledgment. We thank Dr. David Clarke of Newcastle University for continued support and guidance, in particular for drawing our attention to many of the references cited.

REFERENCES

[1] H. J. Allen and E. W. Perkins, *A Study of Effects of Viscosity on Flow over Slender Inclined Bodies of Revolution*, report 1048, Nat. Adv. Comm. Aero., 1951.

[2] G. K. Batchelor, *On steady laminar flow with closed streamlines at large Reynolds number*, J. Fluid Mech., 1 (1956), pp. 177–190.

[3] G. K. Batchelor, *Axial flow in trailing line vortices*, J. Fluid Mech., 20 (1964), pp. 645–658.
[4] G. K. Batchelor, *An Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge, 1967.
[5] E. Chadwick, *The far field Oseen velocity expansion*, Proc. R. Soc. London Ser. A, 454 (1998), pp. 2059–2082.
[6] E. Chadwick, *A slender body theory in Oseen flow*, Proc. R. Soc. London Ser. A, 458 (2002), pp. 2007–2016.
[7] E. Chadwick, *A slender-wing theory in potential flow*, Proc. R. Soc. London Ser. A, 461 (2005), pp. 415–432.
[8] E. Chadwick, *The vortex line in steady, incompressible Oseen flow*, Proc. R. Soc. London Ser. A, 462 (2006), pp. 391–401.
[9] D. Clarke, *A two-dimensional strip method for surface ship hull derivatives*, J. Mech. Engng. Science, 14 (1972), pp. 53–61.
[10] N. J. Fishwick, *Manoeuvring of Slender Bodies through Fluids*, Ph.D. thesis, University of Salford, 2006.
[11] S. Goldstein, *The forces on a solid body moving through a viscous fluid*, Proc. R. Soc. London Ser. A, 123 (1929), pp. 216–230.
[12] S. Goldstein, *The forces on a solid body moving through a viscous fluid*, Proc. R. Soc. London Ser. A, 131 (1931), pp. 198–208.
[13] G. Hearn, D. Clarke, H. Chan, A. Incecik, and K. Varyani, *The influence of vorticity upon estimation of manoeuvring derivatives*, in Proceedings of the 20th Symposium on Naval Hydrodynamics, National Academic Press, 1994, pp. 669–681.
[14] J. R. Jack, *Aerodynamic Characteristics of a Slender Cone-Cylinder Body of Revolution at a Mach Number of* 3.85, research memo, RM E51H17, Nat. Adv. Comm. Aero., 1951.
[15] R. T. Jones, *Properties of Low-Aspect-Ratio Pointed Wings at Speeds Below and Above the Speed of Sound*, report 835, Nat. Adv. Comm. Aero., 1945.
[16] L. H. Jorgensen, *Inclined Bodies of Various Cross Sections at Supersonic Speeds*, NASA technical memo, NASA MEMO 10-3-58A, 1958.
[17] L. H. Jorgensen, *Prediction of Static Aerodynamic Characteristics for Space-Shuttle-like and Other Bodies at Angles of Attack from* 0° *to* 180°, NASA technical note, NASA TN D-6996, 1973.
[18] J. Katz and A. Plotkin, *Low-Speed Aerodynamics*, 2nd ed., Cambridge University Press, Cambridge, 2001.
[19] P. A. Lagerstrom, *Laminar Flow Theory, Vol.* 6: *High Speed Aerodynamics and Jet Propulsion*, F. K. Moore, ed., Princeton University Press, Princeton, NJ, 1964.
[20] H. Lamb, *Hydrodynamics*, Cambridge University Press, Cambridge, 1932.
[21] M. J. Lighthill, *Note on the swimming of slender fish*, J. Fluid Mech., 9 (1960), pp. 305–317.
[22] M. M. Munk, *The Aerodynamic Forces on an Airship Hull*, report 184, Nat. Adv. Comm. Aero., Washington, 1924.
[23] J. N. Newman, *Marine Hydrodynamics*, MIT Press, Cambridge, MA, 1978.
[24] J. N. Nielsen, *Missile Aerodynamics*, McGraw-Hill, New York, 1960.
[25] C. W. Oseen, *Neure Methoden und Ergebnisse in der Hydrodynamik*, Akad. Verlagsgesellschaft, Leipzig, 1927.
[26] A. Sigal, *Methods of analysis and experiments for missiles with noncircular fuselages*, in Tactical Missile Aerodynamics, Chap. 5, M. R. Mendenhall, ed., 1992, pp. 171–223.
[27] J. H. B. Smith, *Improved calculations of leading-edge separation from slender, thin, delta wings*, Proc. R. Soc. London Ser. A, 306 (1968), pp. 67–90.
[28] B. Spencer and W. P. Phillips, *Transonic Aerodynamic Characteristics of a Series of Bodies Having Variations in Fineness Ratio and Cross-Sectional Ellipticity*, NASA Technical note TN D-2622, 1965.

SIAM J. APPL. MATH.
Vol. 67, No. 5, pp. 1479–1503

STATIONARY PATTERN OF A RATIO-DEPENDENT FOOD CHAIN MODEL WITH DIFFUSION*

RUI PENG†, JUNPING SHI‡, AND MINGXIN WANG§

Abstract. In the paper, we investigate a three-species food chain model with diffusion and ratio-dependent predation functional response. We mainly focus on the coexistence of the three species. For this coupled reaction-diffusion system, we study the persistent property of the solution, the stability of the constant positive steady state solution, and the existence and nonexistence of nonconstant positive steady state solutions. Both the general stationary pattern and Turing pattern are observed as a result of diffusion. Our results also exhibit some interesting effects of diffusion and functional responses on pattern formation.

Key words. food chain model, diffusion, ratio-dependent functional response, stationary pattern, Turing pattern, steady state solution

AMS subject classifications. 35J55, 92C15, 92D40

DOI. 10.1137/05064624X

1. Introduction.

Understanding of spatial and temporal behaviors of interacting species in ecological systems is a central issue in population ecology. One aspect of great interest for a model with multispecies interactions is whether the involved species can persist or even stabilize at a coexistence steady state. In the case where the species are homogeneously distributed, this would be indicated by a constant positive solution of an ordinary differential equation (ODE) system. In the spatially inhomogeneous case, the existence of a nonconstant time-independent positive solution, also called stationary pattern, is an indication of the richness of the corresponding partial differential equation (PDE) dynamics. In recent years, stationary pattern induced by diffusion has been studied extensively, and many important phenomena have been observed.

In particular, starting with Turing's seminal 1952 paper [34], diffusion has been regarded as the driving force of the spontaneous emergence of spatiotemporal structure in a variety of nonequilibrium situations. To verify the influence of diffusion on this aspect, in the past decades, biologists and applied mathematicians have proposed a number of models, and much work has been devoted to the investigation of the existence of stationary pattern in chemical and biological dynamics theoretically as well as numerically. For example, chemical models include the activator-inhibitor Gierer–Meinhardt model [10, 23], the Sel'kov model [7, 35], the Gray–Scott model [32, 37], the Brusselator model [3, 30], the Noyes–Field model for Belousov–Zhabotinskii

*Received by the editors November 29, 2005; accepted for publication (in revised form) April 26, 2007; published electronically August 24, 2007.
http://www.siam.org/journals/siap/67-5/64624.html

†Corresponding author. Institute of Nonlinear Complex System, College of Science, China Three Gorges University, Yichang, 443002, Hubei, People's Republic of China (pengrui_seu@163.com). The work of this author was supported by the Scientific Research Projects of Hubei Provincial Department of Education Q200713001.

‡Department of Mathematics, College of William and Mary, Williamsburg, VA 23187-8795 and Department of Mathematics, Harbin Normal University, Harbin, 150080, Helongjiang, People's Republic of China (jxshix@wm.edu). The work of this author was partially supported by US-NSF grants DMS-0314736 and EF-0436318, NSFC grant 10671049, and a Longjiang scholar grant.

§Department of Mathematics, Southeast University, Nanjing, 210018, Jiangsu, People's Republic of China (mxwang@seu.edu.cn). The work of this author was supported by NSFC grant 10471022.

reaction [29], and the chemotactic diffusion model [4, 18, 19, 22, 24, 37], and biological models include the competition model [14, 20] and the predator-prey model [8, 15, 16, 27, 28, 31, 36] (see also the references therein).

In his original paper [34], Turing proposed the notion of diffusion-driven instability (also called Turing instability) in his attempt at modeling, among other things, the regeneration phenomenon of hydra—one of the earliest examples of morphogenesis. That is, Turing claimed that the formation of spatial pattern during morphogenesis could be explained in terms of the instability of a homogeneous steady state solution to a reaction-diffusion network describing the growth and movement of a set of morphogens. Turing's original work was primarily concerned with the stability analysis of the uniform steady state solution of the system for the interacting morphogens.

In biology and chemistry, the more interesting question, however, is whether the spatially inhomogeneous solution may be generated by such instability. Strikingly, in some cases, Turing instability can indeed lead to stationary pattern (also called Turing pattern), a fascinating phenomenon in nonlinear science, which has been found in various mechanisms [4, 18, 26, 27, 28, 30, 35, 36, 37]. While linear stability analysis of the homogeneous steady state is a straightforward method for calculating conditions for the onset of Turing instability, the analysis of the existence of resulting nonhomogeneous steady states is mathematically challenging. In this paper, it is the question of the existence of nonhomogeneous steady states that we focus on.

In the present work, we will investigate a coupled reaction-diffusion food chain model with ratio-dependent functional response and analyze the coexistence of the three species. We attempt to further understand the influences of diffusion and functional responses on pattern formation. As a consequence, the existence and nonexistence results for nonconstant positive steady state solutions to this system indicate that stationary pattern arises as the diffusion coefficients enter into certain regions. In other words, diffusion does help to create stationary pattern. For this model, we also show that Turing instability occurs and prove the generation of Turing pattern in some cases.

On the other hand, our results also demonstrate that diffusion and functional response can become determining factors in the formation of pattern. Although our model is very different from the one considered by Lou, Martinez, and Ni in [20], their interesting observation that the introduction of a new species may qualitatively change the pattern structure of the original system is again present in our study. At the same time, our work corroborates recent numerical results implemented by Alonso, Bartumeus, and Catalan in [2]. We refer the reader to more detailed discussions in section 6.

Our paper is organized as follows. In section 2, we propose our mathematical model. In section 3, we discuss the persistence and stability of the unique constant positive steady state for the ODE and PDE systems. In section 4, we consider the nonexistence of nonconstant positive steady state solutions, while section 5 is devoted to the existence of nonconstant positive steady state solutions. In section 6, from the biological viewpoint we make some comments on our studies, indicating some interesting influences of diffusion and functional responses on pattern formation. Finally, in the appendix, we analyze some conditions, which are imposed in section 5 to obtain the nonconstant positive steady state solutions to the PDE system.

2. The derivation of the mathematical model. Numerous examples from biological control indicate that the classical prey-dependent predator-prey model is often contrary to actual observations, such as the well-known paradox of enrichment formulated by Rosenzweig [33]. The theory of Rosenzweig states that enriching a

predator-prey system (increasing the prey's carrying capacity) will cause an increase in the equilibrium density of the predator but not in that of the prey; it will destabilize the positive equilibrium as the prey's carrying capacity increases, and thus will increase the possibility of stochastic extinction of the predator. Recently there is growing evidence that in some situations, especially when predators have to search, share, and compete for food, a more suitable general predator-prey model should be a so-called ratio-dependent one (namely, the functional responses are ratio-dependent). Roughly speaking, this model states that the per capita predator growth rate should be a function of the ratio of prey to predator abundance (see, e.g., [1]).

In the case of multiple species interaction, the prey-dependent models such as those studied in [5, 9, 11, 17], while mathematically interesting, inherit the mechanism that generates the factitious paradox of enrichment and fail to produce the often observed extinction dynamics resulting in the collapse of the system. Consequently, a ratio-dependent food chain model, which is an ODE system with three equations whose species are hence assumed to be spatially homogeneous, was proposed by Hsu, Hwang, and Kuang in [13] to describe the growth of plant, pest, and top predator.

More precisely, the authors of [13] considered the following three–trophic-level food chain system with ratio-dependent functional response:

$$\begin{cases} \dfrac{\mathrm{d}u_1}{\mathrm{d}t} = ru_1\left(1-\dfrac{u_1}{k}\right) - \dfrac{1}{\eta_1}\dfrac{m_1u_1u_2}{u_1+c_1u_2}, & t>0,\\ \dfrac{\mathrm{d}u_2}{\mathrm{d}t} = \dfrac{m_1u_1u_2}{u_1+c_1u_2} - b_1u_2 - \dfrac{1}{\eta_2}\dfrac{m_2u_2u_3}{u_2+c_2u_3}, & t>0,\\ \dfrac{\mathrm{d}u_3}{\mathrm{d}t} = \dfrac{m_2u_2u_3}{u_2+c_2u_3} - b_2u_3, & t>0,\\ u_1(0)>0, \quad u_2(0)>0, \quad u_3(0)>0, \end{cases} \tag{2.1}$$

where u_i $(i=1,2,3)$ are the respective population densities of prey, predator, and top predator. For $i=1,2$, η_i, m_i, c_i, and b_i represent the yield constants, maximal predator growth rates, half-saturation constants, and predator's death rates, respectively. Constants r and k are the prey intrinsic growth rate and carrying capacity, respectively. As observed in [13], u_3 preys on u_2 and only on u_2, and u_2 preys on u_1 and nutrient recycling is not accounted for, which produces the so-called simple food chain. A distinct feature of the simple food chain is the domino effect: if one species dies out, all the species at higher trophic levels die out as well.

As in [13], for simplicity, we use the following scaling to (2.1):

$$\begin{array}{cccc} t\to rt, & u_1\to u_1/k, & u_2\to c_1u_2/k, & u_3\to c_1c_2u_3/k,\\ m_1\to m_1/r, & b_1\to b_1/r, & m_2\to m_2/r, & b_2\to b_2/r, \end{array}$$

and (2.1) becomes the form

$$\begin{cases} \dfrac{\mathrm{d}u_1}{\mathrm{d}t} = u_1\left(1-u_1\right) - \dfrac{a_1u_1u_2}{u_1+u_2}, & t>0,\\ \dfrac{\mathrm{d}u_2}{\mathrm{d}t} = \dfrac{m_1u_1u_2}{u_1+u_2} - b_1u_2 - \dfrac{a_2u_2u_3}{u_2+u_3}, & t>0,\\ \dfrac{\mathrm{d}u_3}{\mathrm{d}t} = \dfrac{m_2u_2u_3}{u_2+u_3} - b_2u_3, & t>0,\\ u_1(0)>0, \quad u_2(0)>0, \quad u_3(0)>0, \end{cases} \tag{2.2}$$

where $a_i = m_i/(\eta_i c_i r)$ $(i = 1, 2)$, can be regarded as the respective predation rate of u_2 and u_3.

From [13], it is easily shown that (2.2) has a unique positive steady state solution if and only if the following are satisfied:

$$m_2 > b_2, \quad A > 1 \quad \text{and} \quad 0 < a_1 < A/(A-1), \tag{2.3}$$

where

$$A \equiv m_1/(a_2(m_2 - b_2)/m_2 + b_1).$$

Moreover, the unique positive steady state $(u_1, u_2, u_3) = (\tilde{u}_1, \tilde{u}_2, \tilde{u}_3)$ can be expressed as

$$\tilde{u}_1 = [a_1 + A(1 - a_1)]/A, \quad \tilde{u}_2 = (A-1)\tilde{u}_1, \quad \text{and} \quad \tilde{u}_3 = (m_2 - b_2)\tilde{u}_2/b_2.$$

We also note that $m_2 > b_2$ and $A > 1$ imply $m_1 > b_1$.

In [13], the authors dealt with (2.2). In particular, they obtained the extinction conditions of certain species and discussed the local asymptotical stability of $(\tilde{u}_1, \tilde{u}_2, \tilde{u}_3)$ and various scenarios where distinct solutions can be attracted to the origin, the pest-free steady state, and the positive steady state $(\tilde{u}_1, \tilde{u}_2, \tilde{u}_3)$. For more detail, we refer the reader to [13]. From their results, the authors pointed out that this ODE system is very rich in dynamics.

To take into account the inhomogeneous distribution of the predators and the prey in different spatial locations within a fixed bounded domain Ω in $\mathbf{R}^N$ with smooth boundary at any given time, and the natural tendency of each species to diffuse to a smaller population concentration, instead of (2.2), we need to consider the following reaction-diffusion (PDE) system:

$$\begin{cases} u_{1t} - d_1\Delta u_1 = u_1(1 - u_1) - \dfrac{a_1 u_1 u_2}{u_1 + u_2} & \text{in } \Omega \times (0, \infty), \\ u_{2t} - d_2\Delta u_2 = \dfrac{m_1 u_1 u_2}{u_1 + u_2} - b_1 u_2 - \dfrac{a_2 u_2 u_3}{u_2 + u_3} & \text{in } \Omega \times (0, \infty), \\ u_{3t} - d_3\Delta u_3 = \dfrac{m_2 u_2 u_3}{u_2 + u_3} - b_2 u_3 & \text{in } \Omega \times (0, \infty), \\ \partial_\nu u_i = 0, \quad i = 1,\ 2,\ 3, & \text{on } \partial\Omega \times (0, \infty), \\ u_i(x, 0) = u_{i0}(x) \geq 0, \not\equiv 0,\ i = 1, 2, 3, & \text{in } \Omega. \end{cases} \tag{2.4}$$

Here ν is the outward unit normal vector on the boundary $\partial\Omega$ and $\partial_\nu = \partial/\partial\nu$, and d_i $(i = 1, 2, 3)$ are called the diffusion coefficients of the corresponding species u_i and hence are assumed to be positive constants. The initial data u_{i0} $(i = 1, 2, 3)$ are continuous functions, and the homogeneous Neumann boundary condition means that model (2.4) is self-contained and has no population flux across the boundary $\partial\Omega$.

In our work here, we are mainly concerned with the effect of diffusion on stationary pattern generated by (2.4). Hence, this leads us to study the steady state problem of

(2.4), which satisfies

$$
(2.5)\qquad \begin{cases} -d_1\Delta u_1 = u_1(1-u_1) - \dfrac{a_1u_1u_2}{u_1+u_2} & \text{in } \Omega,\\ -d_2\Delta u_2 = \dfrac{m_1u_1u_2}{u_1+u_2} - b_1u_2 - \dfrac{a_2u_2u_3}{u_2+u_3} & \text{in } \Omega,\\ -d_3\Delta u_3 = \dfrac{m_2u_2u_3}{u_2+u_3} - b_2u_3 & \text{in } \Omega,\\ \partial_\nu u_i = 0,\ i=1,2,3, & \text{on } \partial\Omega. \end{cases}
$$

It is evident that only nonnegative solutions of (2.5) are of real interest. The positive solution (u_1, u_2, u_3) of (2.5) to be mentioned throughout this paper always refers to a classical solution with $u_i > 0$ $(i = 1, 2, 3)$ on $\bar{\Omega}$. It should also be noted that the well-known maximum principle ensures that a nonnegative classical solution of (2.5) with $u_i \not\equiv 0$ $(i = 1, 2, 3)$ must be a positive one.

For (2.4) and the steady state problem (2.5), we will mainly concentrate on the coexistence of the three species and consider the case of $a_1 < 1$. In particular, some results for the existence and nonexistence of nonconstant positive solutions to (2.5) are derived. In establishing the existence of nonconstant positive solutions, due to the lack of variational structure for (2.5), our mathematical tool is the topology degree theory incorporated with the calculation of the fixed point index.

3. Persistence and stability. For simplicity of presentation, we introduce some notation. Throughout this section, let

$$
\mathbf{u}(t) = (u_1(t), u_2(t), u_3(t))^T \quad \text{and} \quad \mathbf{u}(x,t) = (u_1(x,t), u_2(x,t), u_3(x,t))^T
$$

be the respective solutions of (2.2) and (2.4). Denote $\mathbf{u} = (u_1, u_2, u_3)^T$, $\tilde{\mathbf{u}} = (\tilde{u}_1, \tilde{u}_2, \tilde{u}_3)^T$. From classical theories of ODEs and parabolic equations, $\mathbf{u}(t)$ and $\mathbf{u}(x,t)$ exist globally and are positive; namely, $u_i(t),\ u_i(x,t) > 0$ $(i = 1, 2, 3)$ for all $t > 0$ and $x \in \bar{\Omega}$.

First we state some simple facts about the asymptotical behavior of solutions to (2.4). The proof is similar to that of Theorem 2.5 in [28] and so is omitted here.

PROPOSITION 3.1. *The solution* $(u_1(x,t), u_2(x,t), u_3(x,t))$ *of* (2.4) *satisfies the following:*

(i) *If* $m_1 \le b_1$, *then* $(u_2(x,t), u_3(x,t)) \to (0,0)$ *uniformly on* $\bar{\Omega}$ *as* $t \to \infty$.

(ii) *If* $m_2 \le b_2$, *then* $u_3(x,t) \to 0$ *uniformly on* $\bar{\Omega}$ *as* $t \to \infty$.

(iii) *If* $a_1 \le 1$ *and* $m_1 \le b_1$, *then* $u_1(x,t) \to 1$ *and* $(u_2(x,t), u_3(x,t)) \to (0,0)$ *uniformly on* $\bar{\Omega}$ *as* $t \to \infty$. *As a consequence, if* $m_1 \le b_1$ *or* $m_2 \le b_2$, *problem* (2.5) *has no positive solutions.*

As shown in Proposition 3.1, if $m_1 \le b_1$ or $m_2 \le b_2$, the two predators or the top predator will become extinct, respectively. Moreover, if $a_1 \le 1$ and $m_1 \le b_1$, then only the plant will exist eventually.

In this paper, since our main goal is to analyze the coexistence of the three species, from now on, unless otherwise specified, it is always assumed that $(\tilde{u}_1, \tilde{u}_2, \tilde{u}_3)$ exists, which implies that $m_1 > b_1$ and $m_2 > b_2$ as indicated in section 2.

We have the following basic persistence property of the solutions $\mathbf{u}(t)$ and $\mathbf{u}(x,t)$, which shows that the three species always coexist at any time and any location of the habitat domain, no matter how fast or slowly they diffuse, under certain conditions on parameters. This result is even new for the ODE system (2.2).

PROPOSITION 3.2. *Assume that* $a_1 < 1$, $a_2 + b_1 < m_1$ *hold. Then, for any* $0 < \varepsilon \ll 1$, *there exists* $T \gg 1$ *such that* $\mathbf{u}(t)$ *and* $\mathbf{u}(x,t)$ *satisfy*

$$K - \varepsilon < u_1(t), \quad u_1(x,t) < 1 + \varepsilon,$$
$$\frac{K(m_1 - (a_2 + b_1))}{a_2 + b_1} - \varepsilon < u_2(t), \quad u_2(x,t) < \frac{m_1 - b_1}{b_1} + \varepsilon,$$
$$\frac{K(m_1 - (a_2 + b_1))(m_2 - b_2)}{(a_2 + b_1)b_2} - \varepsilon < u_3(t), \quad u_3(x,t) < \frac{(m_1 - b_1)(m_2 - b_2)}{b_1 b_2} + \varepsilon$$

for all $x \in \bar{\Omega}$ *and* $t > T$. *Here,* K *is given by*

$$K = \frac{1}{2}\left\{ 2 - \frac{m_1}{b_1} + \sqrt{\Big(2 - \frac{m_1}{b_1}\Big)^2 + 4(1 - a_1)\Big(\frac{m_1}{b_1} - 1\Big)} \ \right\}.$$

Proof. The proof is based on comparison principles. We first prove that the estimates hold for $\mathbf{u}(t)$. For $0 < \varepsilon \ll 1$ and $t \gg 1$, from the first equation in (2.2) it is clear that $u_1(t) < 1 + \varepsilon$ by the comparison principle for ODEs.

In the following, we always consider that $0 < \varepsilon \ll 1$ and $t \geq T \gg 1$, and the values of ε and T may be different from line to line. Since $u_2(t)$ satisfies

$$u_2'(t) < \frac{(m_1 - b_1)(1 + \varepsilon) - b_1 u_2}{1 + \varepsilon + u_2} u_2,$$

by the comparison principle for ODEs again, we have that

$$u_2(t) < \frac{(m_1 - b_1)(1 + \varepsilon)}{b_1} + \varepsilon = \frac{m_1 - b_1}{b_1} + \frac{m_1}{b_1}\varepsilon.$$

Thus we can assume the following holds:

$$u_2(t) < \frac{m_1 - b_1}{b_1} + \varepsilon. \tag{3.1}$$

Combining (3.1) and the first equation in (2.2), we deduce that

$$u_1'(t) > \frac{-u_1^2 + (2 - m_1/b_1 - \varepsilon)u_1 + (1 - a_1)[(m_1 - b_1)/b_1 + \varepsilon]}{(m_1 - b_1)/b_1 + \varepsilon + u_1} u_1.$$

Therefore

$$u_1(t) > \frac{1}{2}\left\{ 2 - \frac{m_1}{b_1} - \varepsilon + \sqrt{\Big(2 - \frac{m_1}{b_1} - \varepsilon\Big)^2 + 4(1 - a_1)\Big(\frac{m_1}{b_1} - 1 + \varepsilon\Big)} \ \right\} \tag{3.2}$$
$$-\varepsilon > K - \varepsilon.$$

Similarly, applying (3.2) to the second equation in (2.2), we obtain

$$u_2(t) > \frac{K(m_1 - (a_2 + b_1))}{a_2 + b_1} - \varepsilon. \tag{3.3}$$

Together with (3.1) and (3.3), the third equation in (2.2) results in

$$\frac{K[m_1 - (a_2 + b_1)](m_2 - b_2)}{(a_2 + b_1)b_2} - \varepsilon < u_3(t) < \frac{(m_1 - b_1)(m_2 - b_2)}{b_1 b_2} + \varepsilon. \tag{3.4}$$

To sum up, (3.1)–(3.4) deduce our result for $\mathbf{u}(t)$. In a similar manner, by the comparison principle for parabolic equations, one can establish the desired estimates for $\mathbf{u}(x,t)$. □

In particular Proposition 3.2 and the maximum principle imply a priori upper and lower bounds for the positive solutions of (2.5), which will play crucial roles in the later sections. To prove that we recall the following maximum principle (for example, Lemma 2.1 in [21]).

LEMMA 3.1. *Suppose that* $g \in C(\bar{\Omega} \times \mathbf{R})$.

(i) *Assume that* $w \in C^2(\Omega) \cap C^1(\bar{\Omega})$ *and satisfies*

$$\Delta w(x) + g(x, w(x)) \geq 0 \quad \text{in } \Omega, \quad \partial_\nu w \leq 0 \quad \text{on } \partial\Omega.$$

If $w(x_0) = \max_{\bar{\Omega}} w$, *then* $g(x_0, w(x_0)) \geq 0$.

(ii) *Assume that* $w \in C^2(\Omega) \cap C^1(\bar{\Omega})$ *and satisfies*

$$\Delta w(x) + g(x, w(x)) \leq 0 \quad \text{in } \Omega, \quad \partial_\nu w \geq 0 \quad \text{on } \partial\Omega.$$

If $w(x_0) = \min_{\bar{\Omega}} w$, *then* $g(x_0, w(x_0)) \leq 0$.

Now we have the following a priori estimates for steady state solutions.

THEOREM 3.1. *Assume that* $a_1 < 1$ *and* $a_2 + b_1 < m_1$ *hold. Let* K *be defined as in Proposition* 3.2. *Then any positive solution* (u_1, u_2, u_3) *of* (2.5) *satisfies the following: for all* $x \in \bar{\Omega}$,

$$K < u_1(x) < 1,$$

$$\frac{K(m_1 - (a_2 + b_1))}{a_2 + b_1} < u_2(x) < \frac{m_1 - b_1}{b_1},$$

$$\frac{K(m_1 - (a_2 + b_1))(m_2 - b_2)}{(a_2 + b_1)b_2} < u_3(x) < \frac{(m_1 - b_1)(m_2 - b_2)}{b_1 b_2}.$$

Proof. From Proposition 3.2, stated results hold if strict inequalities are replaced by nonstrict inequalities. Thus we only need to show the strict inequalities. Let (u_1, u_2, u_3) be a positive solution of (2.5) and set

$$u_i(x_i) = \max_{\bar{\Omega}} u_i \quad \text{and} \quad u_i(y_i) = \min_{\bar{\Omega}} u_i, \quad i = 1, 2, 3.$$

Applying Lemma 3.1 to the first equation in (2.5), we find that

$$1 - u_1(x_1) - \frac{a_1 u_2(x_1)}{u_1(x_1) + u_2(x_1)} \geq 0.$$

Thus when $a_1 < 1$, it follows that $u_1(y_1) < 1$. Following the same order in the proof of Proposition 3.2, we can show that the stated results with strict inequalities hold. □

When the population persistence holds for the food chain, the constant steady state $\tilde{\mathbf{u}}$ is always in the attracting region given in Proposition 3.2 and Theorem 3.1. Next we discuss the stability of $\tilde{\mathbf{u}}$ with respect to (2.4). To this end, we need to collect some known facts from [13]. For sake of simplicity, we denote

$$\mathbf{G}(\mathbf{u}) = \begin{pmatrix} u_1(1 - u_1) - \dfrac{a_1 u_1 u_2}{u_1 + u_2} \\ \dfrac{m_1 u_1 u_2}{u_1 + u_2} - b_1 u_2 - \dfrac{a_2 u_2 u_3}{u_2 + u_3} \\ \dfrac{m_2 u_2 u_3}{u_2 + u_3} - b_2 u_3 \end{pmatrix} \quad \text{and} \quad \mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}}) = \begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{pmatrix},$$

where

$$\begin{cases} a_{11} = \tilde{u}_1\Big[-1+\dfrac{a_1\tilde{u}_2}{(\tilde{u}_1+\tilde{u}_2)^2}\Big], \quad a_{22} = \tilde{u}_2\Big[-\dfrac{m_1\tilde{u}_1}{(\tilde{u}_1+\tilde{u}_2)^2}+\dfrac{a_2\tilde{u}_3}{(\tilde{u}_2+\tilde{u}_3)^2}\Big], \\ a_{33} = -\dfrac{m_2\tilde{u}_2\tilde{u}_3}{(\tilde{u}_2+\tilde{u}_3)^2} < 0, \quad a_{12} = -\dfrac{a_1\tilde{u}_1^2}{(\tilde{u}_1+\tilde{u}_2)^2} < 0, \quad a_{21} = \dfrac{m_1\tilde{u}_2^2}{(\tilde{u}_1+\tilde{u}_2)^2} > 0, \\ a_{23} = -a_2\tilde{u}_2^2/(\tilde{u}_2+\tilde{u}_3)^2 < 0, \quad a_{32} = m_2\tilde{u}_3^2/(\tilde{u}_2+\tilde{u}_3)^2 > 0. \end{cases}$$

In Proposition 3.1 in [13], it was proved that if $a_{11} \le 0$ and $a_{22} \le 0$, $\tilde{\mathbf{u}}$ is locally asymptotically stable for (2.2). Indeed even with the presence of the diffusion, $\tilde{\mathbf{u}}$ is uniformly asymptotically stable for (2.4) under the same conditions. More precisely, we have the following theorem.

THEOREM 3.2. *Assume that $a_{11} \le 0$ and $a_{22} \le 0$ hold; then $\tilde{\mathbf{u}}$ is locally uniformly asymptotically stable for* (2.4) *in the sense of* [12]. *As a consequence*, (2.5) *has no nonconstant positive solution in a neighborhood of $\tilde{\mathbf{u}}$. Moreover, if $a_1 < 1$ and $a_2 + b_1 < m_1$, then $a_{11} < 0$ and $a_{22} < 0$; hence $\tilde{\mathbf{u}}$ is locally uniformly asymptotically stable.*

Proof. The proof of stability when $a_{11} \le 0$ and $a_{22} \le 0$ is similar to that of Theorem 2 in [36], and we omit the details here. We note that if $a_1 < 1$, then $a_{11} < 0$. Moreover, the inequality $a_{22} \le 0$ is equivalent to

$$m_1 \ge \Big(b_1 + a_2\frac{m_2-b_2}{m_2}\Big)^2 \Big/ \Big(b_1 + a_2\Big(\frac{m_2-b_2}{m_2}\Big)^2\Big). \tag{3.5}$$

It is also noted that

$$\Big(b_1 + a_2\frac{m_2-b_2}{m_2}\Big)^2 \Big/ \Big(b_1 + a_2\Big(\frac{m_2-b_2}{m_2}\Big)^2\Big) < a_2 + b_1.$$

Hence if $a_1 < 1$ and $a_2 + b_1 < m_1$, we have a_{11}, $a_{22} < 0$ by (3.5). ∎

Remark 3.1. Theorem 3.2 and the previous arguments show that no Turing instability or diffusion-driven instability phenomenon occurs when $a_{11} \le 0$ and $a_{22} \le 0$ hold. On the other hand, if we take $m_1 = (b_1 + a_2(m_2-b_2)/m_2)^2/(b_1 + a_2(m_2-b_2)^2/m_2^2) - o(b_1)$, and $b_1 \to 0$, a_2, b_2, m_2 are properly chosen and either $a_1 \to 1/2$ or $a_1 \to 1$, as in Proposition 3.1 in [13], together with some meticulous computations, the well-known Roth–Hurwitz criterion ensures that $\tilde{\mathbf{u}}$ is still stable for the ODE system (2.2). However, by fixing these parameters including d_1 and d_2, and then letting the diffusion d_3 be large enough, similar to the proof of Theorem 2 in [36], one can show that $\tilde{\mathbf{u}}$ is unstable with respect to the PDE system (2.4). Thus Turing instability could occur when the conditions of Theorem 3.2 are not satisfied.

From Theorem 3.2, $\tilde{\mathbf{u}}$ is locally uniformly asymptotically stable when $a_1 < 1$ and $a_2 + b_1 < m_1$. In this case it is unlikely that nonconstant positive solutions (stationary pattern) of (2.5) exist. Indeed with more restrictive conditions on the parameters, we can show the global stability of $\tilde{\mathbf{u}}$ for systems (2.2) and (2.4). Our result below is independent of the diffusion rates d_i; that is, the constant coexistence state $\tilde{\mathbf{u}}$ is globally asymptotically stable. Hence when the conditions on the parameters are satisfied, $\tilde{\mathbf{u}}$ is stabilized under arbitrary spatially inhomogeneous perturbation.

THEOREM 3.3. *Let K be defined as in Proposition* 3.2. *Assume that the following hold*:

(i) $a_1 < 1$ *and* $a_2 + b_1 < m_1$;
(ii) $a_1(A-1)/A < m_1K/(a_2+b_1)$;

(iii) $a_2b_2m_1(m_2-b_2)(a_2+b_1) < b_1m_2K[m_1-(a_2+b_1)][b_1m_2+a_2(m_2-b_2)]$.
Then the constant positive steady state $\tilde{\mathbf{u}}$ *is globally asymptotically stable for systems* (2.2) *and* (2.4) *for all initial nonnegative conditions which are not steady states. In particular*, (2.5) *has no nonconstant positive solution if conditions* (i)–(iii) *hold.*

Proof. We use Lyapunov functionals for the proof. First, we verify the result for system (2.2). For our purpose, we first recall the following basic Lyapunov functionals:

$$E(u_i) = u_i - \tilde{u}_i - \tilde{u}_i \ln \frac{u_i}{\tilde{u}_i}, \qquad i = 1,\ 2,\ 3.$$

Note that $E(u_i(t))$ are nonnegative, and $E(u_i(t)) = 0$ $(i = 1, 2, 3)$ if and only if $(u_1(t), u_2(t), u_3(t)) = (\tilde{u}_1, \tilde{u}_2, \tilde{u}_3)$. Hence, letting

$$E(t) = E(u_1(t)) + \frac{a_1\tilde{u}_1}{m_1\tilde{u}_2}E(u_2(t)) + \frac{a_1a_2\tilde{u}_1}{m_1m_2\tilde{u}_3}E(u_3(t)),$$

we have

$$\begin{aligned}\frac{dE}{dt} = {} & \left\{ -1 + \frac{a_1\tilde{u}_2}{(\tilde{u}_1+\tilde{u}_2)(u_1+u_2)} \right\}(u_1-\tilde{u}_1)^2 + \frac{a_1\tilde{u}_1}{m_1\tilde{u}_2}\left\{ -\frac{m_1\tilde{u}_1}{(\tilde{u}_1+\tilde{u}_2)(u_1+u_2)} \right. \\ & \left. + \frac{a_2\tilde{u}_3}{(\tilde{u}_2+\tilde{u}_3)(u_2+u_3)} \right\}(u_2-\tilde{u}_2)^2 - \frac{a_1a_2\tilde{u}_1\tilde{u}_2}{m_1\tilde{u}_3(\tilde{u}_2+\tilde{u}_3)(u_2+u_3)}(u_3-\tilde{u}_3)^2. \end{aligned} \tag{3.6}$$

Under our assumptions (i)–(iii), we can claim that for $t \gg 1$ the following hold:

$$\frac{a_1\tilde{u}_2}{(\tilde{u}_1+\tilde{u}_2)(u_1+u_2)} \le 1 \quad \text{and} \quad \frac{a_2\tilde{u}_3}{(\tilde{u}_2+\tilde{u}_3)(u_2+u_3)} \le \frac{m_1\tilde{u}_1}{(\tilde{u}_1+\tilde{u}_2)(u_1+u_2)}. \tag{3.7}$$

In fact, by Proposition 3.2, to satisfy (3.7), for $t \gg 1$ it is sufficient to require

$$\frac{a_1\tilde{u}_2}{(\tilde{u}_1+\tilde{u}_2)} < \frac{m_1K}{a_2+b_1} \quad \text{and} \quad \frac{a_2\tilde{u}_3}{(\tilde{u}_2+\tilde{u}_3)} \cdot \frac{m_1}{b_1} < \frac{m_1\tilde{u}_1}{(\tilde{u}_1+\tilde{u}_2)} \cdot \frac{Km_2(m_1-(a_2+b_1))}{(a_2+b_1)b_2}.$$

Therefore by the definition of $(\tilde{u}_1, \tilde{u}_2, \tilde{u}_3)$, we easily see that the above two inequalities are equivalent to assumptions (ii) and (iii), respectively. Thus (3.6) implies that $E'(t) < 0$ for $t \gg 1$. Now for $t \gg 1$, $E(t)$ is a Lyapunov functional for system (2.2); namely, for $t \gg 1$, $E'(t) < 0$ along trajectories and $E(t) > 0$ except at $\tilde{\mathbf{u}}$. Hence $\tilde{\mathbf{u}}$ is globally asymptotically stable for (2.2) following the well-known theorem of Lyapunov stability.

Based on the proof of Theorem 3.3, by Proposition 3.2, it is not hard to see that for $t \gg 1$

$$E^*(t) = \int_\Omega \left\{ E(u_1(x,t)) + \frac{a_1\tilde{u}_1}{m_1\tilde{u}_2}E(u_2(x,t)) + \frac{a_1a_2\tilde{u}_1}{m_1m_2\tilde{u}_3}E(u_3(x,t)) \right\} \mathrm{d}x$$

is a Lyapunov functional for system (2.4) and $\tilde{\mathbf{u}}$ is globally asymptotically stable for system (2.4) under our assumptions. ∎

Remark 3.2. Simple analysis shows that Theorem 3.3 holds if one of the following holds: (1) $a_1 \to 0$, $a_2 \to 0$; (2) $a_1 < 1$, m_1 is large and $m_2 \to b_2$; or (3) $a_1 < 1$, m_1 is large, and $a_2b_2(m_2-b_2)(a_2+b_1) < (1-a_1)b_1m_2[b_1m_2+a_2(m_2-b_2)]$. Indeed in case (1), K defined in Proposition 3.2 tends to 1 as $a_1 \to 0$, and the lower and upper bounds in Proposition 3.2 and Theorem 3.1 tend to the same value as $a_2 \to 0$.

This shows that the a priori estimate in Proposition 3.2 and Theorem 3.1 are sharp when a_1 and a_2 are small.

Results in this section have interesting and significant biological implications. Regarding the impact of the diffusion, all results in this section (Propositions 3.1 and 3.2 and Theorems 3.1, 3.2, and 3.3) are independent of diffusion coefficients d_i, $i = 1, 2, 3$. In these parameter ranges, diffusion usually enhances the stability of the constant steady states. Proposition 3.1 gives conditions of total extinction of all three species and conditions of the extinction of both middle and top predators. Comparison can be made with results in section 2 of [13], where the ODE system is studied in more detail.

When the constant coexistence steady state $\tilde{\mathbf{u}}$ exists, our main persistence and stability results are proved under the assumptions

$$a_1 < 1 \quad \text{and } a_2 + b_1 < m_1. \tag{3.8}$$

These conditions are evidently stronger than the conditions (2.3) under which $\tilde{\mathbf{u}}$ exists. But with (3.8) satisfied, persistence holds for the whole food chain, and all three species coexist regardless of initial conditions (see Proposition 3.2). The persistence question is even open for the same ODE system, and here we prove it for the more general reaction-diffusion system with no-flux boundary condition. This answers an open question raised in [13] (see discussion on p. 80). Moreover, under (3.8), $\tilde{\mathbf{u}}$ is also locally uniformly asymptotically stable with respect to (2.4), and under strong conditions in Theorem 3.3, $\tilde{\mathbf{u}}$ is globally asymptotically stable. For the ODE systems, these results complement those in [13] in which the main concern is successful biological control. Indeed our results show that under (3.8), biological control of the pest cannot be achieved.

4. Nonexistence of nonconstant positive solutions of (2.5). In Theorem 3.3, the global stability of the constant coexistence steady state implies the nonexistence of nonconstant positive solutions of (2.5) regardless of diffusions. Several nonexistence results of nonconstant positive solutions to (2.5) will be presented in this section, and in these results, the diffusion coefficients do play important roles. The mathematical techniques to be employed are the implicit function theorem method and the energy method, respectively. From now on, let $0 = \mu_0 < \mu_1 \le \mu_2 \le \cdots$ be the eigenvalues of the operator $-\Delta$ on Ω with the homogeneous Neumann boundary condition.

4.1. The energy method. In this subsection, we apply the energy method to establish some results on the nonexistence of nonconstant positive solutions of (2.5). For convenience, let us denote the constants a_i, b_i, m_i $(i = 1, 2)$ collectively by Λ.

THEOREM 4.1. *Assume that* $a_1 < 1$ *and* $a_2 + b_1 < m_1$.

(i) *There exists* $\hat{D}_{1,2} = \hat{D}_{1,2}(\Lambda)$ *which is independent of* d_3 *and* Ω, *such that* (2.5) *has no nonconstant positive solution provided that* $\min\{\mu_1 d_1,\ \mu_1 d_2\} \ge \hat{D}_{1,2}$.

(ii) *If, in addition,* $a_1(a_2 + b_1)^2(m_1 - b_1) \le (1 - a_1)^2 b_1 m_1^2$, *then there exists* $\hat{D}_2 = \hat{D}_2(\Lambda)$ *which is independent of* d_1, d_3, *and* Ω, *such that* (2.5) *has no nonconstant positive solution provided that* $\mu_1 d_2 \ge \hat{D}_2$.

(iii) *If, in addition,* $a_2(a_2 + b_1)^2 b_2 m_1(m_1 - b_1)(m_2 - b_2) \le (1 - a_1)^3 b_1^3 m_2^2(m_1 - (a_2 + b_1))^2$, *then there exists* $\hat{D}_{1,3} = \hat{D}_{1,3}(\Lambda)$ *which is independent of* d_2 *and* Ω, *such that* (2.5) *has no nonconstant positive solution provided that* $\min\{\mu_1 d_1,\ \mu_1 d_3\} \ge \hat{D}_{1,3}$.

Proof. Let (u_1, u_2, u_3) be a positive solution of (2.5) and let $\bar{g} = |\Omega|^{-1}\int_\Omega g\,\mathrm{d}x$. Then, multiplying the corresponding equation in (2.5) by $\frac{1}{u_i}(u_i - \bar{u}_i)$, $i = 1, 2, 3$, integrating over Ω, and adding the results, we get

$$\begin{aligned}
&\int_\Omega \left\{ \sum_{i=1}^{3} \frac{d_i \bar{u}_i |\nabla(u_i - \bar{u}_i)|^2}{u_i^2} \right\} \mathrm{d}x \\
&\quad = \int_\Omega \Bigg\{ (u_1 - \bar{u}_1)^2 \Big(-1 + \frac{a_1 \bar{u}_2}{(u_1 + u_2)(\bar{u}_1 + \bar{u}_2)} \Big) \\
&\qquad + (u_1 - \bar{u}_1)(u_2 - \bar{u}_2) \frac{-a_1 \bar{u}_1 + m_1 \bar{u}_2}{(u_1 + u_2)(\bar{u}_1 + \bar{u}_2)} \\
&\qquad + (u_2 - \bar{u}_2)^2 \Big(\frac{-m_1 \bar{u}_1}{(u_1 + u_2)(\bar{u}_1 + \bar{u}_2)} + \frac{a_2 \bar{u}_3}{(u_2 + u_3)(\bar{u}_2 + \bar{u}_3)} \Big) \\
&\qquad + (u_2 - \bar{u}_2)(u_3 - \bar{u}_3) \frac{-a_2 \bar{u}_2 + m_2 \bar{u}_3}{(u_2 + u_3)(\bar{u}_2 + \bar{u}_3)} - (u_3 - \bar{u}_3)^2 \frac{m_2 \bar{u}_2}{(u_2 + u_3)(\bar{u}_2 + \bar{u}_3)} \Bigg\} \mathrm{d}x. \qquad (4.1)
\end{aligned}$$

By Theorem 3.1 and the Young inequality, from (4.1) it follows that

$$\begin{aligned}
\int_\Omega \sum_{i=1}^{3} d_i |\nabla(u_i - \bar{u}_i)|^2 \,\mathrm{d}x \le C \int_\Omega \Bigg\{ & (u_1 - \bar{u}_1)^2 \left(-1 + \frac{a_1 \bar{u}_2}{(u_1 + u_2)(\bar{u}_1 + \bar{u}_2)} + \varepsilon \right) \\
& + C(\varepsilon)(u_2 - \bar{u}_2)^2 \\
& + (u_3 - \bar{u}_3)^2 \left(- \frac{m_2 \bar{u}_2}{(u_2 + u_3)(\bar{u}_2 + \bar{u}_3)} + \varepsilon \right) \Bigg\} \mathrm{d}x. \qquad (4.2)
\end{aligned}$$

Here, C depends only on Λ, and $C(\varepsilon)$ depends only on Λ and ε. By Theorem 3.1 again, we can choose $0 < \varepsilon \ll 1$ which depends only on Λ such that

$$- \frac{m_2 \bar{u}_2}{(u_2 + u_3)(\bar{u}_2 + \bar{u}_3)} + \varepsilon < 0.$$

Thus, with (4.2) and the Poincaré inequality,

$$\mu_1 \int_\Omega (g - \bar{g})^2 \,\mathrm{d}x \le \int_\Omega |\nabla(g - \bar{g})|^2 \,\mathrm{d}x,$$

we find that

$$\mu_1 \int_\Omega \sum_{i=1}^{3} d_i (u_i - \bar{u}_i)^2 \,\mathrm{d}x \le C(\varepsilon) \int_\Omega \sum_{i=1}^{2} (u_i - \bar{u}_i)^2 \,\mathrm{d}x.$$

By the above inequality, it is clear that there exists $\hat{D}_{1,2}$ depending only on Λ, such that when $\min\{\mu_1 d_1,\ \mu_1 d_2\} \ge \hat{D}_{1,2}$, $u_i \equiv \bar{u}_i = \text{constant}$, $i = 1, 2, 3$, which asserts our result (i).

If, in addition, we assume $a_1(a_2 + b_1)^2(m_1 - b_1) \le (1 - a_1)^2 b_1 m_1^2$, then Theorem 3.1 implies

$$-1 + \frac{a_1 \bar{u}_2}{(u_1 + u_2)(\bar{u}_1 + \bar{u}_2)} < 0.$$

Therefore, for $0 < \varepsilon \ll 1$ satisfying

$$-1 + \frac{a_1\bar{u}_2}{(u_1+u_2)(\bar{u}_1+\bar{u}_2)} + \varepsilon < 0 \quad \text{and} \quad -\frac{m_2\bar{u}_2}{(u_2+u_3)(\bar{u}_2+\bar{u}_3)} + \varepsilon < 0,$$

as before, (4.2) implies

$$\int_\Omega \sum_{i=1}^{3} d_i |\nabla(u_i - \bar{u}_i)|^2 \, \mathrm{d}x \le C(\varepsilon) \int_\Omega (u_2 - \bar{u}_2)^2 \, \mathrm{d}x. \tag{4.3}$$

Similar to arguments above, from (4.3) and the Poincaré inequality, there exists $\hat{D}_2 = \hat{D}_2(\Lambda)$ such that (2.5) has no nonconstant positive solution if $\mu_1 d_2 > \hat{D}_2$. Thus (ii) holds.

To prove (iii), as in the arguments above, it is enough to verify that

$$a_2\bar{u}_3(u_1+u_2)(\bar{u}_1+\bar{u}_2) < m_1\bar{u}_1(u_2+u_3)(\bar{u}_2+\bar{u}_3). \tag{4.4}$$

By Theorem 3.1 again, to ensure (4.4), it suffices to require that the third condition in (iii) holds. This completes our proof. ∎

THEOREM 4.2.

(i) *Let d_1^*, d_3^* be fixed positive constants satisfying $\mu_1 d_1^* > 1$ and $\mu_1 d_3^* > m_2 - b_2$. Then there exists a positive constant $D_2^* = D_2^*(d_1^*, d_3^*, \Lambda)$ such that (2.5) has no nonconstant positive solution provided that $\mu_1 d_2 \ge D_2^*$, $d_1 \ge d_1^*$, and $d_3 \ge d_3^*$.*

(ii) *Let d_2^* be a fixed positive constant satisfying $\mu_1 d_2^* > m_1 - b_1$. Then there exists a positive constant $D_{1,3}^* = D_{1,3}^*(d_2^*, \Lambda)$ such that (2.5) has no nonconstant positive solution provided that $\min\{\mu_1 d_1, \mu_1 d_3\} \ge D_{1,3}^*$ and $d_2 \ge d_2^*$.*

Proof. We prove only (i), and the verification of (ii) is similar. Suppose that (u_1, u_2, u_3) and $(\bar{u}_1, \bar{u}_2, \bar{u}_3)$ are the same as in the proof of Theorem 4.1. Multiplying the corresponding equation of (2.5) by $u_i - \bar{u}_i$, $i = 1, 2, 3$, the analysis similar to the proof of Theorem 4.1 deduces

$$\mu_1 \sum_{i=1}^{3} \int_\Omega d_i (u_i - \bar{u}_i)^2 \mathrm{d}x \le \int_\Omega \{(1+\varepsilon)(u_1 - \bar{u}_1)^2 + C(u_2 - \bar{u}_2)^2 + (m_2 - b_2 + \varepsilon)(u_3 - \bar{u}_3)^2\} \mathrm{d}x$$

for some positive constant $C = C(\Lambda, \varepsilon)$. Choose $\varepsilon > 0$ to be so small that $d_1\mu_1 \ge 1 + \varepsilon$, $d_3\mu_1 \ge m_2 - b_2 + \varepsilon$; then there exists D_2^* such that $(u_1, u_2, u_3) = (\bar{u}_1, \bar{u}_2, \bar{u}_3)$ must hold if $d_2 \ge D_2^*$, and so our conclusion holds. ∎

The results in this subsection demonstrate such a phenomenon: when all diffusion coefficients are large, no patterns exist. Here either d_1, d_3, or d_2 has a lower bound (see Theorem 4.2). If, in addition, the conditions (3.8) are satisfied, then the patterns do not exist even if only one or two diffusion coefficients are large. Such results for general reaction-diffusion systems appeared in [6], and our results here show more delicate dependence on the diffusion coefficients only for the food chain system (2.4) and (2.5).

4.2. The implicit function theorem method. In this subsection, we use the implicit function theorem to obtain some further results for the nonexistence of nonconstant positive solutions of (2.5). We will need the following a priori estimate.

THEOREM 4.3. *Let $a_1 < 1$ and let d be a fixed positive number. Assume that for any positive constants $\tilde{d}_2$ and $\tilde{d}_3$, the boundary value problem*

$$\begin{cases} -\tilde{d}_2\Delta w_2 = (m_1 - b_1)w_2 - \dfrac{a_2 w_2 w_3}{w_2 + w_3} & \text{in } \Omega, \\ -\tilde{d}_3\Delta w_3 = \dfrac{m_2 w_2 w_3}{w_2 + w_3} - b_2 w_3 & \text{in } \Omega, \\ \partial_\nu w_2 = \partial_\nu w_3 = 0 & \text{on } \partial\Omega \end{cases} \tag{4.5}$$

has no positive solution satisfying $|w_2|_\infty + |w_3|_\infty = 1$. Then there exist positive constants $C_1(\Lambda, \Omega, d)$ and $C_2(\Lambda, \Omega, d)$ such that any positive solution (u_1, u_2, u_3) of (2.5) satisfies

$$C_1(\Lambda, \Omega, d) \le u_i \le C_2(\Lambda, \Omega, d), \quad i = 1, 2, 3,$$

provided that d_1, d_2, $d_3 \ge d$.

Proof. Since $a_1 < 1$, from the proof of Theorem 3.1, we see that

$$1 - a_1 < u_1 < 1, \quad u_2 < (m_1 - b_1)/b_1, \quad \text{and } u_3 < (m_1 - b_1)(m_2 - b_2)/(b_1 b_2), \tag{4.6}$$

so $C_2(\Lambda, \Omega, d)$ has been found. Similarly to the proof of Theorem 3.4 in [28], from the second and third equations in (2.5), the desired $C_1(\Lambda, \Omega, d)$ can be obtained. □

The assumption that (4.5) has no positive solution is satisfied in some important parameter ranges.

LEMMA 4.1. *Problem (4.5) has no positive solution if one of the following holds*:

(i) $a_2 + b_1 \le m_1$; *or*

(ii) $a_2 + b_1 > m_1$ *and* $\sqrt{a_2 + m_2} < \sqrt{m_1 - b_1} + \sqrt{b_2}$.

In particular, if $a_1 < 1$ and either (i) *or* (ii) *holds, the a priori estimate in Theorem 4.3 holds.*

Proof. If condition (i) holds, our conclusion is derived from (ii) of Lemma 3.1; if condition (ii) is satisfied, the proof is the same as that of Corollary 3.5 in [28]. □

In this subsection, we will prove a result which considerably improves Theorem 4.2 if the estimates in Theorem 4.3 hold. We note that the conditions (i) and (ii) include (3.8); thus the results are along the same lines as those in the last subsection. To prove our result, we first prepare two lemmas.

LEMMA 4.2. *Assume that $f(u)$ is a continuous function in $[0, \infty)$ and for some positive constant a, $f(u) > 0$ in $(0, a)$ and $f(u) < 0$ in (a, ∞). Then the problem*

$$-\Delta u = u f(u) \quad \text{in } \Omega, \qquad \partial_\nu u = 0 \quad \text{on } \partial\Omega$$

has a unique positive solution $u(x) \equiv a$.

Proof. The above result is easily obtained by the direct application of Lemma 3.1. □

LEMMA 4.3. (i) *Assume that $a_1 < 1$ and that assumptions in Theorem 4.3 hold. Let (u_{1i}, u_{2i}, u_{3i}) be a sequence of positive solutions of (2.5) with $d_2 = d_{2i}$ and $d_{2i} \to \infty$ as $i \to \infty$. Then (u_{1i}, u_{2i}, u_{3i}) converges to $\tilde{\mathbf{u}}$ in $[C(\bar{\Omega})]^3$ as $i \to \infty$.*

(ii) *Assume that $a_1 < 1$ and that assumptions in Theorem 4.3 hold. Let (u_{1i}, u_{2i}, u_{3i}) be a sequence of positive solutions of (2.5) with $(d_1, d_3) = (d_{1i}, d_{3i})$ and $d_{1i}, d_{3i} \to \infty$ as $i \to \infty$. Then (u_{1i}, u_{2i}, u_{3i}) converges to $\tilde{\mathbf{u}}$ in $[C(\bar{\Omega})]^3$ as $i \to \infty$.*

Proof. We prove only (i), and (ii) can be proved similarly by using Theorem 3.1 and Lemma 3.1.

From Theorem 4.3, the sequence $\{(u_{1i}, u_{2i}, u_{3i})\}$ is bounded in $[C(\bar{\Omega})]^3$ with the bound independent of d_2. Then some standard arguments show that there is a subsequence of (u_{1i}, u_{2i}, u_{3i}) (still labelled by itself), such that $(u_{1i}, u_{2i}, u_{3i}) \to (u_1, u_2, u_3)$ in $[C(\bar{\Omega})]^3$ as $i \to \infty$. Furthermore, $u_2 \equiv c$, which is a positive constant; $u_1, u_3 > 0$ on $\bar{\Omega}$; and (u_1, c, u_3) solves

$$
\begin{cases}
-d_1\Delta u_1 = u_1(1-u_1) - \dfrac{a_1 c u_1}{u_1 + c} & \text{in } \Omega, \quad \partial_\nu u_1 = 0 \ \text{ on } \partial\Omega, \\
\displaystyle\int_\Omega \left\{ \frac{m_1 u_1}{u_1 + c} - b_1 - \frac{a_2 u_3}{c + u_3} \right\} \mathrm{d}x = 0, & \\
-d_3\Delta u_3 = \dfrac{c m_2 u_3}{c + u_3} - b_2 u_3 & \text{in } \Omega, \quad \partial_\nu u_3 = 0 \ \text{ on } \partial\Omega.
\end{cases} \tag{4.7}
$$

By Lemma 4.2, from the first and third equations in (4.7), we find that u_1 and u_3 are both constants:

$$
u_1 \equiv \frac{1}{2}\Big\{ 1 - c + \sqrt{(1-c)^2 + 4c(1-a_1)} \Big\} \ \text{ and } \ u_3 \equiv \frac{m_2 - b_2}{b_2} c. \tag{4.8}
$$

Substituting (4.8) into the second equation in (4.7), we find that $(u_1, c, u_3) = \tilde{\mathbf{u}}$. This verifies that the convergence holds for a subsequence of (u_{1i}, u_{2i}, u_{3i}). But the limit is a fixed point; thus the convergence holds for the whole sequence (u_{1i}, u_{2i}, u_{3i}). ∎

Now we state our main result in this subsection.

THEOREM 4.4. *Assume that $a_1 < 1$ and that assumptions in Theorem 4.3 hold.*

(i) *Let ϵ_1 be an arbitrary positive constant. Then there exists $D_2 = D_2(\epsilon_1, \Lambda, \Omega)$ such that (2.5) has no nonconstant positive solution provided that $\min\{d_1, d_3\} \geq \epsilon_1$ and $d_2 \geq D_2$.*

(ii) *Let ϵ_2 be an arbitrary positive constant. Then there exists $D_{1,3} = D_{1,3}(\epsilon_2, \Lambda, \Omega)$ such that (2.5) has no nonconstant positive solution provided that $d_2 \geq \epsilon_2$ and $\min\{d_1, d_3\} \geq D_{1,3}$.*

Proof. We first prove (i). By (i) of Theorem 4.2, for a fixed large constant $D_{1,3}$ depending only on Λ and Ω, there exists $D_2^* = D_2^*(\Lambda, \Omega)$ such that (2.1) has no positive nonconstant solution when $d_1, d_3 \geq D_{1,3}$ and $d_2 \geq D_2^*$. As a result, it suffices to consider the case $d_1, d_3 \in [\epsilon_1/2, D_{1,3}]$.

We make a decomposition: $u_2 = w_2 + \xi$ with $\int_\Omega w_2 = 0$ and $\xi \in \mathbf{R}^+$. We observe that finding the positive solution of (2.5) is equivalent to solving the following problem:

$$
\begin{cases}
d_1\Delta u_1 + u_1(1-u_1) - \dfrac{a_1 u_1 (w_2 + \xi)}{u_1 + w_2 + \xi} = 0 & \text{in } \Omega,\ \partial_\nu u_1 = 0 \text{ on } \partial\Omega, \\
\Delta w_2 + \rho \left\{ \dfrac{m_1 u_1 (w_2 + \xi)}{u_1 + w_2 + \xi} \right. & \\
\qquad \left. - b_1(w_2 + \xi) - \dfrac{a_2 (w_2 + \xi) u_3}{w_2 + \xi + u_3} \right\} = 0 & \text{in } \Omega,\ \partial_\nu w_2 = 0 \text{ on } \partial\Omega, \\
\displaystyle\int_\Omega \left\{ \frac{m_1 u_1 (w_2 + \xi)}{u_1 + w_2 + \xi} - b_1(w_2 + \xi) \right. & \\
\qquad \left. - \frac{a_2 (w_2 + \xi) u_3}{w_2 + \xi + u_3} \right\} \mathrm{d}x = 0, & \\
d_3\Delta u_3 + \dfrac{m_2 (w_2 + \xi) u_3}{w_2 + \xi + u_3} - b_2 u_3 = 0 & \text{in } \Omega,\ \partial_\nu u_3 = 0 \text{ on } \partial\Omega, \\
\xi > 0, u_1, u_3 > 0 & \text{in } \Omega,
\end{cases} \tag{4.9}
$$

where $\rho = d_2^{-1}$. Clearly, $(u_1, w_2, \xi, u_3) = (\tilde{u}_1,\, 0, \tilde{u}_2, \tilde{u}_3)$ is a solution of (4.9) for $\rho > 0$.

To prove our theorem, by the finite covering argument, it is sufficient to prove that, for any fixed $\tilde{d}_1,\, \tilde{d}_3 \in [\epsilon_1/2, D_{1,3}]$, there exists $\delta_0 > 0$ such that if $\rho \in (0, \delta_0)$, $(d_1, d_3) \in (\tilde{d}_1 - \delta_0, \tilde{d}_1 + \delta_0) \times (\tilde{d}_3 - \delta_0, \tilde{d}_3 + \delta_0)$, then $(\tilde{u}_1, 0, \tilde{u}_2, \tilde{u}_3)$ is the unique solution of (4.9). To this end, we define the following Banach spaces:

$$W_\nu^{2,2}(\Omega) = \{g \in W^{2,2}(\Omega) \,\big|\, \partial_\nu g = 0 \ \text{ on } \ \partial\Omega\}, \quad L_0^2(\Omega) = \left\{ g \in L^2(\Omega) \,\Big|\, \int_\Omega g \,\mathrm{d}x = 0 \right\},$$

and denote

$$F(d_1,\, d_3,\, \rho,\, u_1,\, w_2,\, \xi,\, u_3) = (f_1,\, f_2,\, f_3,\, f_4)(d_1,\, d_3,\, \rho,\, u_1,\, w_2,\, \xi,\, u_3)$$

with

$$f_1(d_1,\, d_3,\, \rho,\, u_1,\, w_2,\, \xi,\, u_3) = d_1 \Delta u_1 + u_1\,(1 - u_1) - \frac{a_1 u_1 (w_2 + \xi)}{u_1 + w_2 + \xi},$$

$$f_2(d_1,\, d_3,\, \rho,\, u_1,\, w_2,\, \xi,\, u_3) = \Delta w_2 + \rho \left\{ \frac{m_1 u_1 (w_2 + \xi)}{u_1 + w_2 + \xi} - b_1 (w_2 + \xi) - \frac{a_2 (w_2 + \xi) u_3}{w_2 + \xi + u_3} \right\},$$

$$f_3(d_1,\, d_3,\, \rho,\, u_1,\, w_2,\, \xi,\, u_3) = \int_\Omega \left\{ \frac{m_1 u_1 (w_2 + \xi)}{u_1 + w_2 + \xi} - b_1 (w_2 + \xi) - \frac{a_2 (w_2 + \xi) u_3}{w_2 + \xi + u_3} \right\} \mathrm{d}x,$$

$$f_4(d_1,\, d_3,\, \rho,\, u_1,\, w_2,\, \xi,\, u_3) = d_3 \Delta u_3 + \frac{m_2 (w_2 + \xi) u_3}{w_2 + \xi + u_3} - b_2 u_3.$$

Then

$$F : \mathbf{R}^+ \times \mathbf{R}^+ \times \mathbf{R}^+ \times W_\nu^{2,2}(\Omega) \times (L_0^2(\Omega) \cap W_\nu^{2,2}(\Omega)) \times \mathbf{R}^+ \times W_\nu^{2,2}(\Omega) \to L^2(\Omega) \times L_0^2(\Omega) \times \mathbf{R} \times L^2(\Omega)$$

is a well-defined mapping. It is clear that the solutions of (4.9) satisfy $F(d_1, d_3, \rho, u_1, w_2, \xi, u_3) = 0$. Moreover, (4.9) has a unique solution $(u_1,\, w_2,\, \xi,\, u_3) = (\tilde{u}_1,\, 0,\, \tilde{u}_2,\, \tilde{u}_3)$ when $\rho = 0$ and $(d_1,\, d_3) = (\tilde{d}_1,\, \tilde{d}_3)$ from the proof of (i) of Lemma 4.3. Obviously, F is a differentiable mapping, and its partial derivative with respect to the last four arguments is

$$\Psi \equiv D_{(u_1,\, w_2,\, \xi,\, u_3)} F(\tilde{d}_1,\, \tilde{d}_3,\, 0,\, \tilde{u}_1,\, 0,\, \tilde{u}_2,\, \tilde{u}_3),$$

$$\Psi : W_\nu^{2,2}(\Omega) \times (L_0^2(\Omega) \cap W_\nu^{2,2}(\Omega)) \times \mathbf{R} \times W_\nu^{2,2}(\Omega) \to L^2(\Omega) \times L_0^2(\Omega) \times \mathbf{R} \times L^2(\Omega)$$

with

$$\Psi(v_1, v_2, \tau, v_3) = \begin{pmatrix} \tilde{d}_1 \Delta v_1 + a_{11} v_1 + a_{12}(v_2 + \tau) \\ \Delta v_2 \\ \int_\Omega \{a_{21} v_1 + a_{22}(v_2 + \tau) + a_{23} v_3\} \,\mathrm{d}x \\ \tilde{d}_3 \Delta v_3 + a_{32}(v_2 + \tau) + a_{33} v_3 \end{pmatrix},$$

where a_{ij} are given in section 3.

We claim that Ψ is an isomorphism operator. Assume that $\Psi(v_1, v_2, \tau, v_3) = (0, 0, 0, 0)$; then $v_2 = 0$. Note that $a_1 < 1$ implies $a_{11} < 0$. Then from the equation of v_1, it follows that $v_1 \equiv -a_{12}\tau/a_{11}$. Similarly, $v_3 \equiv -a_{32}\tau/a_{33}$ since $a_{33} < 0$ and $\tau \in \mathbf{R}$. We substitute these results into the integral equations satisfied by (v_1, v_2, τ, v_3) and obtain that

$$\left(-\frac{a_{12}a_{21}}{a_{11}} + a_{22} - \frac{a_{23}a_{32}}{a_{33}}\right)\tau = 0.$$

This is equivalent to $\det\{\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})\}\tau = 0$, where

$$\det\{\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})\} = -(a_{12}a_{21}a_{33} + a_{11}a_{23}a_{32} - a_{11}a_{22}a_{33}) = -\frac{m_1 m_2 \tilde{u}_1^2 \tilde{u}_2^2 \tilde{u}_3}{(\tilde{u}_1 + \tilde{u}_2)^2(\tilde{u}_2 + \tilde{u}_3)^2} < 0$$

by some basic computations. Therefore $\tau = 0$, which implies that $(v_1, v_2, \tau, v_3) = (0, 0, 0, 0)$ and Ψ is injective. On the other hand, for a given $h_2 \in L_0^2(\Omega)$, the problem

$$-\Delta u_2 = h_2 \ \text{ in } \ \Omega, \qquad u \in L_0^2(\Omega) \cap W_\nu^{2,2}(\Omega)$$

has a unique solution. By using $\det\{\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})\} < 0$ again, one can also check that Ψ is also surjective. Consequently Ψ is an isomorphism.

By the implicit function theorem, there exist positive constants ρ_0 and δ_0 such that, for each $\rho \in [0, \rho_0]$ and $(d_1, d_3) \in (\tilde{d_1} - \delta_0, \tilde{d_1} + \delta_0) \times (\tilde{d_3} - \delta_0, \tilde{d_3} + \delta_0)$, $(\tilde{u}_1, 0, \tilde{u}_2, \tilde{u}_3)$ is the unique solution of $F(d_1, d_3, \rho, u_1, w_2, \xi, u_3) = 0$ in $B_{\delta_0}(\tilde{u}_1, 0, \tilde{u}_2, \tilde{u}_3)$, where $B_{\delta_0}(\tilde{u}_1, 0, \tilde{u}_2, \tilde{u}_3)$ is the ball in $W_\nu^{2,2}(\Omega) \times (L_0^2(\Omega) \cap W_\nu^{2,2}(\Omega)) \times \mathbf{R} \times W_\nu^{2,2}(\Omega)$ centered at $(\tilde{u}_1, 0, \tilde{u}_2, \tilde{u}_3)$ with radius δ_0. Taking smaller ρ_0 and δ_0 if necessary, we can conclude (i) by use of Lemma 4.3(i).

In a similar manner, (ii) can be proved. In fact, we write $u_i = w_i + \xi_i$ with $\int_\Omega w_i = 0$ and $\xi_i \in \mathbf{R}^+$ $(i = 1, 3)$ and construct analogous operator

$$F(d_2, \rho_1, \rho_3, w_1, \xi_1, u_2, w_3, \xi_3) = (f_1, f_2, f_3, f_4, f_5)(d_2, \rho_1, \rho_3, w_1, \xi_1, u_2, w_3, \xi_3)$$

with

$$f_1(d_2, \rho_1, \rho_3, w_1, \xi_1, u_2, w_3, \xi_3) = \Delta w_1 + \rho_1 \left\{(w_1 + \xi_1)(1 - w_1 - \xi_1) - \frac{a_1(w_1 + \xi_1)u_2}{w_1 + \xi_1 + u_2}\right\},$$

$$f_2(d_2, \rho_1, \rho_3, w_1, \xi_1, u_2, w_3, \xi_3) = \int_\Omega \left\{(w_1 + \xi_1)(1 - w_1 - \xi_1) - \frac{a_1(w_1 + \xi_1)u_2}{w_1 + \xi_1 + u_2}\right\} \mathrm{d}x,$$

$$f_3(d_2, \rho_1, \rho_3, w_1, \xi_1, u_2, w_3, \xi_3) = d_2\Delta u_2 + \frac{m_1(w_1 + \xi_1)u_2}{w_1 + \xi_1 + u_2} - b_1 u_2 - \frac{a_2(w_3 + \xi_3)u_2}{w_3 + \xi_3 + u_2},$$

$$f_4(d_2, \rho_1, \rho_3, w_1, \xi_1, u_2, w_3, \xi_3) = \Delta w_3 + \rho_3 \left\{\frac{m_2(w_3 + \xi_3)u_2}{w_3 + \xi_3 + u_2} - b_2(w_3 + \xi_3)\right\},$$

$$f_5(d_2, \rho_1, \rho_3, w_1, \xi_1, u_2, w_3, \xi_3) = \int_\Omega \left\{\frac{m_2(w_3 + \xi_3)u_2}{w_3 + \xi_3 + u_2} - b_2(w_3 + \xi_3)\right\} \mathrm{d}x,$$

where $\rho_i = d_i^{-1}$ (i=1, 3). For fixed $\tilde{d}_2 > 0$, we can verify that

$$D_{(w_1,\xi_1,u_2,w_3,\xi_3)}F(\tilde{d}_2, 0, 0, 0, \tilde{u}_1, \tilde{u}_2, 0, \tilde{u}_3) :$$
$$(L_0^2(\Omega) \cap W_\nu^{2,2}(\Omega)) \times \mathbf{R} \times W_\nu^{2,2}(\Omega) \times (L_0^2(\Omega) \cap W_\nu^{2,2}(\Omega))$$
$$\times \mathbf{R} \to L_0^2(\Omega) \times \mathbf{R} \times L^2(\Omega) \times L_0^2(\Omega) \times \mathbf{R}$$

is an isomorphism. As in the discussion of (i), by the implicit function theorem, Lemma 4.3(ii) and Theorem 4.2(ii), our result is obtained. The proof of Theorem 4.4 is complete. □

5. Existence of nonconstant positive solutions of (2.5). This section is devoted to the existence of nonconstant positive solutions of (2.5) for certain values of diffusion coefficients d_1 and d_3, respectively, while the other parameters are fixed. Our results will show that, if the parameters are properly chosen, both the general stationary pattern and more interesting Turing pattern can arise as a result of diffusion.

For our purposes, we start with some preliminary results. First we study the linearization of (2.5) at $\tilde{\mathbf{u}}$. We denote

$$\mathbf{X} = \{\mathbf{u} \in [C^2(\bar{\Omega})]^3 \mid \partial_\nu \mathbf{u} = 0 \ \text{ on } \partial\Omega\}$$

and

$$\mathbf{X}^+ = \{\mathbf{u} \in \mathbf{X} \mid u_i > 0 \ \text{ on } \ \bar{\Omega}, \ \ i = 1, 2, 3\},$$
$$B(C) = \{\mathbf{u} \in \mathbf{X} \mid C^{-1} < u_i < C \ \text{ on } \ \bar{\Omega}, \ \ i = 1, 2, 3\}, \ \ C > 0.$$

With the diffusion matrix $\mathcal{D} = \text{diag}(d_1, d_2, d_3)$, (2.5) can be written as

$$\begin{cases} -\mathcal{D}\Delta\mathbf{u} = \mathbf{G}(\mathbf{u}) & \text{in } \ \Omega, \\ \partial_\nu \mathbf{u} = 0 & \text{on } \ \partial\Omega, \end{cases} \tag{5.1}$$

and $\mathbf{u}$ is a positive solution to (5.1) if and only if

$$\mathbf{F}(\mathbf{u}) \equiv \mathbf{u} - (\mathbf{I} - \Delta)^{-1}\{\mathcal{D}^{-1}\mathbf{G}(\mathbf{u}) + \mathbf{u}\} = 0 \ \text{ for } \ \mathbf{u} \in \mathbf{X}^+,$$

where $(\mathbf{I} - \Delta)^{-1}$ is the inverse of $\mathbf{I} - \Delta$ in $\mathbf{X}$. As $\mathbf{F}(\cdot)$ is a compact perturbation of the identity operator, for any $B = B(C)$, the Leray–Schauder degree $\deg(\mathbf{F}(\cdot), 0, B)$ is well defined if $\mathbf{F}(\mathbf{u}) \neq 0$ on ∂B.

We also note that

$$D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}}) = \mathbf{I} - (\mathbf{I} - \Delta)^{-1}\{\mathcal{D}^{-1}\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}}) + \mathbf{I}\},$$

and recall that if $D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})$ is invertible, the index of $\mathbf{F}$ at $\tilde{\mathbf{u}}$ is defined as $\text{index}(\mathbf{F}(\cdot), \tilde{\mathbf{u}}) = (-1)^\gamma$, where γ is the multiplicity of negative eigenvalues of $D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}})$ [25, Theorem 2.8.1].

For the sake of convenience, we denote

$$H(d_1, d_2, d_3; \mu) \equiv \det\{\mu\mathbf{I} - \mathcal{D}^{-1}\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})\} = \frac{1}{d_1 d_2 d_3}\det\{\mu\mathcal{D} - \mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})\}, \tag{5.2}$$

By arguments similar to those in [28], it can be shown that the following proposition holds.

PROPOSITION 5.1. *Suppose that, for all* $n \geq 0$, *the matrix* $\mu_n \mathbf{I} - \mathcal{D}^{-1}\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})$ *is nonsingular. Then*

$$\text{index}(\mathbf{F}(\cdot),\, \tilde{\mathbf{u}}) = (-1)^{\gamma}, \quad \textit{where} \;\; \gamma = \sum_{n\geq 0,\, H(d_1,d_2,d_3;\mu_n)<0} \dim E(\mu_n).$$

To compute $\text{index}(\mathbf{F}(\cdot),\, \tilde{\mathbf{u}})$, we have to consider the sign of $H(d_1, d_2, d_3; \mu)$. Direct calculation gives

$$\begin{aligned} \det\{\mu\mathcal{D} - \mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})\} &= A_3(d_1,d_3)\mu^3 + A_2(d_1,d_3)\mu^2 + A_1(d_1,d_3)\mu \\ &\quad - \det\{\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})\} \equiv \mathcal{A}(d_1, d_3;\, \mu), \end{aligned} \tag{5.3}$$

with

$$\begin{cases} A_3(d_1,d_3) = d_1d_2d_3, \quad A_2(d_1,d_3) = -\{a_{33}d_1d_2 + (a_{11}d_2 + a_{22}d_1)d_3\}, \\ A_1(d_1,d_3) = a_{11}a_{33}d_2 + (a_{22}a_{33} - a_{23}a_{32})d_1 + (a_{11}a_{22} - a_{12}a_{21})d_3, \end{cases}$$

where a_{ij} are defined in section 3.

We first consider the dependence of $\mathcal{A}$ on d_1. Let $\tilde{\mu}_i(d_1;\, d_2, d_3)$, $i = 1, 2, 3$, be the three roots of $\mathcal{A}(d_1, d_3;\, \mu) = 0$ satisfying $\text{Re}\{\tilde{\mu}_1(d_1;\, d_2, d_3)\} \leq \text{Re}\{\tilde{\mu}_2(d_1;\, d_2, d_3)\} \leq \text{Re}\{\tilde{\mu}_3(d_1;\, d_2, d_3)\}$. Since $\det\{\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})\} < 0$ and $A_3(d_1, d_3) > 0$, one of $\tilde{\mu}_i(d_1;\, d_2, d_3)$ is real and negative, and the product of the other two is positive.

In addition, we have

$$\lim_{d_1\to\infty} \mathcal{A}(d_1, d_3;\, \mu)/d_1 = \mu[d_2d_3\mu^2 - (a_{33}d_2 + a_{22}d_3)\mu + a_{22}a_{33} - a_{23}a_{32}].$$

Note that $a_{22}a_{33} - a_{23}a_{32} > 0$. If $a_{22} > 0$ or the reverse inequality of (3.5),

$$m_1 < \Big(b_1 + a_2\frac{m_2 - b_2}{m_2}\Big)^2 \Big/ \Big(b_1 + a_2\Big(\frac{m_2 - b_2}{m_2}\Big)^2\Big), \tag{5.4}$$

holds, and the parameters d_2 and d_3 satisfy

$$a_{33}d_2 + a_{22}d_3 > 0, \quad \Delta_1 \equiv (a_{33}d_2 + a_{22}d_3)^2 - 4d_2d_3(a_{22}a_{33} - a_{23}a_{32}) > 0, \tag{5.5}$$

we can establish the following proposition.

PROPOSITION 5.2. *Assume that* (5.4) *holds and that* d_2 *and* d_3 *satisfy* (5.5). *Then there exists a positive constant* D_1^* *such that when* $d_1 \geq D_1^*$, *the three roots* $\tilde{\mu}_i(d_1;\, d_2,\, d_3)$, $i = 1, 2, 3$, *of* $\mathcal{A}(d_1, d_3;\, \mu) = 0$ *are all real and satisfy*

$$\begin{cases} \lim\limits_{d_1\to\infty} \tilde{\mu}_1(d_1;\, d_2, d_3) = 0, \\ \lim\limits_{d_1\to\infty} \tilde{\mu}_2(d_1;\, d_2, d_3) = \dfrac{1}{2d_2d_3}\Big\{a_{33}d_2 + a_{22}d_3 - \sqrt{\Delta_1}\Big\} \equiv \mu_2^*(d_2, d_3) > 0, \\ \lim\limits_{d_1\to\infty} \tilde{\mu}_3(d_1;\, d_2, d_3) = \dfrac{1}{2d_2d_3}\Big\{a_{33}d_2 + a_{22}d_3 + \sqrt{\Delta_1}\Big\} \equiv \mu_3^*(d_2, d_3) > 0. \end{cases} \tag{5.6}$$

Moreover, when $d_1 \geq D_1^*$,

$$\begin{cases} -\infty < \tilde{\mu}_1(d_1; d_2, d_3) < 0 < \tilde{\mu}_2(d_1; d_2, d_3) < \tilde{\mu}_3(d_1; d_2, d_3), \\ \mathcal{A}(d_1, d_3; \mu) < 0 \;\; \textit{if} \;\; \mu \in (-\infty, \tilde{\mu}_1(d_1; d_2, d_3)) \cup (\tilde{\mu}_2(d_1; d_2, d_3), \tilde{\mu}_3(d_1; d_2, d_3)), \\ \mathcal{A}(d_1, d_3;\, \mu) > 0 \;\; \textit{if} \;\; \mu \in (\tilde{\mu}_1(d_1; d_2, d_3), \tilde{\mu}_2(d_1; d_2, d_3)) \cup (\tilde{\mu}_3(d_1; d_2, d_3), \infty). \end{cases} \tag{5.7}$$

Similarly, we consider d_3 as the parameter, and d_1 and d_2 satisfy

$$a_{11}d_2 + a_{22}d_1 > 0, \quad \Delta_2 \equiv (a_{11}d_2 + a_{22}d_1)^2 - 4d_1d_2(a_{11}a_{22} - a_{12}a_{21}) > 0; \tag{5.8}$$

then we have the following proposition.

PROPOSITION 5.3. *Assume that* (5.4) *holds and that* d_1 *and* d_2 *satisfy* (5.8). *Then there exists a positive constant* D_3^* *such that when* $d_3 \geq D_3^*$, *the three roots* $\tilde{\mu}_1(d_3; d_1, d_2)$, $i = 1, 2, 3$, *of* $\mathcal{A}(d_1, d_3; \mu) = 0$ *are all real and satisfy*

$$\begin{cases} \lim_{d_3\to\infty} \tilde{\mu}_1(d_3;\, d_1, d_2) \leq 0, \\ \lim\limits_{d_3\to\infty} \tilde{\mu}_2(d_3;\, d_1, d_2) = \dfrac{1}{2d_1d_2}\Big\{a_{11}d_2 + a_{22}d_1 - \sqrt{\Delta_2}\Big\} \equiv \mu_2^*(d_1, d_2) \geq 0, \\ \lim\limits_{d_3\to\infty} \tilde{\mu}_3(d_3;\, d_1, d_2) = \dfrac{1}{2d_1d_2}\Big\{a_{11}d_2 + a_{22}d_1 + \sqrt{\Delta_2}\Big\} \equiv \mu_3^*(d_1, d_2) > 0. \end{cases}$$

Moreover, when $d_3 \geq D_3^*$,

$$\begin{cases} -\infty < \tilde{\mu}_1(d_3;\, d_1, d_2) < 0 < \tilde{\mu}_2(d_3;\, d_1, d_2) < \tilde{\mu}_3(d_3;\, d_1, d_2), \\ \mathcal{A}(d_1, d_3;\, \mu) < 0 \quad \textit{if} \quad \mu \in (-\infty,\, \tilde{\mu}_1(d_3;\, d_1, d_2)) \cup (\tilde{\mu}_2(d_3;\, d_1, d_2),\, \tilde{\mu}_3(d_3;\, d_1, d_2)), \\ \mathcal{A}(d_1, d_3;\, \mu) > 0 \quad \textit{if} \quad \mu \in (\tilde{\mu}_1(d_3;\, d_1, d_2),\, \tilde{\mu}_2(d_3;\, d_1, d_2)) \cup (\tilde{\mu}_3(d_3;\, d_1, d_2),\, \infty). \end{cases}$$

Remark 5.1. Simple computations show that $\mu_2^*(d_1, d_2) = 0$ if and only if $a_{11}a_{22} - a_{12}a_{21} \leq 0$.

In virtue of Theorems 4.3 and 4.2 and Propositions 5.1 and 5.2, the first result of the existence of nonconstant positive solutions of (2.5) can be stated as follows.

THEOREM 5.1. *Assume that* $a_1 < 1$, (5.4), (5.5), *and assumptions in Theorem* 4.3 *hold. If* $\mu_2^*(d_2, d_3) \in (\mu_i,\, \mu_{i+1})$ *and* $\mu_3^*(d_2, d_3) \in (\mu_j,\, \mu_{j+1})$ *for some* $j > i \geq 0$, *where* $\mu_2^*(d_2, d_3)$, $\mu_3^*(d_2, d_3)$ *are defined in Proposition* 5.2, *and the sum* $\sum_{n=i+1}^{j} \dim E(\mu_n)$ *is odd, then there exists a positive constant* $\tilde{D}_1$ *such that, if* $d_1 \geq \tilde{D}_1$, (2.5) *admits at least one nonconstant positive solution.*

Proof. By Proposition 5.2 and our assumptions, there exists a positive constant $\tilde{D}_1$, such that when $d_1 \geq \tilde{D}_1$, (5.7) holds and

$$\mu_i < \tilde{\mu}_2(d_1;\, d_2, d_3) < \mu_{i+1}, \quad \mu_j < \tilde{\mu}_3(d_1;\, d_2, d_3) < \mu_{j+1}. \tag{5.9}$$

According to Theorem 4.2, for $\hat{d}_1$ and $\hat{d}_3$ satisfying $\mu_1\hat{d}_1 > 1$, $\mu_1\hat{d}_3 > m_2 - b_2$, there exists a large $\hat{d}_2$ such that (2.5) has no constant positive solutions when $d_1 \geq \hat{d}_1$, $\mu_1 d_2 \geq \hat{d}_2$, and $d_3 \geq \hat{d}_3$. In addition, since $\det\{\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})\} < 0$ and $\lim_{n\to\infty} \mu_n = \infty$, from (5.3), we can further choose $\hat{d}_1$, $\hat{d}_2$, and $\hat{d}_3$ to be so large that

$$H(\hat{d}_1,\, \hat{d}_2,\, \hat{d}_3;\, \mu_n) > 0 \quad \text{for all} \quad n \geq 0. \tag{5.10}$$

Now we show that for any $d_1 \geq \tilde{D}_1$, (2.5) has at least one nonconstant positive solution. The proof, which is accomplished by a contradiction argument, is based on the homotopy invariance of the topological degree. Suppose on the contrary that the assertion is not true for some $d_1 = \tilde{d}_1 \geq \tilde{D}_1$.

We fix $d_1 = \tilde{d}_1$. Let $d_i(t) = td_i + (1 - t)\hat{d}_i$, $i = 1, 2, 3$, and define $\mathcal{D}(t) = \operatorname{diag}(d_1(t), d_2(t), d_3(t))$. Now we consider the problem

$$\begin{cases} -\mathcal{D}(t)\Delta\mathbf{u} = \mathbf{G}(\mathbf{u}) & \text{in} \quad \Omega, \\ \partial_\nu \mathbf{u} = 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{5.11}$$

Then $\mathbf{u}$ is a positive solution of (2.5) if and only if it is a positive solution of (5.11) for $t = 1$. It is obvious that $\tilde{\mathbf{u}}$ is the unique constant positive solution of (5.11) for any $0 \le t \le 1$. For any $0 \le t \le 1$, $\mathbf{u}$ is a positive solution of (5.11) if and only if

$$\mathbf{F}(t;\, \mathbf{u}) \equiv \mathbf{u} - (\mathbf{I} - \Delta)^{-1}\Big\{\mathcal{D}^{-1}(t)\mathbf{G}(\mathbf{u}) + \mathbf{u}\Big\} = 0 \quad \text{for} \quad \mathbf{u} \in \mathbf{X}^{+}.$$

Clearly, $\mathbf{F}(1;\, \mathbf{u}) = \mathbf{F}(\mathbf{u})$. Theorem 4.2 shows that the only positive solution of $\mathbf{F}(0;\, \mathbf{u}) = 0$ is $\tilde{\mathbf{u}}$. From direct calculation,

$$D_{\mathbf{u}}\mathbf{F}(t;\, \tilde{\mathbf{u}}) = \mathbf{I} - (\mathbf{I} - \Delta)^{-1}\{\mathcal{D}^{-1}(t)\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}}) + \mathbf{I}\}.$$

In particular,

$$\begin{aligned} D_{\mathbf{u}}\mathbf{F}(0;\, \tilde{\mathbf{u}}) &= \mathbf{I} - (\mathbf{I} - \Delta)^{-1}\{\hat{\mathcal{D}}^{-1}\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}}) + \mathbf{I}\}, \\ D_{\mathbf{u}}\mathbf{F}(1;\, \tilde{\mathbf{u}}) &= \mathbf{I} - (\mathbf{I} - \Delta)^{-1}\{\mathcal{D}^{-1}\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}}) + \mathbf{I}\} = D_{\mathbf{u}}\mathbf{F}(\tilde{\mathbf{u}}), \end{aligned}$$

where $\hat{\mathcal{D}} = \mathrm{diag}(\hat{d}_1,\, \hat{d}_2,\, \hat{d}_3)$. From (5.2) and (5.3) we see that

$$H(d_1, d_2, d_3;\, \mu) = \frac{1}{d_1 d_2 d_3}\mathcal{A}(d_1, d_3;\, \mu). \tag{5.12}$$

In view of (5.7) and (5.9), it follows from (5.12) that

$$\begin{cases} H(d_1, d_2, d_3;\, \mu_0) = H(0) > 0, \\ H(d_1, d_2, d_3;\, \mu_n) < 0, \quad i+1 \le n \le j, \\ H(d_1, d_2, d_3;\, \mu_n) > 0, \quad 1 \le n \le i \ \text{ and } \ n \ge j+1. \end{cases}$$

Therefore, zero is not an eigenvalue of the matrix $\mu_i\mathbf{I} - \mathcal{D}^{-1}\mathbf{G}_{\mathbf{u}}(\tilde{\mathbf{u}})$ for all $n \ge 0$, and

$$\sum_{n\ge 0, H(d_1,d_2,d_3;\,\mu_n)<0} \dim E(\mu_n) = \sum_{n=i+1}^{j} \dim E(\mu_n) = \text{an odd number}.$$

Then Proposition 5.1 shows that

$$\mathrm{index}(\mathbf{F}(1;\, \cdot),\, \tilde{\mathbf{u}}) = (-1)^{\gamma} = -1. \tag{5.13}$$

On the other hand, by (5.10) and Proposition 5.1 again, we obtain that

$$\mathrm{index}(\mathbf{F}(0;\, \cdot),\, \tilde{\mathbf{u}}) = (-1)^{0} = 1. \tag{5.14}$$

In view of $\tilde{d}_1 > \tilde{D}_1$, by Theorem 4.3, there exists a positive constant $C = C(\tilde{D}_1, d_2, d_3, \hat{d}_1, \hat{d}_2, \hat{d}_3, \Lambda)$ such that, for all $0 \le t \le 1$, the positive solutions of (5.11) satisfy $1/C < u_1,\, u_2,\, u_3 < C$. Therefore, $\mathbf{F}(t;\, \mathbf{u}) \ne 0$ on $\partial B(C)$ for all $0 \le t \le 1$. By the homotopy invariance of the topological degree,

$$\deg\left(\mathbf{F}(1;\, \cdot),\, 0,\, B(C)\right) = \deg\left(\mathbf{F}(0;\, \cdot),\, 0,\, B(C)\right). \tag{5.15}$$

Moreover, under our assumptions, the only positive solution of both $\mathbf{F}(1;\mathbf{u}) = 0$ and $\mathbf{F}(0;\, \mathbf{u}) = 0$ in $B(C)$ is $\tilde{\mathbf{u}}$, and hence, by (5.13) and (5.14),

$$\deg\left(\mathbf{F}(0;\, \cdot),\, 0,\, B(C)\right) = \mathrm{index}(\mathbf{F}(0;\, \cdot),\, \tilde{\mathbf{u}}) = 1$$

and

$$\deg\left(\mathbf{F}(1;\, \cdot),\, 0,\, B(C)\right) = \mathrm{index}(\mathbf{F}(1;\, \cdot),\, \tilde{\mathbf{u}}) = -1.$$

This contradicts (5.15), and the proof is complete. □

Remark 5.2. When d_3 is fixed, we note that $\lim_{d_2\to 0}\mu_2^*(d_2,d_3) = (a_{22}a_{33} - a_{23}a_{32})/(a_{22}d_3)$, $\lim_{d_2\to 0}\mu_3^*(d_2,d_3)=\infty$, and (5.5) is automatically fulfilled for small d_2. Therefore, if for all $i=0,1,2,\dots,\mu_i$ are simple and $(a_{22}a_{33}-a_{23}a_{32})/(a_{22}d_3)\neq\mu_i$, by Theorem 5.1, when $a_1<1$ and the assertion of Theorem 4.3 hold, there exist two sequences of intervals $\{(\theta_n^1,\theta_n^2)\}_{n=1}^\infty$ and $\{(\Theta_n^1,\Theta_n^2)\}_{n=1}^\infty$ satisfying $\theta_{n+1}^2<\theta_n^1$, $\Theta_n^2<\Theta_{n+1}^1$, and $\theta_n^1,\ \theta_n^2\to 0^+$ while $\Theta_n^1,\ \Theta_n^2\to\infty$ as $n\to\infty$ such that (2.5) admits at least one nonconstant positive solution for all $d_1\in(\Theta_n^1,\Theta_n^2)$ and $d_2\in(\theta_n^1,\theta_n^2)$, $n=1,2,3,\dots$. Recall that Theorem 4.3 holds when condition (i) or (ii) in Lemma 4.1 holds. When (i) (same as (3.8)) holds, (5.4) is not satisfied. But when (ii) in Lemma 4.1 holds, conditions in Theorem 5.1 can be satisfied.

Similarly, let us consider the case of large d_3. By Proposition 5.3 and Remark 5.1, we have the following theorem.

THEOREM 5.2. *Assume that* $a_1<1$, (5.4), (5.8), *and Theorem* 4.3 *hold.*

(i) *If* $a_{11}a_{22}-a_{12}a_{21}>0$, *then* $\mu_2^*(d_1,d_2)\in(\mu_i,\mu_{i+1})$, $\mu_3^*(d_1,d_2)\in(\mu_j,\mu_{j+1})$ *for some* $j>i\geq 0$, *and* $\sum_{n=i+1}^{j}\dim E(\mu_n)$ *is odd.*

(ii) *If* $a_{11}a_{22}-a_{12}a_{21}\leq 0$, $\mu_3^*(d_1,d_2)\in(\mu_j,\mu_{j+1})$ *for some* $j>0$, *and* $\sum_{n=1}^{j}\dim E(\mu_n)$ *is odd, where* $\mu_2^*(d_1,d_2)$, $\mu_3^*(d_1,d_2)$ *are defined in Proposition* 5.3, *then there exists a positive constant* $\tilde{D}_3$ *such that, if* $d_3\geq\tilde{D}_3$, (2.5) *admits at least one nonconstant positive solution.*

Remark 5.3. By Proposition 5.3, regardless of the sign of $a_{11}a_{22}-a_{12}a_{21}$, we have a conclusion similar to that in Remark 5.2. In addition, we mention that the sign of $a_{11}a_{22}-a_{12}a_{21}$ is indefinite when $a_1<1$, $a_{22}>0$, $\tilde{\mathbf{u}}$ exists, and the assertion of Theorem 4.3 holds (note that, if $\sqrt{a_2+m_2}<\sqrt{m_1-b_1}+\sqrt{b_2}$, then Theorem 4.3 is true by Lemma 4.1). The detailed analysis on this claim is left to the appendix.

Remark 5.4. Fix d_1 and d_2; by Remark 3.1, if $m_1=(b_1+a_2(m_2-b_2)/m_2)^2/(b_1+a_2(m_2-b_2)^2/m_2^2)-o(b_1)$, and $b_1\to 0$, a_2, b_2, m_2 are properly chosen and either $a_1\to 1/2$ or $a_1\to 1$, and d_3 is sufficiently large, Turing instability actually happens. Furthermore, combined with the analysis of the appendix, Proposition A.2(i) also holds for such chosen parameters. With proper choices of d_1 and d_2, we can find certain parameter ranges guaranteeing the existence of both Turing instability and the nonconstant positive solution to (2.5) by Theorem 5.2(i). As a consequence, Turing patterns exist for these parameter ranges.

6. Conclusions. In this paper, we analyze a reaction-diffusion food chain model with ratio-dependent functional response. We are mainly concerned with the coexistence of the three species and focus on the case of a weak predation rate for the pest species (i.e., $a_1<1$). In particular, the existence and nonexistence of nonconstant positive steady states have been established. The existence results provide a theoretical support for pattern formation caused by diffusion.

We summarize our investigation here and hope to reveal some interesting phenomena of pattern formation in population ecology. We always assume the existence of a constant coexistence. The main results of sections 3 and 4 show that this constant coexistence steady state $\tilde{\mathbf{u}}$ is the only steady state if (a) both of the predation rates a_1 and a_2 are small; (b) a_1 is small while a_2 is suitably chosen, and either the pest or the other two species diffuse quickly (Theorem 4.4). In the former case, we are also able to find a more restrictive parameter range so that $\tilde{\mathbf{u}}$ is globally asymptotically stable (Theorem 3.3). This can also be seen from a bifurcation point of view. Here if we assume that a_1 is small, then stronger stability results of $\tilde{\mathbf{u}}$ can be proved when a_2 is smaller. When a_2 is close to zero, then $\tilde{\mathbf{u}}$ is globally asymptotically stable; when a_2

increases, $\tilde{\mathbf{u}}$ is still locally asymptotically stable but may not be globally asymptotically stable, and it is still the only steady state; and when a_2 further increases, $\tilde{\mathbf{u}}$ becomes unstable for both the ODE and the reaction-diffusion system, and nonconstant patterns exist in this case. In the latter case, the diffusion of the first or third species must be large enough (see Theorems 5.1 and 5.2). Thus for small a_1 and suitable a_2, the quick migration of the plant or top predator enhances the formation of spatial pattern for the system. In contrast, the quick migration of the pest or both the plant and top predators tends to prevent the system from generating pattern. It is well known that fast diffusion of all species in a biological system will not lead to spatial inhomogeneous patterns; see [6]. Our result shows the importance of the diffusion rate of the middle species in a food chain. The large diffusion rate of the pest (middle species) alone can lead to the nonexistence of spatial patterns, but if the pest diffusion rate is not large, then all other diffusion rates must be large to prevent the occurrence of patterns. On the other hand, a large diffusion rate of the top species or bottom species will help the generation of patterns. This demonstrates that, in an ecological model, different diffusions may play essentially different roles in developing spatial patterns. In addition, taking into account the close relationship between the time-dependent solutions to a reaction-diffusion system and the corresponding steady state solutions, to a great extent, the dynamical behaviors of (2.4) will be determined by the diffusions of the three species.

These conclusions can also be compared with those in [28] and [36]. In the absence of u_3, (2.4) becomes the prey-predator model studied by Pang and Wang in [28]. The results of the existence and nonexistence of nonconstant positive solutions there indicate that large d_2 contributes to the evolution of heterogeneousness for the dynamics, while large d_1 tends to increase the possibility of spatial uniform distribution. Therefore, combined with the our conclusions for (2.5), this suggests that the structure of solutions to the model in [28] will be significantly different due to the emergence of the top predator, which in turn leads to the qualitative change of the biological mechanism of the system. Such a phenomenon was also discussed by Lou, Martinez, and Ni for the classical Lotka–Volterra competition model in [20].

In [36], Wang investigated a three-species prey-predator model. In that model, the interaction between the lower and middle species is described by Holling II-type functional response (prey-dependent), while the functional response between the middle and top species is ratio-dependent (predator-dependent). It was proved that Turing pattern may appear if both d_1 and d_3 are large, but will not if d_2 is large. Therefore the results of the present paper and [36] show that the formation of Turing pattern in the biological models with the same degree of complexity depends on the choices of functional responses. In other words, the feeding strategy of predators may be one of the determining factors in producing Turing pattern. In a very recent work [2], Alonso, Bartumeus, and Catalan performed some numerical calculations indicating that predator-dependent models are sometimes capable of generating Turing pattern, while similar prey-dependent models are not. Hence our theoretical analysis for the food chain model rigorously confirms the outcome of computer simulation in [2].

Finally we point out that some of our mathematical techniques in sections 4 and 5 can be applied to deal with the prey-predator model proposed by Pang and Wang in [28] and derive some new a priori estimates for positive steady state solutions and nonexistence results for nonconstant positive steady state solutions.

Appendix A. In section 6, to prove the existence of nonconstant positive solutions to (2.5), we have made some hypotheses, namely, $a_1 < 1,\ a_{22} > 0$, $\tilde{\mathbf{u}}$ exists, and The-

orem 4.3 holds (in particular, $\sqrt{a_2+m_2} < \sqrt{m_1-b_1}+\sqrt{b_2}$ means that Theorem 4.3 is true). We list these conditions as follows:

$$\text{(A.1)}\quad \begin{cases} a_1 < 1, \quad m_2 > b_2, \quad A > 1 \Longleftrightarrow a_2(m_2-b_2)/m_2 + b_1 < m_1, \\ \sqrt{a_2+m_2} < \sqrt{m_1-b_1}+\sqrt{b_2} \Longleftrightarrow (\sqrt{a_2+m_2}-\sqrt{b_2}\,)^2 + b_1 < m_1, \\ a_{22} > 0 \Longleftrightarrow m_1 < (b_1 + a_2(m_2-b_2)/m_2)^2/(b_1 + a_2(m_2-b_2)^2/m_2^2). \end{cases}$$

In the following, we will verify the claim made in Remark 5.3, which says that $a_{11}a_{22} - a_{12}a_{21}$ is indefinite when (A.1) holds. First of all, by the definitions of a_{11}, a_{22}, a_{12}, a_{21}, the direct computations yield the following proposition.

PROPOSITION A.1. *Define*

$$\begin{aligned} Q \equiv {} & -(1-a_1)a_2b_2(m_2-b_2)A^3 + (1-a_1)m_1m_2^2A^2 \\ & + [(2a_1-1)m_1m_2^2 - a_1a_2b_2(m_2-b_2)]A - a_1m_1m_2^2; \end{aligned}$$

then $a_{11}a_{22} - a_{12}a_{21} > 0 \Longleftrightarrow Q > 0$. *Moreover, we note that*

(i) *as* $a_1 \to 0$, $Q \to (m_1m_2^2(A-1) - a_2b_2(m_2-b_2)A^2)A$;

(ii) *as* $a_1 \to 1/2$, $Q \to -\frac{1}{2}a_2b_2(m_2-b_2)A^3 - \frac{1}{2}a_2b_2(m_2-b_2)A + \frac{1}{2}m_1m_2^2(A+1)(A-1)$;

(iii) *as* $a_1 \to 1$, $Q \to m_1m_2^2(A-1) - a_2b_2(m_2-b_2)A$.

PROPOSITION A.2. *The following results hold*:

(i) *If we take* $m_1 = (b_1 + a_2(m_2-b_2)/m_2)^2/(b_1 + a_2(m_2-b_2)^2/m_2^2) - o(b_1)$, *then* (A.1) *can be satisfied and* $Q > 0$ *if* $b_1 \to 0$, a_2, b_2, m_2 *are properly chosen and either* $a_1 \to 1/2$ *or* $a_1 \to 1$.

(ii) *If we take* $m_1 = b_1 + 1$, *then* (A.1) *can be satisfied and* $Q < 0$ *if* $b_1 \to \infty$, $a_2 > 1$, b_2, m_2 *are properly chosen and either* $a_1 \to 0$ *or* $a_1 \to 1/2$ *or* $a_1 \to 1$.

Proof. (i) As $b_1 \to 0$, $m_1 \to a_2$, and $A \to m_2/(m_2-b_2)$, it is clear that $Q > 0$ provided that $b_1 \to 0$ and either $a_1 \to 1/2$ or $a_1 \to 1$ by Proposition A.1. On the other hand, it is clear that there are a_2, b_2, m_2 such that (i) holds.

Now, we verify (ii). If $a_1 \to 1$, $Q \to m_1m_2^2(A-1) - a_2b_2(m_2-b_2)A$. Choosing $m_1 = b_1+1$ and letting $b_1 \to \infty$, we note that $A \to 1$ and $m_1(A-1) \to 1 - a_2(m_2-b_2)/m_2$. Therefore,

$$\text{(A.2)}\qquad Q < 0 \Longleftrightarrow 1 - a_2(m_2-b_2)/m_2 < a_2b_2(1-b_2/m_2)/m_2.$$

By the above choice, (A.1) becomes equivalent to

$$\text{(A.3)}\qquad \begin{aligned} a_2(m_2-b_2)/m_2 &< 1, \\ 1 + a_2(m_2-b_2)^2/m_2^2 &< 2a_2(m_2-b_2)/m_2, \\ \sqrt{a_2+m_2} &< 1 + \sqrt{b_2}. \end{aligned}$$

Claim. There exist a_2, b_2, m_2, and $b_2 < m_2$ such that (A.2) and (A.3) are true. In fact, let $b_2 = \alpha m_2$, where $\alpha \in (0,1)$ will be determined later. If $a_2 > 1$, solving (A.3), we have

$$(a_2-1)/a_2 < \alpha < \sqrt{(a_2-1)/a_2} \quad \text{and} \quad \alpha > \left(\sqrt{1+a_2/m_2} - \sqrt{1/m_2}\,\right)^2.$$

It is evident that if

$$\sqrt{(a_2-1)/a_2} > \left(\sqrt{1+a_2/m_2} - \sqrt{1/m_2}\right)^2, \tag{A.4}$$

there is α such that (A.3) holds. We verify that there exist a_2 and m_2 such that (A.4) is valid. Let $a_2 = \beta m_2$, where $\beta \in (0,\infty)$ will be determined later, and denote

$$f(\beta) = \sqrt[4]{1-1/a_2} + \sqrt{\beta/a_2} - \sqrt{1+\beta}.$$

So, for some $\beta > 0$, $f(\beta) > 0 \Longleftrightarrow$ (A.4) holds for some a_2 and m_2. Simple analysis shows that $f(\beta)$ attains its maximal value at $\beta = 1/(a_2-1)$ and $f(1/(a_2-1)) > 0$. Take $\beta = 1/(a_2-1)$. According to our notation, (A.2) becomes equivalent to $a_2(1-\alpha^2) - 1 > 0$. If $\alpha = \sqrt{(a_2-1)/a_2}$, $a_2(1-\alpha^2) - 1 = 0$. Hence, we can find α which is close to but less than $\sqrt{(a_2-1)/a_2}$ such that (A.2) is valid. Consequently, our claim holds.

For $a_1 \to 0$ or $a_1 \to 1/2$, as above, similar analysis can be done. Therefore, from these arguments, it can follow that (ii) is also true under our requirements. □

Acknowledgment. The authors wish to thank the referees for their constructive comments, which led to a significant improvement of the original manuscript.

REFERENCES

[1] P. Abrams and L. Ginzburg, *The nature of predation: Prey dependent, ratio dependent or neither?*, Trends Ecol. Evol., 15 (2000), pp. 337–341.
[2] D. Alonso, F. Bartumeus, and J. Catalan, *Mutual interference between predators can give rise to Turing spatial patterns*, Ecology, 83 (2002), pp. 28–34.
[3] K. J. Brown and F. A. Davidson, *Global bifurcation in the Brusselator system*, Nonlinear Anal., 24 (1995), pp. 1713–1725.
[4] T. K. Callahan and E. Knobloch, *Pattern formation in three-dimensional reaction-diffusion systems*, Phys. D, 132 (1999), pp. 339–362.
[5] C. H. Chiu and S. B. Hsu, *Extinction of top predator in a three-level food-chain model*, J. Math. Biol., 37 (1998), pp. 372–380.
[6] E. Conway, D. Hoff, and J. Smoller, *Large time behavior of solutions of systems of nonlinear reaction-diffusion equations*, SIAM J. Appl. Math., 35 (1978), pp. 1–16.
[7] F. A. Davidson and B. P. Rynne, *A priori bounds and global existence of solutions of the steady-state Sel'kov model*, Proc. Roy. Soc. Edinburgh Sect. A, 130 (2000), pp. 507–516.
[8] Y. H. Du and Y. Lou, *Qualitative behavior of positive solutions of a predator-prey model: Effects of saturation*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 321–349.
[9] H. I. Freedman and P. Waltman, *Mathematical analysis of some three-species food-chain models*, Math. Biosci., 33 (1977), pp. 257–277.
[10] A. Gierer, *Generation of biological patterns and form: Some physical, mathematical and logical aspects*, Prog. Biophys. Biol., 37 (1981), pp. 1–47.
[11] A. Hastings, T. Powell, and S. B. Hsu, *Chaos in a three-species food chain*, Ecology, 72 (1991), pp. 896–903.
[12] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, New York, 1993.
[13] S. B. Hsu, T. W. Hwang, and Y. Kuang, *A ratio-dependent food chain model and its applications to biological control*, Math. Biosci., 181 (2003), pp. 55–83.
[14] S. B. Hsu, H. L. Smith, and P. Waltman, *Competitive exclusion and coexistence for competitive systems on ordered Banach spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 4083–4094.
[15] Y. Kan-on, *Existence and instability of Neumann layer solutions for a 3-component Lotka-Volterra model with diffusion*, J. Math. Anal. Appl., 243 (2000), pp. 357–372.
[16] Y. Kan-on and M. Mimura, *Singular perturbation approach to a 3-component reaction-diffusion system arising in population dynamics*, SIAM J. Math. Anal., 29 (1998), pp. 1519–1536.

[17] A. Klebanoff and A. Hastings, *Chaos in one-predator, two-prey models: General results from bifurcation theory*, Math. Biosci., 122 (1994), pp. 221–233.
[18] J. A. Leach and J. C. Wei, *Pattern formation in a simple chemical system with general orders of autocatalysis and decay. I. Stability analysis*, Phys. D, 180 (2003), pp. 185–209.
[19] C. S. Lin, W. M. Ni, and I. Takagi, *Large amplitude stationary solutions to a chemotaxis system*, J. Differential Equations, 72 (1988), pp. 1–27.
[20] Y. Lou, S. Martinez, and W. M. Ni, *On 3×3 Lotka-Volterra competition systems with cross-diffusion*, Discrete Contin. Dyn. Syst., 6 (2000), pp. 175–190.
[21] Y. Lou and W. M. Ni, *Diffusion vs cross-diffusion: An elliptic approach*, J. Differential Equations, 154 (1999), pp. 157–190.
[22] M. Mimura and Y. Nishiura, *Pattern formation in coupled reaction-diffusion systems*, Japan J. Indust. Appl. Math., 12 (1995), pp. 385–424.
[23] W. M. Ni and I. Takagi, *On the Neumann problem for some semilinear elliptic equations and systems of activator-inhibitor type*, Trans. Amer. Math. Soc., 297 (1986), pp. 351–368.
[24] W. M. Ni and J. C. Wei, *On positive solutions concentrating on spheres for the Gierer-Meinhardt system*, J. Differential Equations, 221 (2006), pp. 158–189.
[25] L. Nirenberg, *Topics in Nonlinear Functional Analysis*, American Mathematical Society, Providence, RI, 2001.
[26] K. Page, P. K. Maini, and N. A. M. Monk, *Pattern formation in spatially heterogeneous Turing reaction-diffusion models*, Phys. D, 181 (2003), pp. 80–101.
[27] P. Y. H. Pang and M. X. Wang, *Non-constant positive steady states of a predator-prey system with non-monotonic functional response and diffusion*, Proc. London Math. Soc. (3), 88 (2004), pp. 135–157.
[28] P. Y. H. Pang and M. X. Wang, *Qualitative analysis of a ratio-dependent predator-prey system with diffusion*, Proc. Roy. Soc. Edinburgh Sect. A, 133 (2003), pp. 919–942.
[29] R. Peng and M. X. Wang, *Positive steady-state solutions of the Noyes-Field model for Belousov-Zhabotinskii reaction*, Nonlinear Anal., 56 (2004), pp. 451–464.
[30] R. Peng and M. X. Wang, *Pattern formation in the Brusselator system*, J. Math. Anal. Appl., 309 (2005), pp. 151–166.
[31] R. Peng and M. X. Wang, *Positive steady-states of the Holling-Tanner prey-predator model with diffusion*, Proc. Roy. Soc. Edinburgh Sect. A, 135 (2005), pp. 149–164.
[32] R. Peng and M. X. Wang, *On pattern formation in the Gray-Scott model*, Sci. China Ser. A, 50 (2007), pp. 377–386.
[33] M. L. Rosenzweig, *Paradox of enrichment: Destabilization of exploitation systems in ecological time*, Science, 171 (1969), pp. 385–387.
[34] A. Turing, *The chemical basis of morphogenesis*, Philos. Trans. R. Soc. Lond. Ser. B., 237 (1952), pp. 37–72.
[35] M. X. Wang, *Non-constant positive steady states of the Sel'kov model*, J. Differential Equations, 190 (2003), pp. 600–620.
[36] M. X. Wang, *Stationary patterns for a prey-predator model with prey-dependent and ratio-dependent functional responses and diffusion*, Phys. D, 196 (2004), pp. 172–192.
[37] J. C. Wei and M. Winter, *Existence and stability of multiple-spot solutions for the Gray-Scott model in R^2*, Phys. D, 176 (2003), pp. 147–180.

SIAM J. APPL. MATH.
Vol. 67, No. 5, pp. 1504–1521

A SIMPLE ILLUSTRATION OF A WEAK SPECTRAL CASCADE*

DAVID J. MURAKI†

Abstract. The textbook first encounter with nonlinearity in a partial differential equation (PDE) is the first-order wave equation: $u_t + uu_x = 0$. Often referred to as the inviscid Burgers equation, this equation is familiar to many in the theoretical contexts of characteristics, wavebreaking, or shock propagation. Another canonical behavior contained within this simplest of PDEs is the *spectral cascade*. Surprisingly, buried in a little-known 1964 article by G.W. Platzman is an elegant example of an exact Fourier series solution associated with a purely sinusoidal initial condition. This Fourier representation, valid prior to wavebreaking, is generalized to arbitrary continuous initial conditions on both the periodic and infinite domains. For the specific example of Platzman's original problem, the Fourier coefficients decay exponentially with increasing wavenumber, and the decay rate flattens to zero precisely at the time of wavebreaking. It is demonstrated that two simplified descriptions, a downscale truncation and a linearization from initial conditions, also produce an exponential spectral cascade uniformly to large wavenumbers. This weak cascade is responsible for the initial generation of Fourier harmonics in the viscous Burgers equation.

Key words. spectral cascade, nonlinear wave, inviscid Burgers equation

AMS subject classifications. 35L60, 76M45

DOI. 10.1137/040619090

1. Introduction.

One of the first nonlinear partial differential equations (PDEs) typically encountered in the applied mathematical canon is the wave equation

$$u_t + uu_x = 0, \tag{1.1}$$

which, though elementary, provides a rich introduction to nonlinearity. As a first-order PDE, it provides an example with exact representations for the quasi-linear characteristics. Convergence of these characteristics leads to wavebreaking, multivaluedness, and the development of shock structures. Subsequent propagation of discontinuities is governed by Rankine–Hugoniot conditions obtained from conservation law properties of weak solutions. Beyond this, there is a vast literature associated with this equation whose early references include the simple wave of advection in one-dimensional fluid flow [7], the inviscid limit of the Burgers equation [2], and the kinematic wavespeed equation [21].

Without the advantages of linearity, the usual applications of Fourier methods do not generate modal solutions to (1.1). Rather the opposite occurs, as the forward time evolution from a sinusoidal initial condition, via the wave steepening process, immediately generates a solution with nonzero Fourier amplitudes at all scales. This is an example of a spectral cascade, whereby the nonlinear interaction of Fourier modes leads to an increase in the Fourier amplitudes at shorter spatial scales (higher wavenumbers). While this imagery of the downscale cascade is quite intuitive, as the textbook Fourier methods do not apply to nonlinear PDEs, the absence of illustrative examples is one barrier to elementary-level analysis of this process. It is relatively

*Received by the editors November 17, 2004; accepted for publication (in revised form) May 4, 2007; published electronically August 24, 2007. This work was supported by NSERC RGPIN-238928 and NSF CMG-0327658.

http://www.siam.org/journals/siap/67-5/61909.html

†Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6 Canada (muraki@math.sfu.ca).

unknown, however, that a Fourier series solution, whose coefficients are expressed as Bessel functions, can be elegantly derived for the evolution of (1.1) in the special case of a sinusoidal initial condition. This surprising result, by Platzman in 1964, appeared in *Tellus*, a journal for dynamic meteorology and oceanography [15].

In this article, we generalize this result to obtain an integral representation of the Fourier coefficients for arbitrary periodic initial conditions, which is valid up to the time of first wavebreaking. This gives an exact formula for each Fourier amplitude as a function of wavenumber and time, which requires only a spatial quadrature over the initial condition. For the wave equation (1.1), the Fourier spectrum is characterized by an exponential decay with wavenumber [18]. The increase in the decay rate with time is a convenient measure of the developing cascade. The weak cascade process is further investigated from the perspectives of spectral dynamics and linearized PDE dynamics about small amplitude initial conditions. For the specific case of sinusoidal initial conditions, both these perspectives on the cascade dynamics also produce short-time approximations where the exponential decay of the Fourier spectrum is uniform to large wavenumbers. The inviscid cascade is shown to be consistent with initial growth of the exponential spectra observed for the viscous Burgers equation. Finally, the Fourier solution of the wave equation is extended to the infinite line, where it is applied to the downscale cascade from a Gaussian initial condition.

The primary intent here is the presentation of explicit PDE solutions which illustrate the downscale cascade. First, we examine an exact integral formula for obtaining the Fourier coefficients of solutions to the nonlinear wave initial value problem (1.1). Additionally, the special Platzman solution, whose Fourier coefficients are expressible using Bessel functions, provides a benchmark against which we can compare various approximate descriptions of the cascade process. As it happens, the concepts required to relate this particular story nearly read as an introductory syllabus of applied mathematics: characteristics, Fourier representations, special functions, perturbation series, contour integration, and integral asymptotics. So, in keeping with the illustrative nature of this problem, these calculations have been presented in a manner to emphasize its more expository aspects.

2. From characteristics to Fourier series. Consider the general initial value problem of the nonlinear wave equation

$$u_t + uu_x = 0, \qquad u(x,0) = f(x), \tag{2.1}$$

periodic on a domain $-\pi \le x \le \pi$. The characteristics are curves in x-t space which are defined by the ordinary differential equation (ODE)

$$\frac{dx}{dt} = u, \qquad x(0) = x_0, \tag{2.2}$$

where x_0 labels the originating initial point at $(x,t) = (x_0, 0)$. Along this characteristic, the PDE (2.1) is now seen to be the perfect derivative

$$\frac{du}{dt} = 0, \qquad u(0) = f(x_0), \tag{2.3}$$

which shows that u maintains the constant value established at its initial point $(x_0, 0)$. Solutions to the ODEs (2.2) and (2.3) produce the wave solution

$$u = f(x_0), \qquad x = ut + x_0, \tag{2.4}$$

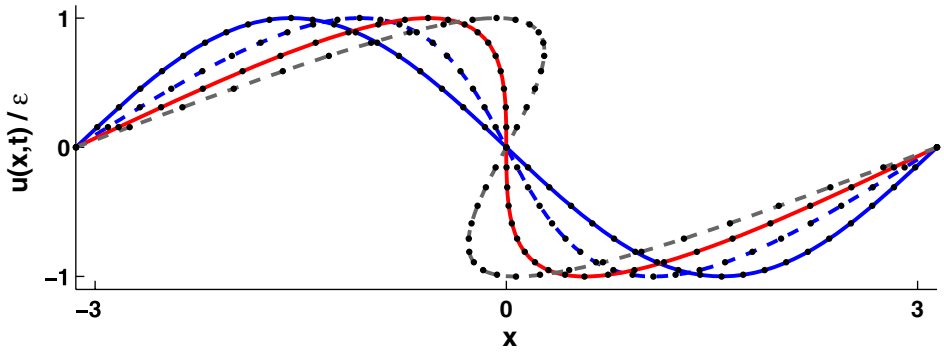


FIG. 2.1. *Wavebreaking evolution of $u(x,t)/\epsilon$ beginning from a sinusoidal initial condition* (2.6). *Shown are scaled times $\epsilon t = 0, 1/2, 1, 3/2$ with the initial and critical wavebreaking profiles in solid and an overturning profile in gray dashed. Obtained from the parametric solution* (2.4), *the dots track values of $u = f(x_0)$ corresponding to characteristics labeled by x_0 at intervals of $\pi/20$.*

expressed as a parametrization on x_0. Eliminating the parameter immediately produces the well-known implicit general solution for $u(x,t)$:

$$u = f(x - ut). \tag{2.5}$$

It is a consequence of the nonlinearity in (1.1) that (nontrivial) solutions beginning from smooth initial conditions will eventually develop a finite-time derivative singularity. Figure 2.1 shows the solution $u(x,t)$ beginning from the sinusoidal initial condition

$$f(x) = -\epsilon \sin x \tag{2.6}$$

at times $\epsilon t = 0, 1/2, 1, 3/2$, where the critical wavebreaking event occurs at $\epsilon t_c = 1$. Although the ϵ can be removed by rescaling, it is retained for future convenience in the short-time analyses in later sections.

At first glance, construction of a Fourier series solution directly from the PDE (1.1) seems unlikely since nonlinearity precludes the usual application of Fourier transforms. It is a truly remarkable consequence from Platzman's original analysis that the Fourier series representation of $u(x,t)$,

$$u(x,t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n(t) \cos nx + b_n(t) \sin nx], \tag{2.7}$$

$$a_n(t) = \frac{1}{\pi} \int_{-\pi}^{+\pi} u(x,t) \cos nx \, dx, \tag{2.8}$$

$$b_n(t) = \frac{1}{\pi} \int_{-\pi}^{+\pi} u(x,t) \sin nx \, dx, \tag{2.9}$$

has coefficients $a_n(t)$ and $b_n(t)$, which can be manipulated into integrals completely determined by the given initial profile $f(x)$. For the sine coefficient $b_n(t)$, this reformulation begins from an integration by parts of (2.9), followed by a replacement of u_x using the parametrized form of the characteristic $x = ut + x_0$ (2.4),

$$b_n(t) = \frac{1}{\pi n}\int_{-\pi}^{+\pi} u_x(x,t)\cos nx\,dx$$
$$= \frac{1}{\pi nt}\int_{-\pi}^{+\pi}\left(1-\frac{dx_0}{dx}\right)\cos nx\,dx. \tag{2.10}$$

Noting that only the dx_0/dx-term contributes to the full-period integration, changing the variable of integration to x_0 gives

$$b_n(t) = -\frac{1}{\pi nt}\int_{-\pi}^{+\pi}\cos[nx_0 + nt\,f(x_0)]dx_0 \tag{2.11}$$

and achieves a final integral which involves only the initial condition (2.1). Analogous operations obtain the cosine coefficients for $n \geq 0$:

$$a_n(t) = \begin{cases} \dfrac{1}{\pi}\displaystyle\int_{-\pi}^{+\pi} f(x_0)dx_0 & \text{for } n = 0, \\ \dfrac{1}{\pi nt}\displaystyle\int_{-\pi}^{+\pi}\sin[nx_0 + nt\,f(x_0)]dx_0 & \text{for } n > 0, \end{cases} \tag{2.12}$$

where the exceptional $n = 0$ case is simply the conservation of the mean by the PDE (1.1). It is important to note that the use of integration by parts assumes that the solution remains continuous and hence is not valid after wavebreaking.

A further step can be taken by substituting the Fourier coefficients (2.12) and (2.11) back into the series (2.7). First, the Fourier sine and cosine sums collapse into a single sum

$$u(x,t) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\frac{1}{\pi nt}\int_{-\pi}^{+\pi}\sin n[x - x_0 - tf(x_0)]dx_0$$
$$= \frac{a_0}{2} + \frac{1}{t}\int_{-\pi}^{+\pi}\left[\left(\frac{x - x_0 - t\,f(x_0)}{2\pi}\mathrm{mod}1\right) - \frac{1}{2}\right]dx_0; \tag{2.13}$$

then an interchange of sum and integral yields what seems to be a quadrature solution for (2.1). Prior to crossing of characteristics, however, $u(x,t)$ cannot depend globally on the initial condition, but is determined exactly by one value of the initial condition. The resolution of this apparent nonlocality is the presence of the modulus in (2.13), which produces a discontinuous integrand. The discontinuity occurs precisely at the unique value of x_0 parametrizing the characteristic (2.4) that determines $u(x,t)$. Shifting the integration domain to the periodic interval $x_0 - 2\pi \leq y \leq x_0$ allows the removal of the modulus

$$u(x,t) = \frac{a_0}{2} + \frac{1}{t}\int_{x_0-2\pi}^{x_0}\left[\frac{x - y - t\,f(y)}{2\pi} - \frac{1}{2}\right]dy$$
$$= \left[\frac{a_0}{2} - \frac{1}{2\pi}\int_{x_0-2\pi}^{x_0} f(y)\;dy\right] + \int_{x_0-2\pi}^{x_0}\left[\frac{x - x_0}{2\pi t} - \frac{y - (x_0 - \pi)}{2\pi t}\right]dy$$
$$= \frac{x - x_0}{t} = f(x_0) \tag{2.14}$$

and, after some grouping of terms, reduces the integral to the local value $f(x_0)$. Similarly, for the case when several characteristics are involved, the integral then

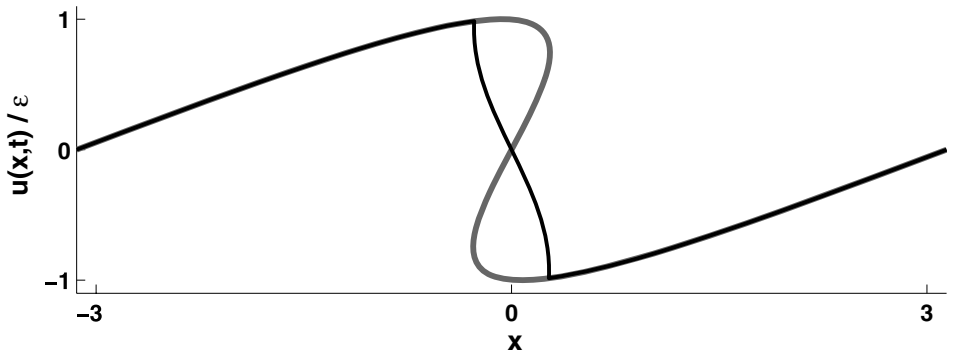


FIG. 2.2. *The characteristic solution* (2.4) *from Figure* 2.1 *(thick, light curve) beyond the wave-breaking time* ($\epsilon t = 3/2$) *compared with the Fourier series representation* (3.2)*, which is single-valued and continuous (thin, dark curve). The two solutions differ only in regions where the characteristic solution is multivalued.*

becomes a weighted sum over all such characteristic values $\pm f(x_0)$, where the sign matches that of dx/dx_0. For instance, when the characteristic solution becomes triple-valued ($u_b < u_m < u_t$), the series adopts the value $u_b - u_m + u_t$. This averaging effect within the Fourier series is illustrated by the thin dark curve in Figure 2.2, in comparison to the multivalued characteristic solution ($\epsilon t = 3/2$) as replicated from Figure 2.1. Thus, although the Fourier series defined by (2.11) and (2.12) no longer satisfies the original PDE (1.1) after wavebreaking, the series retains a meaning related to the multivaluedness of the characteristic solution (2.4), but not one connected with any of the usual entropy solutions [8].

3. Platzman's solution and its downscale cascade. The specific example considered by Platzman [15] was based upon the sinusoidal initial condition (2.6), whose forward evolution is shown as Figure 2.1. It is this solution for which Platzman essentially realized that the Fourier coefficient (2.11),

$$b_n(t) = -\frac{1}{\pi nt}\int_{-\pi}^{+\pi} \cos(nx_0 - nt\epsilon \sin x_0)dx_0 = -2\frac{J_n(\epsilon nt)}{nt}, \tag{3.1}$$

resulted in a standard integral representation of the Bessel function of order n [1]. This produced a solution to the nonlinear wave equation (1.1) having an exact expression for its Fourier sine series

$$u(x,t) = -2\sum_{n=1}^{\infty} \frac{J_n(\epsilon nt)}{nt} \sin nx, \tag{3.2}$$

where, in the $t \to 0^+$ limit, only the $n = 1$ term is nonzero and the initial condition (2.6) is satisfied. In classical analysis, summations whose terms involve Bessel functions of increasing indices and arguments are known as Kapteyn series [19]. As a historical aside, Platzman also recognized that the identical series also appears in the analysis of the Keplerian orbital problem.

It is clear from the coefficients (3.1) that all modes become activated for $t > 0$. This is an illustration of a *downscale spectral cascade* whereby the nonlinear evolution from a single initial Fourier mode leads to the immediate appearance of all smaller

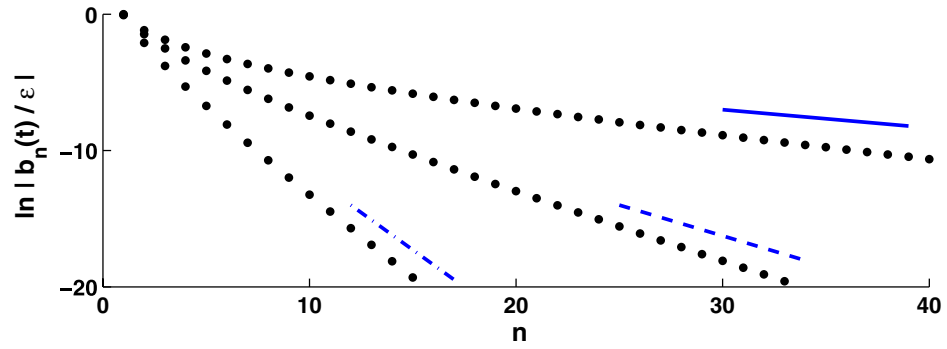


FIG. 3.1. *Semilog plot of spectral amplitudes $b_n(t)/\epsilon$ for Platzman's Fourier series solution* (3.2) *showing the growth of the $n > 1$ modes at scaled times $\epsilon t = 1/4, 1/2, 3/4$. The exponential spectrum is indicated by linear asymptotes* (3.5), *which are becoming flatter with time (dash-dot, dash, solid), and thus illustrates a downscale spectral cascade. The decrease in the fundamental $n = 1$ mode is not discernable on this semilog axis.*

scales. A Bessel recurrence identity [1] gives an alternate expression for (3.1),

$$b_n(t) = -\epsilon \frac{J_{n+1}(\epsilon n t) + J_{n-1}(\epsilon n t)}{n}, \tag{3.3}$$

from which it follows that the $n \geq 2$ amplitudes $|b_n(t)|$ are strictly increasing up to the time of wavebreaking, since $J_n'(z) > 0$ in the interval $0 < z < n$ [19]. The exception is the fundamental amplitude $|b_1(t)|$, the source of the cascade, which decreases steadily and is roughly 88% of its original amplitude at the time of critical wavebreaking. Figure 3.1 shows a semilog plot of the Fourier amplitudes against wavenumber for the times $\epsilon t = 1/4, 1/2, 3/4$.

Also shown in Figure 3.1 are lines indicating the large-n asymptotic slopes of the semilog spectral amplitudes. These are evident from the Debye expansions for the Bessel functions of large index and argument [1],

$$|b_n| \sim \sqrt{\frac{2}{\pi t^2 \tanh\alpha}}\, n^{-3/2} e^{n(-\alpha + \tanh\alpha)} \quad \text{as } n \to \infty, \tag{3.4}$$

where $\cosh\alpha = 1/\epsilon t$. This wavenumber-dependence of the Fourier coefficient reflects the (real) analytic nature of the solution $u(x,t)$ [3]. The spectral slope, $-\alpha + \tanh\alpha$, represents the exponential decay rate with wavenumber and can be explicitly written in terms of ϵt:

$$\frac{\ln|b_n|}{n} \sim \ln\left(\frac{\epsilon t}{2}\right) + \sqrt{1-\epsilon^2 t^2} - \ln\left(\frac{1+\sqrt{1-\epsilon^2 t^2}}{2}\right) \quad \text{as } n \to \infty. \tag{3.5}$$

This expression is equivalent to that deduced by Sulem, Sulem, and Frisch [18] from the pole singularities of the analytic continuation of $u(x,t)$ to the complex x-plane.[1] The early cascade has a spectral slope whose growth is logarithmic in time and corresponds to a geometric decay of the Fourier amplitudes by the factor ϵt (as illustrated

[1] Also in [18] is the identification of a narrow $n^{-4/3}$ spectral regime which occurs just prior to the critical wavebreaking time. This corresponds to a special case of the Bessel asymptotics [1].

later in (5.3)). However, as the wavebreaking $\epsilon t = 1$ is approached, the spectral slope flattens to zero. After this time, the decay becomes algebraic following the development of the derivative singularities such as those shown in Figure 2.2.

The connection of (3.1) to a known integral identity appears to be unique to the Platzman initial condition. More generally, however, the form of (2.12) and (2.11) is such that when the integrands are analytic, extraction of the exponential cascade can be approached by deforming the path of integration into the complex x-plane. For entire $f(x)$, the method of steepest descent applies, as the integration path can be deformed such that the large n contribution is localized to a saddle point. This approach can also be used to obtain (3.4). The saddle-point method is illustrated by the example for the infinite line formulation in section 7.

4. Spectral dynamics and the short-time cascade. A conventional approach for analyzing the cascade is by direct substitution of the series (2.7) into the PDE (1.1). For the special case of a Fourier sine series, the terms involved in the $\sin nx$-mode are

$$\begin{aligned} \cdots + b_n' \sin nx + \cdots & \\ + \sum_{1}^{n-1} \quad & k b_k b_{n-k} \cos kx \sin(n-k)x \\ + \sum_{1}^{\infty} \quad & k b_k b_{n+k} \cos kx \sin(n+k)x \\ + \sum_{1}^{\infty} \quad & (n+k) b_{n+k} b_k \cos(n+k)x \sin kx \quad + \cdots = 0. \end{aligned} \tag{4.1}$$

After applying a trigonometric product identity and reorganizing the terms, a description of the spectral dynamics is obtained as coupled ODEs:

$$b_n' = -\frac{n}{4} \sum_{1}^{n-1} b_k b_{n-k} + \frac{n}{2} \sum_{1}^{n-1} b_k b_{n+k} + \frac{n}{2} \sum_{n}^{\infty} b_k b_{n+k} \tag{4.2}$$

for the amplitudes $b_n(t)$ over wavenumbers n. The first of the three sums corresponds to downscale transfer involving longer waves with wavenumbers from below, $k < n$ and $(n-k) < n$. The second corresponds to mixing transfer involving straddling wavenumbers, $k < n < n+k$, while the third corresponds to upscale transfer involving only shorter waves, $n \le k < n+k$. These last two summations can be combined into a single sum. It is quite unclear as to how the Bessel amplitudes (3.1) could possibly have been directly obtained beginning only from the spectral ODEs (4.2) and the initial conditions $\{b_n(0)\} = \{-\epsilon, 0, 0, \ldots\}$.

Analytical progress is possible, however, in the limit of small ϵ. At $O(1)$ times, the assumption of small amplitude initial condition leads to a wavenumber scaling of the Fourier amplitudes $b_n(t) = O(\epsilon^n)$ and allows a natural truncation of the spectral dynamics (4.2) to involve only the downscale transfer summation

$$\tilde{b}_n' = \begin{cases} 0 & \text{for } n = 1, \\ -\dfrac{n}{4} \displaystyle\sum_{1}^{n-1} \tilde{b}_k \tilde{b}_{n-k} & \text{for } n \ge 2. \end{cases} \tag{4.3}$$

This will be referred to as the downscale cascade truncation. The exact solution to the above truncation must therefore be the small ϵ limit of Platzman's solution (3.1),

$$\tilde{b}_n(t) = -\epsilon \frac{n^{n-1}}{n!} \left(\frac{\epsilon t}{2}\right)^{n-1}, \tag{4.4}$$

which derives from the first nonzero term of the Taylor expansion for the Bessel function [1]. Verification of this, by direct substitution of (4.4) into (4.3), yields a combinatorial identity of uncommon origin—one such instance is found in graph theory as an elementary counting of trees [10]. A direct approach for arriving at expression (4.4) is via a generating function

$$B(z,t) = \sum_{n=1}^{\infty} \tilde{b}_n(t) \frac{e^{inz}}{2i}. \tag{4.5}$$

By virtue of the downscale spectral dynamics (4.3), $B(z,t)$ also satisfies the same nonlinear wave equation (1.1),

$$B_t + BB_z = 0, \qquad B(z,0) = \epsilon \frac{e^{iz}}{2i}, \tag{4.6}$$

but now with a complex-valued initial condition that is exactly the restriction of the original sinusoid to the positive wavenumber modes. Solution by characteristics leads to the implicit relation

$$iBte^{iBt} = \frac{\epsilon t}{2} e^{iz}, \tag{4.7}$$

whose inversion can be expressed in terms of Lambert's transcendental equation (see also [20]), and otherwise designated by the *W-function* [6],

$$B(z,t) = -\frac{i}{t} W\left(\frac{\epsilon t}{2} e^{iz}\right). \tag{4.8}$$

However, explicit recovery of the formula for the coefficients (4.4) follows more directly from (4.7) with the application of the Lagrange inversion theorem. Using the Stirling approximation for the factorial in (4.4) gives the large wavenumber behavior

$$\tilde{b}_n \sim -\sqrt{\frac{2}{\pi t^2}} n^{-3/2}\, e^n \left(\frac{\epsilon t}{2}\right)^n \quad \text{for } n \to \infty \tag{4.9}$$

and implies the downscale spectral slope

$$\frac{\ln|\tilde{b}_n|}{n} \sim \ln\left(\frac{\epsilon t}{2}\right) + 1 \quad \text{as } n \to \infty \text{ for } \epsilon t \ll 1. \tag{4.10}$$

The difference here from the full cascade (3.5) is that the slope from the downscale cascade truncation (4.10) is less steep, as the exclusion of any upscale transfers results in a more rapid generation of a smaller-scale spectrum. Hence, at short times ($\epsilon t \ll 1$), the spectral slope (3.5) for Platzman's example is well described by the downscale cascade approximation (4.3).

5. A linearized description of the weak cascade. One conclusion from the previous section is that even the truncation of the spectral dynamics to the downscale transfer requires the solution of a fully nonlinear problem. As such, the results relied upon considerable good karma in there being an exact solution (4.4) to a system of nonlinear equations (4.3). In this section, a linear approach is investigated for constructing an approximate solution to the PDE (2.1) that involves the full spectrum of wavenumbers.

Consider a weakly nonlinear analysis which seeks the form of a perturbation expansion

$$u(x,t) \sim f(x) + u_2(x,t) + u_3(x,t) + \cdots, \tag{5.1}$$

where the first term is a small amplitude initial condition $f(x) = O(\epsilon) \gg u_2(x,t) \gg u_3(x,t)\ldots$ for $\epsilon \ll 1$. The simplest such expansion assumes that the corrections $u_n(x,t) = O(\epsilon^n)$. Substituting (5.1) into the PDE and collecting on powers of ϵ gives the sequence of equations

$$\frac{\partial u_n}{\partial t} = -\sum_1^{n-1} u_{n-k}\frac{\partial u_k}{\partial x}, \qquad u_n(x,0) = 0, \tag{5.2}$$

which can be solved iteratively for $n \geq 2$ by direct integration for $t > 0$. For the sinusoidal initial conditions, the first two corrections are

$$\begin{aligned} u_2(x,t) &= -\epsilon \left(\frac{\epsilon t}{2}\right) \quad \sin 2x, \\ u_3(x,t) &= -\epsilon \left(\frac{\epsilon t}{2}\right)^2 \quad \left\{\frac{3}{2}\sin 3x - \sin x\right\}; \end{aligned} \tag{5.3}$$

subsequent terms $u_n(x,t)$ contain only $O(\epsilon^n)$ expressions, which include not only the short-time harmonic $\tilde{b}_n(t)\sin nx$ from the downscale transfer (4.4), but also smaller harmonics due to contributions from the straddling and upscale transfers (4.2). Finite application of this method thus produces an $O(\epsilon^n)$ series expansion limited to the first n harmonics. Such a finite expansion is not a uniform approximation over wavenumbers, since for the sinusoidal initial condition the extent to which the spectral cascade is realized is limited by the number of terms in the expansion (5.1).

To develop an approach which involves all harmonics, consider the solution as a disturbance from a small amplitude initial condition

$$u(x,t) = f(x) + \tilde{u}(x,t), \tag{5.4}$$

so that $\tilde{u}(x,t) \ll f(x) = O(\epsilon)$. This results in the exact disturbance equation

$$\tilde{u}_t = -ff_x - (f\tilde{u})_x - \tilde{u}\tilde{u}_x, \quad \tilde{u}(x,0) = 0, \tag{5.5}$$

where the right-hand side terms are nominally $O(\epsilon^2), O(\epsilon^3)$, and $O(\epsilon^4)$. If (5.5) is approximated by keeping only the ff_x-term, then the disturbance $\tilde{u}(x,t)$ is $O(\epsilon^2)$-correct and would be identical to $u_2(x,t)$ as determined by (5.2). Alternatively, an additional order in $\tilde{u}(x,t)$ is achieved if only the last and nonlinear disturbance term is neglected. This truncation can be interpreted as a first Newton iterate, since the $\tilde{u}$-correction is obtained by a linearized solve against a residual error (in the form of the ff_x-term). Thus we consider the linearized problem

$$U_t + (fU)_x = -ff_x, \quad U(x,0) = 0, \tag{5.6}$$

so that $u(x,t) \sim f(x)+U(x,t)$ constitutes an $O(\epsilon^3)$-correct asymptotic representation. Multiplying the equation though by $f(x)$ and defining $v(x,t) = f(x)\,U(x,t)$ gives

$$v_t + f v_x = -\frac{1}{2} f (f^2)_x, \quad v(x,0) = 0, \tag{5.7}$$

which is a first-order but nonconstant coefficient and inhomogeneous PDE. Unlike the original PDE (1.1), the characteristics for the linearization (5.6) do not depend on the solution, but only on the initial condition, via

$$\frac{dx}{dt} = f(x), \quad x(0) = x_0, \tag{5.8}$$

where again x_0 labels the originating initial point at $(x,t) = (x_0, 0)$. Along this characteristic, the PDE (5.8) now becomes the perfect derivative

$$\frac{dv}{dt} = -\frac{1}{2}\frac{dx}{dt}(f^2)_x = -\frac{1}{2}\frac{d(f^2)}{dt}, \quad v(x_0,0) = 0, \tag{5.9}$$

which relies upon the t-independence of f^2. Direct integration from a zero initial condition gives the solutions

$$\begin{aligned} v(x,t) &= -\frac{1}{2}\left(f^2(x) - f^2(x_0)\right), \\ U(x,t) &= -\frac{1}{2}\left(\frac{1 - f^2(x_0)}{f^2(x)}\right) f(x), \end{aligned} \tag{5.10}$$

where the label $x_0 = x_0(x,t)$ is obtained by inverting the solution of the characteristic ODE (5.8). Specifically for Platzman's initial condition, it is shown next that this correction term is no longer spectrally limited to a few harmonics but embodies a cascade across all wavenumbers.

For the case of $f(x) = -\epsilon \sin x$, the characteristic ODE (5.8) is a nonlinear but separable equation; hence

$$\ln\left(\frac{\tan x/2}{\tan x_0/2}\right) = \int_{x_0}^{x} \frac{dx}{\sin x} = -\epsilon \int_0^t dt = -\epsilon t, \tag{5.11}$$

from which the trigonometric relation $\tan(x_0/2) = e^{\epsilon t}\tan(x/2)$ follows. Using this and a half-angle identity gives

$$\sin x_0 = \frac{2\tan(x_0/2)}{1+\tan^2(x_0/2)} = \frac{2e^{\epsilon t}\tan(x/2)}{1+e^{2\epsilon t}\tan^2(x/2)} = \frac{\operatorname{sech}\epsilon t}{1-\tanh\epsilon t\cos x}\sin x, \tag{5.12}$$

which, in the Platzman case, effects the inversion of the characteristic label x_0 into the original x,t-coordinates. Thus, the linearized solution (5.10) leads to the asymptotic approximation

$$u(x,t) \sim -\epsilon \sin x + \frac{\epsilon}{2}\left(1 - \frac{\operatorname{sech}^2\epsilon t}{(1-\tanh\epsilon t\cos x)^2}\right)\sin x + O(\epsilon^4), \tag{5.13}$$

where the second term is actually $O(\epsilon^2)$ with the vanishing of the bracketed factor when $\epsilon = 0$.

Obtaining the spectral cascade requires finding the Fourier series representation of (5.13). The obvious problematic term is the second term in the correction whose Fourier-sine coefficient is the imaginary part of

$$-\frac{\epsilon}{2\pi}\operatorname{sech}^2\epsilon t\int_{-\pi}^{+\pi}\frac{e^{inx}\sin x}{(1-\tanh\epsilon t\cos x)^2}\,dx. \tag{5.14}$$

This expression can be evaluated via complex contour integration around a rectangle whose corners are $\{-\pi,+\pi,+\pi+iY,-\pi+iY\}$. Contributions from the sides parallel to the imaginary axis cancel by the periodicity of the integrand, and the contribution from the side with $\operatorname{Im}(z)=Y$ tends to zero as $Y\to+\infty$ by the decay of the integrand. The closed contour contains only a double pole at z_p, where

$$\begin{aligned}\cos z_p=\cosh iz_p&=\frac{\tanh(\epsilon t2)+\coth(\epsilon t2)}{2}=\coth\epsilon t\\ \Rightarrow\qquad e^{iz_p}&=\tanh(\epsilon t2),\end{aligned} \tag{5.15}$$

and thus is located along the positive imaginary axis for $\epsilon t>0$. The end result of this residue calculation is the series representation for (5.13):

$$u(x,t)\sim-\frac{\epsilon}{2}\left(\sin x+\operatorname{sech}^2(\epsilon t2)\sum_1^\infty n\tanh^{n-1}(\epsilon t2)\sin nx\right), \tag{5.16}$$

which reveals that again the spectrum has exponential decay, whose spectral slope is

$$\text{slope}\ \sim\ \ln\left|\tanh\frac{\epsilon t}{2}\right|\quad\text{as } n\to\infty\ \text{ for } \epsilon t\ll 1. \tag{5.17}$$

Thus the logarithmic part of the $\epsilon t\ll 1$ spectral slope is obtained. It is emphasized that this is really just a scaling result on the amplitudes, indicating only that harmonics decay as powers of ϵt. This limited result is not too surprising since the downscale cascade within the short-time and linear approximation is still a fully nonlinear process (4.3). Nonetheless, for the sinusoidal initial condition, the linearization (5.6) does produce, after just one perturbative calculation, an explicit short-time correction (5.13) that is asymptotically valid only to $O(\epsilon^3)$ yet captures the exponential decay of the spectrum uniformly to large wavenumbers.

More generally, the linearization generates Fourier harmonics through the action of the nonconstant coefficient (5.7), as the Fourier harmonics are no longer the linear modes. In essence, this allows a coupling between Fourier coefficients, but one where the strength of the interaction is determined by the nonconstant coefficient and hence is fixed in time by the initialized state. The above example shows that even when the coupling is established by an initial condition consisting only of a single initial sinusoid, the cascade described by (5.6) captures the downscale cascade with an exponential decay in wavenumber. A second example of a linearized cascade based on a spectrally richer initial Gaussian is computed for the infinite line case of section 7.

6. The weak cascade of the Burgers equation. The exponential spectrum appears in the wave equation (1.1) as a consequence of pole singularities associated with the analytic continuation of the inversion (2.5) over complex x [18]. An exponential spectrum is also common to solutions of the Burgers equation [16, 18], and it includes the effect of viscous, linear dissipation into the nonlinear wave equation [2]:

$$u_t+uu_x=\sigma u_{xx},\qquad u(x,0)=f(x). \tag{6.1}$$

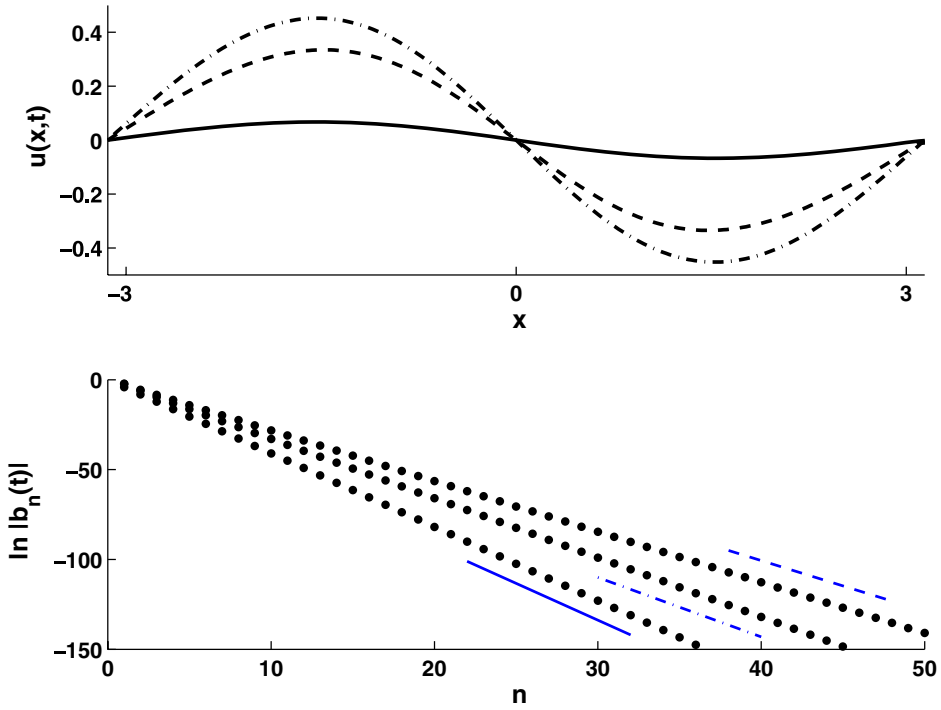


FIG. 6.1. *Top panel shows a Burgers solution beginning from a sinusoidal initial condition for* $\epsilon = 0.5$, $\sigma = 1.0$ *at times* $t = 0.1, 0.4, 2.0$ *(dash-dot, dashed, solid). Bottom panel shows the corresponding semilog Fourier spectra with the asymptotic slopes indicated. The spectral slope initially increases until roughly* $t = 0.4$*, and then decreases linearly in time.*

A familiar spectral result [2, 18, 4] is the exponential spectrum for the steady-state tanh-solution. It is also well known that the Burgers dynamics is equivalent to the linear diffusion equation via the Hopf–Cole transformation [21],

$$u = -2\sigma \frac{\psi_x}{\psi}, \qquad \psi_t = \sigma \psi_{xx}. \tag{6.2}$$

For Hopf–Cole functions $\psi(x,t)$ which are meromorphic over complex x, the evaluation of Fourier coefficients

$$b_n(t) = -\frac{2\sigma}{\pi} \mathrm{Imag} \int_{-\pi}^{+\pi} \frac{\psi_x(x,t)}{\psi(x,t)} e^{inx} dx \tag{6.3}$$

by a contour integration of the type used to obtain (5.14) involves only simple poles. The residue of the pole with smallest imaginary part determines the spectral slope of the exponential spectrum [18].

The upper panel of Figure 6.1 shows the decay of the Burgers solution beginning from an initial sinusoid (2.6) of small amplitude ($\epsilon = 0.5$, $\sigma = 1.0$) as computed by a fully spectral code. Clearly apparent in the lower panel of Figure 6.1 are the linear asymptotes in the corresponding semilog plots of Fourier amplitudes. Inspection of the chronology of the spectra (dash-dot, dashed, solid) reveals that the spectral slopes initially increase in time, and subsequently decrease through the action of viscous dissipation. Figure 6.2 (solid) shows the best-fit spectral slopes, as a function of

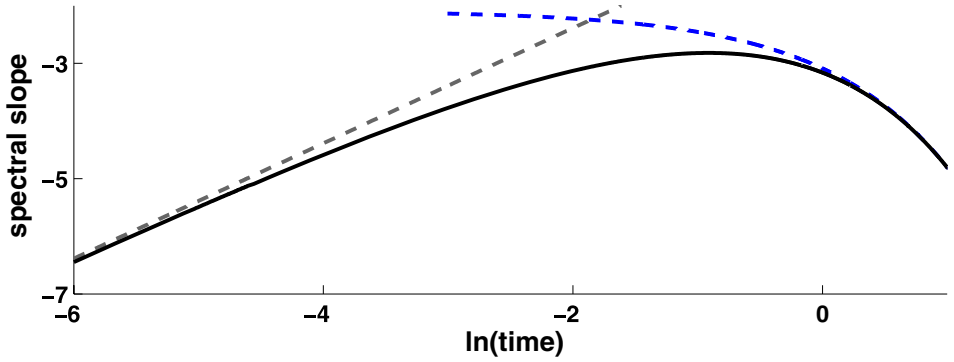


FIG. 6.2. *The semilog spectral slope as a function of log-time for the evolution of Figure* 6.1 *as obtained by the best-fit line over modes* $n = 80$–100 *(solid). At short times, the Platzman spectral slope* (4.10) *illustrates that the growth of the initial cascade scales similarly as the inviscid dynamics (gray dashed). At long times, the slope follows the asymptote as obtained from the Hopf–Cole solution* (6.8) *and shows the erosion of the spectrum by the viscous decay (dashed).*

log-time, for the evolution of Figure 6.1. In terms of the spectral slope, the Burgers cascade exhibits the same growth as the inviscid cascade (4.10) at very early times (gray dashed). At later times, when the amplitudes are decaying, the spectral slope approaches a linear-in-time asymptote (dashed). It is not apparent how to obtain the early-time spectral growth directly from the Hopf–Cole solution, but the long-time behavior is easily extracted.

The initial Hopf–Cole function is given by

$$\psi(x,0) = \exp\Big(-\frac{\epsilon}{2\sigma}\cos x\Big) = I_0\Big(\frac{\epsilon}{2\sigma}\Big) + 2\sum_1^\infty I_n\Big(-\frac{\epsilon}{2\sigma}\Big)\cos nx, \tag{6.4}$$

where the series representation [1] involves a modified Bessel identity which is closely related to that used in the Platzman cascade (3.1). The time-dependent evolution thus has the exact Fourier solution

$$\psi(x,t) = I_0\Big(\frac{\epsilon}{2\sigma}\Big) + 2\sum_1^\infty I_n\Big(-\frac{\epsilon}{2\sigma}\Big)\, e^{-\sigma n^2 t}\cos nx. \tag{6.5}$$

After sufficiently long times (regardless of the values of ϵ and σ), the Hopf–Cole dynamics is dominated by the slowest decaying $n = 1$ mode,

$$\psi(x,t) \sim I_0\Big(\frac{\epsilon}{2\sigma}\Big) - 2I_1\Big(\frac{\epsilon}{2\sigma}\Big)\, e^{-\sigma t}\cos x. \tag{6.6}$$

The corresponding Burgers solution then has the form

$$\begin{aligned} u(x,t) &\sim -4\sigma \frac{I_1(\epsilon/2\sigma)e^{-\sigma t}\sin x}{I_0(\epsilon/2\sigma) - 2I_1(\epsilon/2\sigma)e^{-\sigma t}\cos x} \\ &= -4\sigma\sum_1^\infty e^{n\rho}\sin nx, \end{aligned} \tag{6.7}$$

where the spectral slope, ρ, calculated using the same contour used to obtain (5.14), is given by

$$\rho = -\cosh^{-1}\left(\frac{I_0(\epsilon/2\sigma)}{2I_1(\epsilon/2\sigma)}e^{\sigma t}\right) \sim -\sigma t + \ln\left(\frac{I_1(\epsilon/2\sigma)}{I_0(\epsilon/2\sigma)}\right) \quad \text{for } t \gg 1. \tag{6.8}$$

This is the long-time asymptote (dashed) shown in Figure 6.2 and is the generic spectral behavior once the Burgers evolution is dominated by a largest scale mode.

The special case of the sinusoidal initial condition illustrates that the initial downscale cascade evolves on the inviscid timescale, ϵt, despite the fact that the advective term is a weakly nonlinear effect. The cascade then transitions to dissipation on the viscous timescale, σt, as the solution comes to be dominated by the slowest decaying mode $u(x,t) \sim e^{-\sigma t} \sin x$. The modal dissipation is then driven by weak nonlinearity, as harmonics decay as $e^{-n\sigma t}$, in contrast to the $e^{-n^2\sigma t}$ decay of linear diffusion.

7. Fourier solution on the infinite line. The derivation of the periodic Fourier coefficients (2.11), (2.12) is easily modified to obtain an analogous integral for the Fourier transform solution on the infinite line. Defining the Fourier transform representation of continuous solutions to (1.1) by

$$u(x,t) = \int_{-\infty}^{+\infty} c(k,t) e^{-ikx} dk, \tag{7.1}$$

the coefficients $c(k,t)$ can also be shown to derive from the initial profile $u(x,0) = f(x)$. Beginning from the Fourier integral, an integration by parts is performed,

$$c(k,t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} u(x,t) e^{ikx} dx = \frac{i}{2\pi k} \int_{-\infty}^{+\infty} u_x(x,t) e^{ikx} dx, \tag{7.2}$$

which again assumes continuity as well as sufficiently fast decay of the solution at $x \to \pm\infty$. In a slight departure from the periodic case, the next step introduces the parametric solution $u = f(x_0)$,

$$\begin{aligned} c(k,t) &= \frac{i}{2\pi k} \int_{-\infty}^{+\infty} f'(x_0) \frac{dx_0}{dx} e^{ikx} dx \\ &= \frac{i}{2\pi k} \int_{-\infty}^{+\infty} f'(x_0) \exp[ik(x_0 + t f(x_0))] dx_0, \end{aligned} \tag{7.3}$$

where decay of the integrand is ensured through the initial profile. (Note that an analogous formula can also be derived for the periodic case.)

For example, the solution from an initial Gaussian profile $f(x) = e^{-x^2/2}$ remains single-valued up until the breaking time of $t_c = \sqrt{e}$ (Figure 7.1). In the limit of large wavenumber k, the Fourier integral (7.3) can be approximated by the method of steepest descent. The complex plane for the phase function $\phi(z) = i(z + te^{-z^2/2})$ is shown as Figure 7.2. The saddle points of the phase are determined by the stationary points $\phi'(z_s) = 0$, which for the Gaussian profile can be expressed as the condition

$$(-z_s^2) e^{(-z_s^2)} = -\frac{1}{t^2}. \tag{7.4}$$

Thus the saddle points are complex-valued solutions to Lambert's transcendental equation $z_s^2 = -W(-1/t^2)$. Figure 7.2 shows the four saddle points closest to the real z-axis at time $t/t_c = 3/4$. The integration along the real axis (7.3) can be deformed into a scalloped contour (solid curve in Figure 7.2) in the upper half-plane, so that the dominant contribution will be localized to the saddle point with the maximum $\mathrm{Re}(\phi(z_s))$. The quadratic Taylor expansion of the phase function at a saddle point simplifies to

$$\phi(z) \sim i \left(z_s + \frac{1}{z_s} \right) + \frac{i}{2} \left(z_s - \frac{1}{z_s} \right) (z - z_s)^2 \tag{7.5}$$

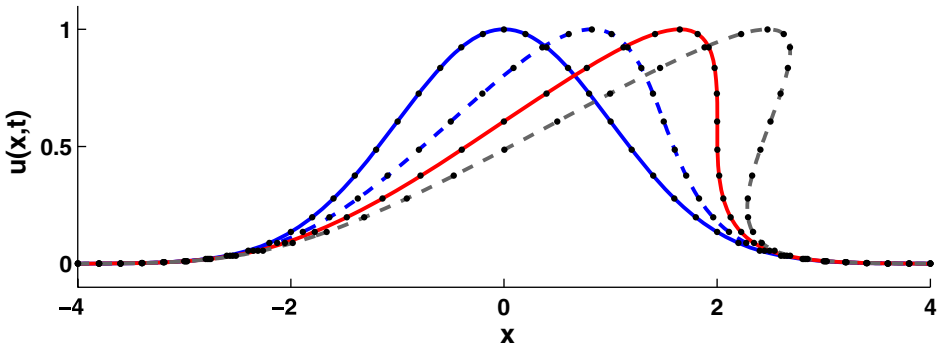


FIG. 7.1. *Wavebreaking evolution of $u(x,t)/\epsilon$ beginning from a Gaussian initial condition. Shown are scaled times $t/t_c = 0, 1/2, 1, 3/2$ with the initial and critical wavebreaking profiles in solid and an overturning profile in gray dashed. Obtained from the parametric solution (2.4), the dots track values of $u = f(x_0)$ corresponding to characteristics labeled by x_0 at intervals of $1/5$.*

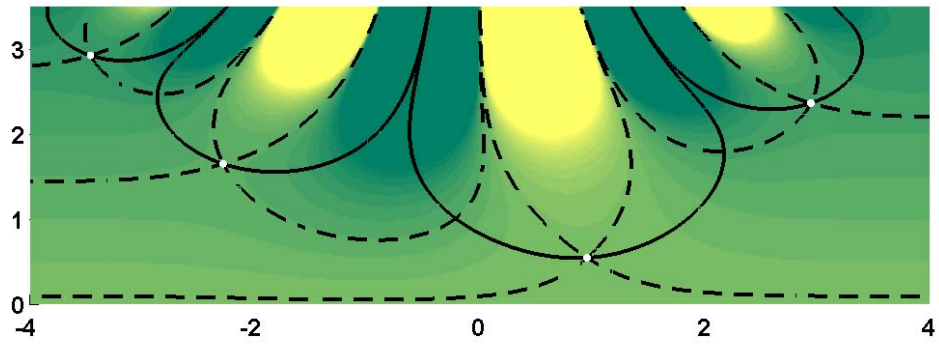


FIG. 7.2. *Complex plane for the phase function $\phi(z;t)$ for $t/t_c = 3/4$. The grayscale indicates $Re(\phi(z;t))$, where darker regions correspond to exponential smallness of the integrand. The contours shown are associated with the four saddle points closest to the real axis (closest $z_s \approx 0.95 + 0.55i$). Solid contours are paths of steepest descent to regions of exponentially small integrand. Dashed contours are level curves of the magnitude.*

and gives the steepest descent contribution

$$\sqrt{\frac{1}{2\pi i(z_s - 1/z_s)t^2}}\, k^{-3/2} e^{i(z_s + 1/z_s)k}, \tag{7.6}$$

where additional time-dependence lies in the location of the saddle point (7.4). Using only the dominant saddle point, this gives an expression for the spectral slope:

$$\frac{\ln|c(k,t)|}{k} \sim -\operatorname{Im}\left(z_s + \frac{1}{z_s}\right) \text{ as } k \to \infty, \tag{7.7}$$

which is verified by the lines in Figure 7.3. As expected, the breaking time $t_c = \sqrt{e}$ coincides with the first real root of the saddle-point condition (7.4), where the spectral decay changes from exponential to algebraic. Thus, at finite times $0 < t < t_c$, the spectrum decays exponentially despite its beginning from more rapid quadratic Gaussian decay (light curve).

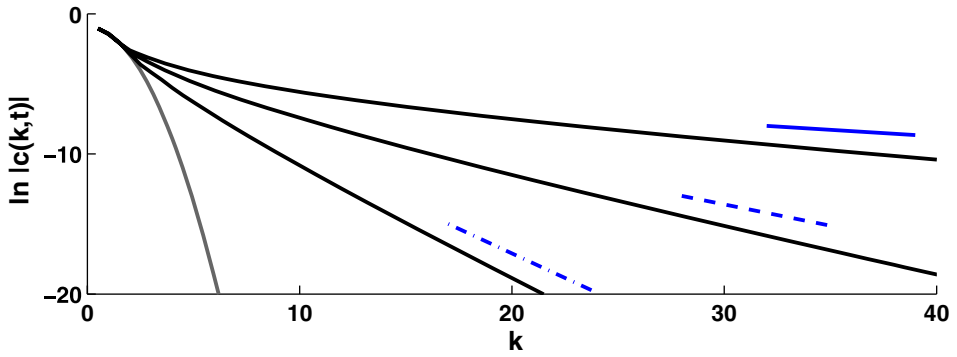


FIG. 7.3. *Semilog plot of spectral amplitudes $c_n(t)/\epsilon$ of a pseudospectral computation (de-aliased to 2048 modes on a 4π-periodic domain) from a Gaussian initial condition (leftmost curve). The downscale spectral cascade is illustrated by the growth of the Fourier amplitudes (other dark solid) over times $t/t_c = 1/4, 1/2, 3/4$. The flattening of the exponential spectrum is indicated by the linear asymptotes (dash-dot, dash, solid) as calculated from the steepest descent contribution* (7.7) *from the saddle point nearest the real axis.*

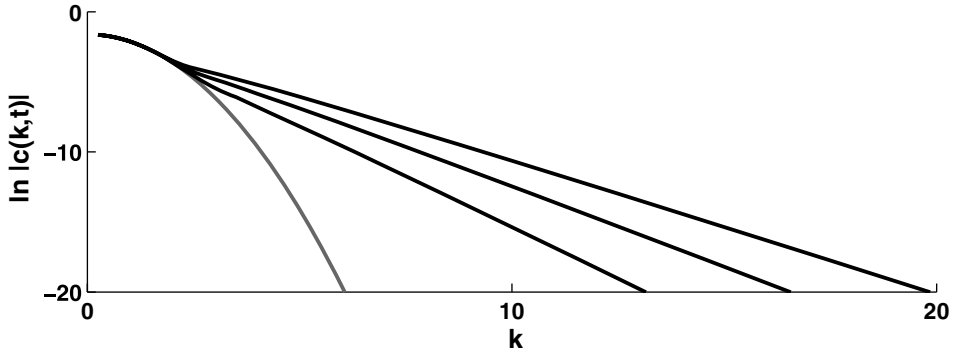


FIG. 7.4. *Semilog plot of spectral amplitudes $c_n(t)/\epsilon$ for the linearized evolution* (5.6) *from an initial Gaussian (gray solid) shows that its cascade also develops an exponential spectrum. The numerical parameters are the same as those of the nonlinear computation of Figure* 7.3. *The flattening of the spectrum is shown for earlier times $t/t_c = 1/12, 1/6, 1/4$, based on the breaking time of the nonlinear evolution. The slope of the linearized cascade at $t = t_c/4$ lags just behind the fully nonlinear cascade shown in Figure* 7.3.

The development of the linearized spectrum from the Gaussian initial condition is shown in Figure 7.4. Although the communication between the Fourier modes is fixed by the spectral characteristics at the initial time in (5.6), the linear evolution still generates a downscale cascade with an exponential decay with wavenumber.

8. Closing thoughts. The Fourier results presented here are spectral identities which follow from Platzman's observation that u_x is simply related to the change of variable dx_0/dx via the parametric solution (2.4) for (1.1). This is a rather unusual situation that does not readily apply beyond the characteristic wave equation. Nonetheless, one generalization for which such spectral formulas can be stated is

$$u_t + g(u,x)u_x = h'(t)u, \quad u(x,0) = f(x). \tag{8.1}$$

On the infinite line, the Fourier transform is expressible as the integral

$$c(k,t) = \frac{i}{2\pi k} e^{h(t)-h(0)} \int_{-\infty}^{+\infty} f'(x_0) e^{ik\,x(x_0,t)} dx_0, \tag{8.2}$$

where $x(x_0,t)$ is determined by the characteristic ODE

$$\frac{dx}{dt} = g(f(x_0)e^{h(t)-h(0)}, x), \quad x(0) = x_0. \tag{8.3}$$

The occurrence of nontrivial examples where the formulas (8.2), (8.3) allow further analysis is a rare event. However, a decaying version of Platzman's example with the additional effect $h' = -\alpha$, a constant value, yields

$$b_n(t) = \frac{2\alpha}{n(1-e^{\alpha t})} J_n\left(\frac{\epsilon n(1-e^{-\alpha t})}{\alpha}\right), \tag{8.4}$$

an exercise that reveals a suppression of the wavebreaking when $\alpha > \epsilon$.

As an explicit Fourier analysis of the inviscid Burgers equation, the spectral formulas (2.12), (2.11), (7.3) can be used to investigate cascades from other initial conditions. Although Platzman's example highlighted the downscale cascade, the dynamics of upscale transfers to large scales from smaller scales can also be addressed by initial sums of sinusoids. For real analytic initial conditions, the method of steepest descent generally applies for obtaining the asymptotic exponential spectrum. In contrast, compactly supported (but piecewise continuous) initial conditions yield simple examples whose spectra have algebraic decay.

The Burgers equation, and its inviscid limit, have long been used as a pedagogical introduction to the nonlinearity of fluid motions [2, 5, 9]. This generalization of Platzman's result provides explicit formulas for one-dimensional, deterministic, and continuous realizations of the spectral cascade — hence the qualification to a *weak* cascade. Large wavenumber asymptotics for the integral formula allow direct calculation of the exponential decay of the Fourier spectrum. The inviscid cascade is observed to characterize the early development of the spectrum in the viscous Burgers equation. The examples shown here are direct spectral illustrations of the nonlinear cascade but are far from those required to understand the multidimensional, statistical nature of fluid turbulence. Nonetheless, they are a clear demonstration of a nonlinear process by which advection can sustain an exponential spectrum in the dissipation range, the scales at which the fields are smooth [17, 12]. The turbulence question aside, these results provide an elementary contribution to the broader effort to understand the spectral signatures of singularities in nonlinear PDEs [14, 18, 16].

Although the existence of an exact solution (2.5) would seemingly render (1.1) as fully understood, recent revisitations in the research literature remind that this textbook equation still serves as a source of inspiration for investigations in nonlinearity. Weideman [20] used complex-valued solutions as tests for a method for tracking singularities using numerical analytical continuation. In particular, the dynamics of the logarithmic branch point was computed for the initial condition $u(x,0) = e^{ix}$. The sinusoidal initial condition has also been used by Majda and Timofeyev [11] to initiate postbreaking ergodic dynamics among the Fourier modes for a Galerkin truncation of the spectral dynamics. The truncated dynamics exhibits a chaos which is shown to have a well-defined statistical equilibrium. Finally, the linearization result is very closely related to some current investigations of Mattingly, Soudian, and Vanden-Eijnden [13], who are constructing linear spectral cascade models with exact solutions.

These models have spectral dynamics which are limited to linear coupling of nearest neighbors (in wavenumber), of which (5.6) is an inviscid example. Their analyses involve an unexpected generalization of the generating function method that is based on orthogonal eigenfunction expansions.

In these investigations the wave equation (1.1) is utilized as a testbed for furthering our understanding of nonlinearity. It is in a similar spirit that these one-dimensional Fourier results, although limited to continuous solutions, are communicated for their novelty as an exact spectral viewpoint.

Acknowledgments. The author acknowledges Ed Spiegel, who recommended the Platzman article some years ago. Special thanks to colleagues Ralf Wittenberg, J. F. Williams, and Youngsuk Lee for their enthusiastic discussions during the course of this work, and for their careful readings of the manuscript. Finally, the author is very grateful to Jim Colliander [18], Elef Gkioulekas [12], and Mike Siegel [14] for alerting me to related works.

REFERENCES

[1] M. Abramowitz and I. Stegun, *Handbook of Mathematical Formulas and Tables*, National Bureau of Standards, Washington, DC, 1964.
[2] J. M. Burgers, *A Mathematical Model Illustrating the Theory of Turbulence*, Adv. Appl. Mech. 1, Academic Press, New York, 1948.
[3] G. F. Carrier, M. Krook, and C. E. Pearson, *Functions of a Complex Variable*, McGraw–Hill, New York, 1966.
[4] A. J. Chorin and O. H. Hald, *Viscosity-dependent inertial spectra of the Burgers and Korteweg-deVries-Burgers equation*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 3921–3923.
[5] J. D. Cole, *On a quasi-linear parabolic equation occurring in aerodynamics*, Quart. Appl. Math., 9 (1951), pp. 225–236.
[6] R. M. Corliss, G. H. Gonnet, D. E. G. Hare, D. J. Jeffery, and D. E. Knuth, *On the Lambert W Function*, Adv. Comput. Math., 5 (1996), pp. 339–359.
[7] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Springer-Verlag, New York, 1948.
[8] L. C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, RI, 1998.
[9] E. Hopf, *The partial differential equation* $u_t + uu_x = \nu u_{xx}$, Comm. Pure Appl. Math., 3 (1950), pp. 201–230.
[10] D. E. Knuth and B. Pittel, *A recurrence related to trees*, Proc. Amer. Math. Soc., 105 (1989), pp. 335–349.
[11] A. J. Majda and I. Timofeyev, *Remarkable statistical behavior for truncated Burgers-Hopf dynamics*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 12413–12417.
[12] O. P. Manley, *The dissipation range spectrum*, Phys. Fluids A, 4 (1992), pp. 1320–1321.
[13] J. C. Mattingly, T. M. Suidan, and E. Vanden-Eijnden, J. Statist. Phys., to appear.
[14] D. W. Moore, *The spontaneous appearance of a singularity in the shape of an evolving vortex sheet*, Proc. Roy. Soc. London Ser. A, 365 (1979), pp. 105–119.
[15] G. W. Platzman, *An exact integral of complete spectral equations for unsteady one-dimensional flow*, Tellus, 16 (1964), pp. 422–431.
[16] D. Senouf, *Dynamics and condensation of complex singularities for Burgers' equation* II, SIAM J. Math. Anal., 28 (1997), pp. 1490–1513.
[17] L. M. Smith and W. C. Reynolds, *The dissipation-range spectrum and the velocity-derivative skewness in turbulent flows*, Phys. Fluids A, 3 (1991), pp. 992–994.
[18] C. Sulem, P.-L. Sulem, and H. Frisch, *Tracing complex singularities with spectral methods*, J. Comput. Phys., 50 (1983), pp. 138–161.
[19] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, Cambridge, UK, 1944.
[20] J. A. C. Weideman, *Computing the dynamics of complex singularities of nonlinear PDEs*, SIAM J. Appl. Dyn. Syst., 2 (2003), pp. 171–186.
[21] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley, New York, 1974.

SIAM J. APPL. MATH.
Vol. 67, No. 6, pp. 1523–1547

P MATRIX PROPERTIES, INJECTIVITY, AND STABILITY IN CHEMICAL REACTION SYSTEMS*

MURAD BANAJI†, PETE DONNELL†, AND STEPHEN BAIGENT‡

Abstract. In this paper we examine matrices which arise naturally as Jacobians in chemical dynamics. We are particularly interested in when these Jacobians are P matrices (up to a sign change), ensuring certain bounds on their eigenvalues, precluding certain behavior such as multiple equilibria, and sometimes implying stability. We first explore reaction systems and derive results which provide a deep connection between system structure and the P matrix property. We then examine a class of systems consisting of reactions coupled to an external rate-dependent negative feedback process and characterize conditions which ensure that the P matrix property survives the negative feedback. The techniques presented are applied to examples published in the mathematical and biological literature.

Key words. chemical reactions, P matrices, injectivity, stability, mass action

AMS subject classifications. 80A30, 15A48, 34D30

DOI. 10.1137/060673412

1. Introduction.

In this paper we will study chemical reaction systems and systems derived from them. Chemical reaction systems have Jacobians with more structure than those of arbitrary dynamical systems. Under mild assumptions we derive a condition on the reaction structure which ensures that a reaction system has Jacobians in a particular class, $P_0^{(-)}$ matrices, to be defined below. This condition is algorithmically easy to check, and immediately implies the absence of multiple equilibria as long as there are appropriate outflow conditions. A weaker condition is then derived specifically for mass action reaction systems, which ensures that they have Jacobians in this class and hence, under appropriate outflow conditions, cannot have multiple equilibria. These conditions are shown to be not only sufficient to preclude multiple equilibria, but also necessary to ensure that the Jacobians can never be singular. Finally a class of systems of particular importance in biochemistry is examined. These systems involve reactions interacting with some external quantity giving rise to a negative feedback process. Necessary and sufficient conditions are derived which ensure that the P matrix properties of the system without feedback persist with the feedback.

2. Basic material.

We start with some basic definitions and observations.

2.1. Chemical reaction systems.

A chemical reaction system in which n reactants participate in m reactions has dynamics governed by the ordinary differential equation

$$\dot{x} = Sv(x). \tag{2.1}$$

*Received by the editors October 26, 2006; accepted for publication (in revised form) May 8, 2007; published electronically September 7, 2007.
http://www.siam.org/journals/siap/67-6/67341.html
†Department of Medical Physics and Bioengineering, University College London, Gower Street, WC1E 6BT London, United Kingdom (m.banaji@ucl.ac.uk, p.donnell@ucl.ac.uk). These authors were funded by an EPSRC/MRC grant to the MIAS IRC (GR/N14248/01).
‡Department of Mathematics, University College London, Gower Street, WC1E 6BT London, United Kingdom (s.baigent@ucl.ac.uk).

Here $x = [x_1, \ldots, x_n]^T$ is the nonnegative n-vector of reactant concentrations, $v = [v_1, \ldots, v_m]^T$ is the m-vector of reaction rates, and S is the $n \times m$ stoichiometric matrix. Equation (2.1) defines a dynamical system on $\mathbb{R}^n_+$ (the nonnegative orthant in $\mathbb{R}^n$). The entries in S are constants—generally integers—with $|S_{ij}|$ describing how many molecules of substrate i are involved in reaction j. The sign of S_{ij} reflects an arbitrary choice of direction for the reaction, with no implication of reversibility or irreversibility. We will generally assume that substrates occur only on one side of a reaction (more on this later), and if $S_{ij} < 0$, we will say that substrate i occurs on the "left-hand side" of reaction j, and on the "right-hand side" if $S_{ij} > 0$.

The same form (2.1) can represent either a closed reaction system, where there is no inflow or outflow of reactants, or an open system. For an open system we simply allow some of the reactions to have empty left- or right-hand sides. We will refer to reactions not involving any inflow or outflow as "true" reactions. S describes a linear mapping between the reaction rates and the time derivatives of the concentrations, and any steady states of (2.1) must correspond to reaction rates lying in the kernel of S. Thus a nontrivial kernel means that there are steady states corresponding to nonzero reaction rates.

The $m \times n$ matrix $V(x)$ defined by $V_{ij}(x) \equiv \frac{\partial v_i}{\partial x_j}$ describes the dependence of the reaction rates on the concentrations. For later notational convenience we will write V instead of $V(x)$. The Jacobian of (2.1) is then just SV.

To make progress, we need to narrow the class of reactions a little. We call a reaction system *nonautocatalytic* (*NAC*) if the stoichiometric matrix S and the matrix V^T have opposite sign structures in the following sense: $S_{ij}V_{ji} \leq 0$ for all i and j, and $S_{ij} = 0 \Rightarrow V_{ji} = 0$. These assumptions are quite general—they mean that if a substrate is used up (created) in a reaction, then increasing the concentration of this substrate, while holding all others constant, cannot cause the reaction rate to decrease (increase). Further, if a substrate does not participate in a reaction, then it is not allowed to influence the reaction rate. As we allow $S_{ij}V_{ji} = 0$, even when $S_{ij} \neq 0$, irreversible reactions are implicitly allowed by this definition.

The assumption that the system is NAC holds for mass action systems, Michaelis–Menten systems, etc., provided that a reactant occurs only on one side of a reaction. It is possible to violate this condition, for example with reactions such as $\mathrm{A}+\mathrm{B} \rightleftharpoons 2\mathrm{A}$, where perhaps for small concentrations of A net flux is to the right, while for large concentrations it is to the left. Sometimes, in practice, such reactions actually represent the amalgamation of several NAC reactions. For example, the above system might actually represent $\mathrm{A} + \mathrm{B} \rightleftharpoons \mathrm{C}$, $\mathrm{C} \rightleftharpoons 2\mathrm{A}$, where C is some short-lived intermediate complex. If a reaction can be rewritten in this way, then it becomes amenable to the analysis presented here.

Most results in this paper are independent of the functional forms chosen for the reaction dynamics, apart from the assumption that reactions are NAC, as just described. However, some of the results which motivated this work are those of Craciun and Feinberg [5, 6] on the possibility of multiple equilibria in *mass action systems*, and the techniques they present to deduce the absence of multiple equilibria from the reaction network structure alone. Since we have included some results on mass action systems, we define these here. Let ν_j be the set of indices of the reactants on the left-hand side of the jth reaction, and ρ_j be the set of indices of the reactants on the right-hand side of the jth reaction. Further, let L_{ij} be the number of molecules of substrate i occurring on the left-hand side of the jth reaction, and R_{ij} be the number of molecules of substrate i occurring on the right-hand side of the jth reaction. Then,

for a mass action system, the reaction rate v_j for the jth reaction takes the form

$$v_j = k_j \prod_{i \in \nu_j} x_i^{L_{ij}} - k_{-j} \prod_{i \in \rho_j} x_i^{R_{ij}},$$

where k_j and k_{-j} are nonnegative constants, known as the forward and backward rate constant for the jth reaction. When the reaction is NAC, this can be rewritten in terms of entries in the stoichiometric matrix to get

$$v_j = k_j \prod_{i \in \nu_j} x_i^{-S_{ij}} - k_{-j} \prod_{i \in \rho_j} x_i^{S_{ij}}.$$

We can clearly write a single reversible reaction as two irreversible reactions.

2.2. *P* matrices and related classes. For some matrix A, $A(\alpha|\gamma)$ will refer to the submatrix of A with rows indexed by the set α and columns indexed by the set γ. A *principal submatrix* of A is a submatrix containing columns and rows from the same index set, i.e., of the form $A(\alpha|\alpha)$, which we will abbreviate to $A(\alpha)$. A *minor* is the determinant of a square submatrix. If $A(\alpha|\gamma)$ is a square submatrix of A (i.e., $|\alpha| = |\gamma|$), then $A[\alpha|\gamma]$ will refer to the corresponding minor, i.e., $A[\alpha|\gamma] = \det(A(\alpha|\gamma))$. A *principal minor* of a matrix is the determinant of a principal submatrix. $A[\alpha]$ will refer to the principal minor corresponding to submatrix $A(\alpha)$.

P matrices are square matrices all of whose principal minors are positive. They are nonsingular, and their eigenvalues are excluded from a certain wedge around the negative real axis [15]. If $-A$ is a P matrix, then we will say that A is a $P^{(-)}$ matrix. These matrices were originally called $N-P$ matrices in [17]. Throughout this paper, when A is a matrix such that $-A$ belongs to some class $\mathcal{C}$, then we will say that A belongs to the class $\mathcal{C}^{(-)}$. If A is a $P^{(-)}$ matrix, this means that each $k \times k$ principal minor of A has sign $(-1)^k$. The problem of checking whether a given matrix is a P matrix is in general NP hard [19].

Another important characterization of P matrices is that a matrix A is a P matrix iff for any nonzero vector y there is some index i such that $y_i(Ay)_i > 0$ [3]. It follows immediately that a matrix A is a $P^{(-)}$ matrix iff for any nonzero vector y there is some index i such that $y_i(Ay)_i < 0$. In other words a $P^{(-)}$ matrix maps each nonzero vector y out of any orthants in which it lies. (As orthants share boundaries, y may lie in several orthants at once.)

P matrices contain other important classes of matrices, such as positive definite matrices and also so-called nonsingular M matrices. As these will be mentioned again later, we define them here. Z matrices are square matrices all of whose off-diagonal entries are less than or equal to zero. Nonsingular M matrices are precisely those matrices which are both Z matrices and P matrices, i.e., matrices whose off-diagonal elements are nonpositive and all of whose principal minors are positive. Using the notational convention defined above, $M^{(-)}$ matrices are matrices which are both $Z^{(-)}$ matrices and $P^{(-)}$ matrices.

A related class of matrices are P_0 matrices consisting of the closure of the set of P matrices. These are matrices all of whose principal minors are nonnegative [14]. Similarly A is a $P_0^{(-)}$ matrix if $-A$ is a P_0 matrix. A matrix A is a P_0 matrix iff for any nonzero vector y there is some index i such that $y_i(Ay)_i \geq 0$, and similarly it is a $P_0^{(-)}$ matrix iff for any nonzero vector y there is some index i such that $y_i(Ay)_i \leq 0$. By definition, P_0 and $P_0^{(-)}$ matrices can be singular.

2.3. Implications of a $P^{(-)}$ Jacobian: Injectivity and stability. In the work of Craciun and Feinberg [5, 6] global injectivity, and hence the absence of multiple equilibria, follows from the nonsingularity of the Jacobian. This is not true for general functions—it is well known that nonsingularity of the Jacobian alone does not imply global injectivity of arbitrary polynomial functions [18]. In this direction there are several results connecting properties of functions with injectivity. A well-known theorem of Hadamard [12] states that nonsingularity of the Jacobian ensures global injectivity, provided that the function is *proper*—i.e., the preimage of any compact set is compact. Recent elegant work such as that in [9] and [20] provides conditions (not all spectral) which ensure that a function is globally injective.

Regarding P matrices, there is a result stating that if the Jacobian of a function is a P matrix (or equivalently a $P^{(-)}$ matrix), this guarantees injectivity of the function on any rectangular region of $\mathbb{R}^n$ [10]. The result for all of $\mathbb{R}^n$ also follows from the geometric fact mentioned in section 2.2 that $P^{(-)}$ matrices map vectors out of the orthants in which they lie. Thus, for a fixed nonzero vector y, every $P^{(-)}$ matrix must rotate y by at least some angle $\theta > \theta_y > 0$, where θ_y is the infimum of the angular distance from y to an orthant to which y does not belong; thus for any unit vector y and any set of $P^{(-)}$ matrices $A(x)$,

$$\sup_x \left\langle y, \frac{A(x)\,y}{|A(x)\,y|} \right\rangle < \cos\theta_y < 1.$$

From Theorem 2 in [20], this condition on the Jacobian guarantees global injectivity of the function.

[10] also contains the following strengthened result, which weakens the condition needed for injectivity: If the Jacobian of a function is a nonsingular P_0 matrix (termed a "weak P matrix" in this reference), this guarantees injectivity of the function on any rectangular region of $\mathbb{R}^n$. The result obviously holds for a nonsingular $P_0^{(-)}$ matrix as well.

While the ruling out of multiple equilibria is the first and perhaps most important consequence of finding that a particular dynamical system gives rise to $P^{(-)}$ matrix Jacobians, sometimes stronger conclusions can be drawn. In particular, if a matrix J is a $P^{(-)}$ matrix, then Hurwitz stability of J may follow from additional observations. We list three of these:

1. If J is similar to a symmetric matrix, and thus has real eigenvalues, then it must be Hurwitz stable.
2. If all off-diagonal elements of J are nonnegative, then it is in fact a nonsingular $M^{(-)}$ matrix [3] and hence Hurwitz stable.
3. A weaker condition is when J is "sign-symmetric," meaning that all symmetrically placed pairs of minors have the same sign: Then it is stable because sign-symmetric $P^{(-)}$ matrices are Hurwitz stable [14]. Certain physical assumptions can give rise to Jacobians which are sign-symmetric.

In this paper we will refer to a reaction system whose Jacobians are always $P^{(-)}$ matrices as $P^{(-)}$ systems, and ones whose Jacobians are always $P_0^{(-)}$ matrices as $P_0^{(-)}$ systems.

2.4. Rate-dependent negative feedback processes. The assumption that a reaction is NAC means, roughly speaking, that every substrate interacts with the reactions in which it participates in the following way: If it is produced by a reaction, then it inhibits the reaction. If it is used up by a reaction, then it activates the

reaction. Physically any scalar quantity ψ which behaves like this participates in *rate-dependent negative feedback*, and adding such a quantity to a system adds a rate-dependent negative feedback process to the system. Although ψ might be the concentration of a chemical, this need not be the case—for example, ψ may take negative values. In an example of biological importance discussed in [1] and used to illustrate our results below, ψ is in fact a chemical and electrical *gradient* with which some of the reactions interact because they pump material across a membrane. This is a frequent occurrence in biochemistry: Quite generally where reactions involve the build-up of gradients between compartments, we get such systems.

Adding a rate-dependent negative feedback process, whether a reactant or not, to a reaction system involves choosing two vectors $x_1, x_2 \in \mathbb{R}^m$ and adding a row x_1^T to S and a column x_2 to V to get augmented versions, S_{aug} and V_{aug}, of these matrices. The negative feedback assumption means that x_1 and x_2 lie in opposite orthants, so that $x_{1,i}x_{2,i} \leq 0$. In general, if if $x_{1,i} = 0$, then $x_{2,i} = 0$, but it is convenient to ignore this and ask the more general question: Given that SV is a $P^{(-)}$ matrix, when will $S_{aug}V_{aug}$ be a $P_0^{(-)}$ matrix for all possible $x_1, x_2 \in \mathbb{R}^m$ lying in (specified) opposite orthants? Given particular orthants, it is possible to state necessary and sufficient conditions on S and V which answer this question, and with appropriate outflow conditions to replace $P_0^{(-)}$ with $P^{(-)}$ in the above statement.

3. $P^{(-)}$ matrices and general reaction systems. We now examine the close connection between $P^{(-)}$ matrices and reaction systems of the form (2.1). After some preliminaries we present a structural result giving a sufficient condition on the stoichiometric matrix S, which will ensure that the Jacobian will be a $P_0^{(-)}$ matrix. In a sense to be made precise this condition is also a necessary condition.

We need some definitions first. A real matrix S determines a *qualitative class* [4] of all matrices with the same sign pattern as S, which we will refer to as $\mathcal{Q}(S)$. It is helpful to think of $\mathcal{Q}(S)$ as a matrix with entries consisting of zeroes and variables of fixed sign, and $\det(\mathcal{Q}(S))$ is then a polynomial in these variables. If $\det(\mathcal{Q}(S))$ is not identically zero, then it is a sum of monomials, each of which is either positive or negative. It also makes sense to refer to $\overline{\mathcal{Q}(S)}$ as the closure of $\mathcal{Q}(S)$ (regarded as a set of matrices), and $\det(\overline{\mathcal{Q}(S)})$ as the same polynomial as $\det(\mathcal{Q}(S))$, with variables now allowed to take the value zero. In this terminology a reaction system is NAC if $V \in \overline{\mathcal{Q}(-S^T)}$.

A square matrix is *sign-nonsingular* (*SNS*) [4] if the sign of its determinant is nonzero and can be determined from the signs of its entries. In other words, it is SNS if the sign of the determinant is the same for every matrix in its qualitative class. For example, any 2-by-2 matrix with a single negative, positive, or zero entry is SNS. On the other hand, a 2-by-2 matrix with two positive and two negative entries is not SNS. If any square matrix T is SNS, then it makes sense to talk about $\text{sign}(\det(\mathcal{Q}(T)))$.

A (not necessarily square) matrix S will be termed *strongly sign determined* (*SSD*) if all square submatrices of S are either SNS or singular. SSD matrices intersect various classes of matrices discussed in [4], for example the so-called totally L-matrices and the S^2NS matrices, and the SSD property is algorithmically quick and easy to check. Some results concerning SSD matrices are collected in Appendix A. These properties show among other things that alternative notational choices in chemical dynamics—for example, the choice to represent one reversible reaction as two irreversible ones, which side of a reaction to consider as the left-hand side, how to order the set of substrates, or how to order the set of reactions—never change whether the stoichiometric matrix is SSD or not.

For the proofs which follow in this section, it is convenient to set up the following notational conventions. S will always refer to some particular, but unspecified, stoichiometric matrix. Since we are interested in NAC, but otherwise unspecified, reaction systems in this section, given a matrix S, it is convenient for V to refer to the closure of a whole class $\overline{\mathcal{Q}(-S^T)}$. Similarly $V(\gamma|\alpha)$ will refer to $\overline{\mathcal{Q}(-S(\alpha|\gamma))}$, and $V[\gamma|\alpha]$ will refer to the polynomial $\det(\overline{\mathcal{Q}(-S(\alpha|\gamma))})$. If we refer to a "choice of V," then this means some particular matrix in $\overline{\mathcal{Q}(-S^T)}$. Objects defined as products will take the appropriate meanings; for example, an object such as $S[\alpha|\gamma]V[\gamma|\alpha]$ is again a polynomial.

It helps to note a few obvious, but important, *preliminaries* about SSD matrices:

1. If a matrix S is SSD, then so is $-S$. In particular, given any square submatrix $S(\alpha|\gamma)$ which is SNS,

$$\operatorname{sign}(\det(-S(\alpha|\gamma))) = (-1)^{|\alpha|}\operatorname{sign}(\det(S(\alpha|\gamma))).$$

   On the other hand, if $S(\alpha|\gamma)$ is singular, then so is $-S(\alpha|\gamma)$.
2. If a stoichiometric matrix S is not SSD, then there is some square submatrix $S(\alpha|\gamma)$ such that $\det(\mathcal{Q}(S(\alpha|\gamma)))$—and hence $V[\gamma|\alpha]$—contains both a positive and a negative term.

We can now state our first theorem.

THEOREM 3.1. *If the stoichiometric matrix S of an NAC reaction system is SSD, then the Jacobian $J = SV$ is a $P_0^{(-)}$ matrix.*

Proof. Let $J[\alpha]$ be the principal minor of J corresponding to the submatrix with rows and columns indexed by a set $\alpha \subset \{1, \dots, n\}$. By the Cauchy–Binet formula (see [11], for example) we get

$$J[\alpha] = (SV)[\alpha] = \sum_{\substack{\gamma \subset \{1,\dots,m\} \\ |\gamma| = |\alpha|}} S[\alpha|\gamma]V[\gamma|\alpha].$$

The sum is over all subsets of $\{1, \dots, m\}$ of size $|\alpha|$, if any such subsets exist. Since the reaction system is NAC and S is SSD, for each γ, either $S[\alpha|\gamma]$ is zero or $S(\alpha|\gamma)$ is SNS, in which case

$$\operatorname{sign}(V[\gamma|\alpha]) = (-1)^{|\alpha|}\operatorname{sign}(S[\alpha|\gamma])$$

by preliminary 1 above. So $J[\alpha]$ is a sum of terms each of which is either zero or has sign $(-1)^{|\alpha|}$, and thus either $J[\alpha] = 0$ or $\operatorname{sign}(J[\alpha]) = (-1)^{|\alpha|}$. ∎

A natural question which arises is whether there is any kind of converse to Theorem 3.1 or, equivalently, whether there could be a weaker condition on the stoichiometric matrix which would still always ensure a $P_0^{(-)}$ Jacobian. The answer, provided in the next theorem, is that there is no weaker condition guaranteeing a $P_0^{(-)}$ Jacobian.

THEOREM 3.2. *Assume that the stoichiometric matrix S of an NAC system is not SSD. Then there is some choice of V for which SV is not a $P_0^{(-)}$ matrix.*

Proof. Since S is not SSD, there are sets $\alpha_0 \subset \{1, \dots, n\}$, $\gamma_0 \subset \{1, \dots, m\}$ with $|\alpha_0| = |\gamma_0|$ such that $S(\alpha_0|\gamma_0)$ is neither SNS nor singular. Consider

$$J[\alpha_0] = \sum_{\substack{\gamma \subset \{1,\dots,m\} \\ |\gamma| = |\alpha_0|}} S[\alpha_0|\gamma]V[\gamma|\alpha_0].$$

Since $S(\alpha_0|\gamma_0)$ is not SNS, $V[\gamma_0|\alpha_0]$ contains a term t such that $S[\alpha_0|\gamma_0]t$ is of the "wrong" sign: $(-1)^{|\alpha_0|+1}$. This follows because, as noted in preliminary 2 earlier, since $S(\alpha_0|\gamma_0)$ is not SNS, $V[\gamma_0|\alpha_0]$ contains both positive and negative terms. However, t is just a term in the determinant of a submatrix of V, i.e., a product of entries of V. Set all entries in V which do not figure in t to 0. Since determinants are homogeneous polynomials in the entries of a matrix, and since no entry has power higher than 1, all terms in $J[\alpha_0]$ other than $S[\alpha_0|\gamma_0]t$ become zero, so that $J[\alpha_0] = S[\alpha_0|\gamma_0]t$, which has sign $(-1)^{|\alpha_0|+1}$. Hence J is not a $P_0^{(-)}$ matrix. By continuity, since the set of matrices which are not $P_0^{(-)}$ matrices is open, the argument still holds if entries in V not occurring in t are sufficiently small but nonzero. □

Incidentally we could phrase the above two results together as the following corollary, possibly of broad interest.

COROLLARY 3.3. *Consider an $n \times m$ matrix A. Then A is SSD iff AB is a P_0 matrix for every $m \times n$ matrix B which satisfies $A_{ij}B_{ji} \geq 0$ and $A_{ij} = 0 \Rightarrow B_{ji} = 0$.*

Proof. The proof is immediate from the previous two results. □

Although the discussion so far has been of $P_0^{(-)}$ matrices, it is clear from the proofs that if, in addition to S being SSD, for $\alpha = \{1, \ldots, n\}$ there is some γ such that $S[\alpha|\gamma]$ and $V[\gamma|\alpha]$ are both nonzero, then J is in fact nonsingular $P_0^{(-)}$, and the function is injective. And if, for every α, there is some γ such that $S[\alpha|\gamma]$ and $V[\gamma|\alpha]$ are both nonzero, then J is in fact a $P^{(-)}$ matrix (and injective). This often arises in practice because there are inflow and outflow processes contributing terms on the diagonal of SV. For example, continuous flow stirred tank reactors (CFSTRs) as presented in [5] have properties which ensure that for nonzero flow rate any Jacobian which is a $P_0^{(-)}$ matrix is in fact a $P^{(-)}$ matrix. Using S to refer to the stoichiometric matrix of the "true" reactions in a CFSTR (excluding the inflow and outflow processes), a CFSTR system can be written as

$$\dot{x} = q(x_{in} - x) + Sv(x), \tag{3.1}$$

where q is a positive scalar representing the flow rate through the reactor and x_{in} is a nonnegative vector representing the "feed" concentration. We then have the following result.

THEOREM 3.4. *Assume that all the reactions in a CFSTR are NAC. If the stoichiometric matrix S is SSD, then the Jacobian of the system is a $P^{(-)}$ matrix.*

Proof. The full stoichiometric matrix S_f of a CFSTR system can be written in block form,

$$S_f = [S| -I_n],$$

where S is the matrix of true reactions and I_n is the $n \times n$ unit matrix. Similarly define V_f by

$$V_f = \left[\frac{V}{qI_n}\right].$$

The Jacobian of the system is

$$J \equiv S_f V_f = -qI + SV,$$

where I is the identity matrix. Since the reactions are NAC and S is SSD this means, by Theorem 3.1, that SV is a $P_0^{(-)}$ matrix. As mentioned in section 2.2, a matrix A

is a P_0 matrix iff for any nonzero vector y there is some index i such that $y_i(Ay)_i \geq 0$, and similarly it is a P matrix iff for any nonzero vector y there is some index i such that $y_i(Ay)_i > 0$. So any P_0 matrix plus a positive diagonal matrix is a P matrix. It follows that a $P_0^{(-)}$ matrix plus a negative diagonal matrix is a $P^{(-)}$ matrix. Thus J is a $P^{(-)}$ matrix. □

Combined with the result of Gale and Nikaido [10], this can be stated as the following corollary.

COROLLARY 3.5. *If the reactions in a CFSTR are NAC, and the stoichiometric matrix S is SSD, then the system does not admit multiple equilibria.*

This result is independent of the nature of the reactions (mass action, Michaelis–Menten, etc.).

For CFSTR systems the result presented in Theorem 3.2 can be strengthened. If the stoichiometric matrix of true reactions in a CFSTR system is not SSD, and hence the Jacobian can fail to be a $P^{(-)}$ matrix, then it can in fact be singular.

THEOREM 3.6. *Assume that all the reactions in a CFSTR are NAC, and that the stoichiometric matrix of true reactions, S, is not SSD. Then there is some choice of entries in V for which* $\det(J)$ *has sign* $(-1)^{n+1}$ *(i.e., the "wrong" sign).*

Proof. The result follows as long as there is a term of the wrong sign in the expansion of the determinant, and this term can be made to dominate all other terms in the expansion.

As in the proof of Theorem 3.2, when S is not SSD, this implies the existence of sets $\alpha_0 \subset \{1, \ldots, n\}$, $\gamma_0 \subset \{1, \ldots, m\}$ with $|\alpha_0| = |\gamma_0|$ such that $V[\gamma_0|\alpha_0]$ contains a term t such that $S[\alpha_0|\gamma_0]t$ has sign $(-1)^{|\alpha_0|+1}$.

Let S_f and V_f be defined as in the proof of Theorem 3.4. The structure of S_f and V_f means that there is a term in $\det(S_f V_f)$ of the form $(-q)^{n-|\alpha_0|}S[\alpha_0|\gamma_0]t$, which is clearly of sign $(-1)^{n+1}$. As the determinant of any submatrix of V_f is a homogeneous polynomial in the entries of V_f, and no entry from V can occur more than once in any term, setting all entries in V other than those which occur in t to zero ensures that

$$\det(S_f V_f) = (-q)^{n-|\alpha_0|}S[\alpha_0|\gamma_0]t \ + \ \text{higher order terms in } q.$$

Choosing any fixed values for entries in t, then for small enough q, the lowest order term $(-q)^{n-|\alpha_0|}S[\alpha_0|\gamma_0]t$ is the dominant term in this expression, and hence $\det(S_f V_f)$ has sign $(-1)^{n+1}$. As in the proof of Theorem 3.2, by continuity, the argument remains true for small nonzero entries in V_f. □

This last theorem is more important than it may at first seem. It implies that if S is not SSD, then the Jacobian can be made singular by choosing entries in V appropriately. Thus finding that a particular reaction system has a stoichiometric matrix which is SSD is a *necessary* condition to ensure that under arbitrary choice of dynamics the Jacobian of the CFSTR system can never be singular.

The astute reader will have noticed that combination of the previous theorems implies that, for a CFSTR system, nonsingularity of the Jacobian (for all entries in V) is equivalent to injectivity of the system. This implies that when checking whether a system is necessarily injective, rather than checking whether S is SSD, one could instead check whether all $n \times n$ submatrices of S_f are either SNS or singular. Although at first glance the second strategy appears easier, the two problems are computationally equivalent, as computing the determinants of all $n \times n$ submatrices of S_f requires computation of the determinants of all square submatrices of S.

4. $P^{(-)}$ matrices and mass action systems. In this section we present some results on mass action systems. It is possible to prove stronger results about mass action systems than arbitrary reaction systems because the matrix V has additional structure beyond its sign structure. Our concern now is with the question of when a reaction system, as a result of its structure combined with the assumption of mass action dynamics, generates a $P_0^{(-)}$ matrix Jacobian (or, in the case of CFSTR systems, a $P^{(-)}$ matrix Jacobian). Of course, if a substrate never occurs on both sides of any reaction, then the mass action form guarantees that all reactions are NAC, and so if the stoichiometric matrix S is SSD, this will ensure a $P_0^{(-)}$ Jacobian. We show, however, that in the case of mass action systems it is possible to weaken the condition that S must be SSD and still get a $P_0^{(-)}$ Jacobian.

It is important at the outset to highlight the close relationship between results in this section and results in [5]. The techniques given in [5] for confirming whether a reaction system is injective are more general than ours in that they apply to autocatalytic reactions as well. We are unable to make claims about injectivity of autocatalytic reactions using our techniques because the stoichiometric matrix "loses information" about reactions which have the same substrate occurring on both sides of the equation—it encodes only net production or loss of a substrate in a reaction, rather than absolute quantities on each side of a reaction. There is some overlap in our methods of proof, although there are also important differences. We will return to this theme at the end of the section.

To formulate the results to follow we need to note that any mass action system can be written as a system of irreversible reactions by considering any reversible reaction as two irreversible reactions. From Lemma A.2 in Appendix A, rewriting the system in this way does not affect whether the stoichiometric matrix is SSD.

We now define a property of stoichiometric matrices weaker than the property of being SSD. Given a matrix S, define S_- to be the matrix S with all positive entries replaced with zeroes. Let a constant matrix S be *weakly sign determined* (*WSD*) if every square submatrix $\tilde{S}$ of S satisfies $\det(\tilde{S})\det(\tilde{S}_-) \geq 0$. In Lemma A.3 of Appendix A it is shown that every SSD matrix is WSD. The two are not equivalent, however—for example, the matrix

$$\tilde{S} = \begin{bmatrix} 1 & -1 \\ -2 & 1 \end{bmatrix}$$

is neither SNS nor singular, but it does satisfy $\det(\tilde{S})\det(\tilde{S}_-) \geq 0$. Results in Appendix A also show that the choice of how to order the set of substrates or reactions does not affect whether the stoichiometric matrix is WSD or not. However, as we shall see later, the choice to represent one reversible reaction as two irreversible ones *can* affect whether the stoichiometric matrix is WSD or not.

We can now restate Theorem 3.1 for mass action systems.

THEOREM 4.1. *Consider the stoichiometric matrix S of an NAC mass action reaction system written as a system of irreversible reactions. If S is WSD, then the Jacobian J is a $P_0^{(-)}$ matrix.*

Proof. The reaction rate for the ith reaction takes the form

$$v_i = k_i \prod_{j \in \nu_i} x_j^{-S_{ji}},$$

where k_i is the rate constant for the ith reaction and ν_i is the set of indices of the reactants on the left-hand side of the ith reaction. Thus the entries in V take the

form

$$V_{ij} = \frac{\partial v_i}{\partial x_j} = \begin{cases} \frac{-S_{ji}}{x_j} v_i & (j \in \nu_i), \\ 0 & (j \notin \nu_i). \end{cases}$$

As above, define S_- to be the matrix S with all positive entries replaced with zeroes. Further, let D_x be the $n \times n$ positive diagonal matrix with entries $\frac{1}{x_j}$ on the diagonal (defined when $x_j > 0$ for all j). Finally, let D_v be the $m \times m$ positive diagonal matrix with entries v_i on the diagonal. With this notation the matrix V can be written

$$V = -D_v S_-^T D_x$$

(again formally defined only when all $x_j > 0$, although of course V exists in the limit as well). Now consider an arbitrary minor of V, $V[\gamma|\alpha]$ with $\alpha \subset \{1, \ldots, n\}$ and $\gamma \subset \{1, \ldots, m\}$, and $|\alpha| = |\gamma|$. Application of the Cauchy–Binet formula combined with the fact that only principal minors of a diagonal matrix are nonzero gives

$$V[\gamma|\alpha] = (-1)^{|\alpha|} D_v[\gamma] S_-^T[\gamma|\alpha] D_x[\alpha].$$

Thus a principal minor of the Jacobian takes the form

$$\begin{aligned} J[\alpha] = (SV)[\alpha] &= \sum_{\substack{\gamma \subset \{1,\ldots,m\} \\ |\gamma| = |\alpha|}} S[\alpha|\gamma] V[\gamma|\alpha] \\ &= (-1)^{|\alpha|} \sum_{\substack{\gamma \subset \{1,\ldots,m\} \\ |\gamma| = |\alpha|}} S[\alpha|\gamma] D_v[\gamma] S_-^T[\gamma|\alpha] D_x[\alpha] \\ &= (-1)^{|\alpha|} D_x[\alpha] \sum_{\substack{\gamma \subset \{1,\ldots,m\} \\ |\gamma| = |\alpha|}} S[\alpha|\gamma] S_-[\alpha|\gamma] D_v[\gamma]. \end{aligned}$$

Since D_x and D_v are positive diagonal matrices, $D_x[\alpha]$ and $D_v[\gamma]$ are positive. Thus $J[\alpha]$ has sign $(-1)^{|\alpha|}$ or is zero, provided that every $S[\alpha|\gamma]$ and $S_-[\alpha|\gamma]$ have the same sign (or one of them is zero).

The argument presented above shows that if S is WSD, then the Jacobian of a mass action system is a $P_0^{(-)}$ matrix in the interior of the positive orthant. However, the set of $P_0^{(-)}$ matrices is closed, and since the Jacobian depends continuously on the values of x_i, it must be $P_0^{(-)}$ everywhere in the closed positive orthant. ∎

The following corollary is immediate.

COROLLARY 4.2. *Assume that all the reactions in a CFSTR are NAC mass action reactions. If the stoichiometric matrix S of the system written as a set of irreversible reactions is WSD, then the Jacobian of the system is a $P^{(-)}$ matrix.*

Proof. The proof is identical to that of Theorem 3.4: A $P_0^{(-)}$ matrix plus a negative diagonal matrix is a $P^{(-)}$ matrix. ∎

There is a kind of converse to Theorem 4.1 showing that the condition of being WSD is *necessary* to guarantee that the Jacobian of a mass action system will be a $P_0^{(-)}$ matrix.

THEOREM 4.3. *Assume that the stoichiometric matrix S of an NAC mass action system written as a set of irreversible reactions is not WSD. Then there is some choice of rate constants k_i for which SV is not a $P_0^{(-)}$ matrix.*

Proof. If S is not WSD, then $S[\alpha_0|\gamma_0]S_-[\alpha_0|\gamma_0] < 0$ for some $\alpha_0 \subset \{1,\dots,n\}$, $\gamma_0 \subset \{1,\dots,m\}$ with $|\alpha_0| = |\gamma_0|$. We have from above

$$J[\alpha_0] = (-1)^{|\alpha_0|} D_x[\alpha_0] \sum_{\substack{\gamma\subset\{1,\dots,m\}\\ |\gamma|=|\alpha_0|}} S[\alpha_0|\gamma]S_-[\alpha_0|\gamma]D_v[\gamma].$$

Since $D_v[\gamma] = \prod_{j\in\gamma} v_j$, choosing $k_j = 0$ for all $j \notin \gamma_0$ and $k_j \neq 0$ for all $j \in \gamma_0$ sets all $D_v[\gamma] = 0$ for $\gamma \not\subseteq \gamma_0$. So with this choice

$$J[\alpha_0] = (-1)^{|\alpha_0|} D_x[\alpha_0]S[\alpha_0|\gamma_0]S_-[\alpha_0|\gamma_0]D_v[\gamma_0],$$

which has sign $(-1)^{|\alpha_0|+1}$ everywhere in the interior of the positive orthant. By continuity, $J[\alpha_0]$ continues to have sign $(-1)^{|\alpha_0|+1}$ in some region of the positive orthant when k_j, $j \notin \gamma_0$, are small but nonzero. ∎

For mass action systems, the condition of being WSD is thus *necessary* to guarantee that the Jacobian will be a $P_0^{(-)}$ matrix. In fact, in the case of CFSTR mass action systems there is an analogue of the general result in Theorem 3.6: The property of S being WSD is necessary to guarantee that the Jacobian will be nonsingular.

THEOREM 4.4. *Assume that the stoichiometric matrix S of the true reactions in an NAC mass action CFSTR system written as a set of irreversible reactions is not WSD. Then there is some choice of flow rate q, rate constants k_i, and concentrations x_i for which* $\det(SV)$ *has sign* $(-1)^{n+1}$ *(i.e., the "wrong" sign).*

Proof. The proof is a little harder than the equivalent proof for general systems, but again, the result follows as long as there is a term of the wrong sign in the expansion of the determinant, and this term can be made to dominate all other terms in the expansion.

Since S is not WSD, $S[\alpha_0|\gamma_0]S_-[\alpha_0|\gamma_0] < 0$ for some sets $\alpha_0 \subset \{1,\dots,n\}$, $\gamma_0 \subset \{1,\dots,m\}$ with $|\alpha_0| = |\gamma_0|$.

The Jacobian $J = SV - qI$, and the determinant of the Jacobian is $\det(SV - qI)$. Expanding this, we get

$$\begin{aligned}
\det(SV - qI) &= \sum_{j=0}^{n} (-1)^j q^j \sum_{\substack{\alpha\subset\{1,\dots,n\}\\ |\alpha|=n-j}} SV[\alpha] \\
&= \sum_{j=0}^{n} (-1)^j q^j \sum_{\substack{\alpha\subset\{1,\dots,n\}\\ |\alpha|=n-j}} (-1)^{n-j} D_x[\alpha] \sum_{\substack{\gamma\subset\{1,\dots,m\}\\ |\gamma|=|\alpha|}} S[\alpha|\gamma]S_-[\alpha|\gamma]D_v[\gamma] \\
&= (-1)^n \sum_{j=0}^{n} q^j \sum_{\substack{\alpha\subset\{1,\dots,n\}\\ |\alpha|=n-j}} D_x[\alpha] \sum_{\substack{\gamma\subset\{1,\dots,m\}\\ |\gamma|=|\alpha|}} S[\alpha|\gamma]S_-[\alpha|\gamma]D_v[\gamma].
\end{aligned}$$

Setting all $k_i \notin \gamma_0$ equal to zero, we get

$$\begin{aligned}
\det(SV - qI) = (-1)^n q^{n-|\gamma_0|} D_v[\gamma_0] &\sum_{\substack{\alpha\subset\{1,\dots,n\}\\ |\alpha|=|\gamma_0|}} D_x[\alpha]S[\alpha|\gamma_0]S_-[\alpha|\gamma_0] \\
&+ \text{higher order terms in } q.
\end{aligned}$$

We know that $S[\alpha_0|\gamma_0]S_-[\alpha_0|\gamma_0] < 0$. Since $D_x[\alpha] = \prod_{i\in\alpha} x_i^{-1}$, by fixing values of x_i for $i \in \alpha_0$ and increasing the values of x_i for $i \notin \alpha_0$ we can make $D_x[\alpha_0]$ much

larger than $D_x[\alpha]$ for any $\alpha \neq \alpha_0$ in the sum above, thus ensuring that the term

$$D_v[\gamma_0]D_x[\alpha_0]S[\alpha_0|\gamma_0]S_-[\alpha_0|\gamma_0]$$

is the dominant term in the coefficient of $q^{n-|\gamma_0|}$ and thus that this coefficient has sign $(-1)^{n+1}$. (Note that increasing the values of $x_i \notin \alpha_0$ affects, but can never decrease, the size of $D_x[\alpha_0]D_v[\gamma_0]$.)

Once we have ensured that the coefficient of $q^{n-|\gamma_0|}$ has sign $(-1)^{n+1}$, we can choose q small so that the term of order $q^{n-|\gamma_0|}$ is the dominant term in $\det(SV - qI)$. Thus for small q, small $x_i \in \alpha_0$ (and all other x_i sufficiently large), large $k_i \in \gamma_0$ (and all other k_i sufficiently small) we can ensure that $\det(SV - qI)$ has sign $(-1)^{n+1}$. ∎

This final result shows that if S is not WSD, then for some choices of rate constants and flow rate the Jacobian of a CFSTR system will be singular. Thus the property of S being WSD is both sufficient and *necessary* to ensure that the Jacobian of an NAC mass action CFSTR system is always nonsingular. It is also sufficient and *necessary* to ensure that the Jacobian is always a $P^{(-)}$ matrix and hence that the system is injective. Together these facts imply that nonsingularity of the Jacobian of an NAC mass action CFSTR system is equivalent to injectivity for these systems. This theorem overlaps with Theorem 3.3 in [5]: Both theorems rely on the fact that for the polynomials which define the determinants in CFSTR systems positivity of the numerical coefficients is necessary to ensure positivity of the polynomial.

There are further close relationships between the theorems here and those in [5]. In Theorem 3.1 of [5] it is proved directly that mass action systems are injective iff their Jacobians are nonsingular for all positive values of the rate constants and concentrations. As just related, we come to the same conclusion for NAC systems via a different route: We have proved that the condition that S is WSD is equivalent both to injectivity and to nonsingularity of the Jacobian in the CFSTR case, and thus that these two are themselves equivalent. This in turn implies that the condition that the stoichiometric matrix of true reactions must be WSD and the requirement that the quantity in (3.4) of [5] must be positive are equivalent for NAC mass action systems embedded in a CFSTR.

One apparent difference between the results here and those in [5] lies in the fact that, in Theorem 3.2 of [5], only determinants of $n \times n$ submatrices of the *full* stoichiometric matrix are needed, whereas when checking the WSD condition we have to check all square submatrices of the stoichiometric matrix. However, this difference is only apparent, and the remark that we made about general systems applies again here: Checking whether S is WSD is computationally equivalent to checking whether all $n \times n$ submatrices T of $S_f = [S| - I_n]$ satisfy $\det(\tilde{T})\det(\tilde{T}_-) \geq 0$.

5. Examples. We present some examples to illustrate the theoretical points in the previous sections.

5.1. Examples from [5]. The phenomenon of S being SSD is more common than it might at first seem. We first examined the reaction system (1.1) in [5] and examples (i) to (viii) presented in Table 1.1 of that reference. Of these, examples (vi), (vii), and (viii) have reactants on both sides of the reactions, and are discussed below in section 5.4. Our analysis of the other examples is presented in Table 5.1. In all cases, we found that whether or not the system had the capacity for multiple equilibria corresponded precisely to whether or not the stoichiometric matrix was SSD. We can thus state for the three systems in which multiple equilibria were ruled out—system (1.1) and systems (ii) and (iv) in Table 1.1—that this remains true if we violate the mass action assumption.

TABLE 5.1

Behavior of some reaction systems presented in [5]. *In all the examples where the systems are WSD, the systems are also in fact SSD, and thus multiple equilibria are ruled out in a CFSTR under arbitrary dynamics.*

	Reaction system	SSD	WSD
(i)	$A + B \rightleftharpoons P$ $B + C \rightleftharpoons Q$ $C \rightleftharpoons 2A$	not SSD	not WSD
(ii)	$A + B \rightleftharpoons P$ $B + C \rightleftharpoons Q$ $C + D \rightleftharpoons R$ $D \rightleftharpoons 2A$	SSD	WSD
(iii)	$A + B \rightleftharpoons P$ $B + C \rightleftharpoons Q$ $C + D \rightleftharpoons R$ $D + E \rightleftharpoons S$ $E \rightleftharpoons 2A$	not SSD	not WSD
(iv)	$A + B \rightleftharpoons P$ $B + C \rightleftharpoons Q$ $C \rightleftharpoons A$	SSD	WSD
(v)	$A + B \rightleftharpoons F$ $A + C \rightleftharpoons G$ $C + D \rightleftharpoons B$ $C + E \rightleftharpoons D$	not SSD	not WSD
Ex. 1.1	$A + B \rightleftharpoons C$ $X \rightleftharpoons 2A + D$ $2A + D \rightleftharpoons Y$ $D \rightleftharpoons C + W$ $B + D \rightleftharpoons Z$	SSD	WSD

It is no surprise that the NAC systems which were proved to be injective in [5] proved to be WSD since, as shown in the previous section, the stoichiometric matrix being WSD is necessary for injectivity of the Jacobian for all values of the rate constants. What is surprising is that all of these examples turned out also to be SSD, and thus that the conclusions about these systems in [5] turn out to be more generally true.

5.2. Systems which are WSD but not SSD. Although the examples taken from [5] and presented in Table 5.1 are all either both SSD and WSD or neither, it is possible to construct examples of systems which are WSD but not SSD. Consider the reaction system

$$A + B \rightleftharpoons C, \qquad 2A + B \rightleftharpoons D,$$

which has stoichiometric matrix, in reversible and irreversible forms,

$$S_r = \begin{bmatrix} -1 & -2 \\ -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad S_{ir} = \begin{bmatrix} -1 & 1 & -2 & 2 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

It is quick to check that S_r (and hence S_{ir}) is not SSD. On the other hand S_{ir} is WSD. Thus if these reactions are embedded in a CFSTR, multiple equilibria can be ruled out as long as the dynamics are mass action dynamics, but not in the general case.

5.3. The reaction system as a reversible/irreversible system. To illustrate that it is essential to consider a system as a set of irreversible reactions when checking whether a stoichiometric matrix is WSD or not, consider the following reaction system:

$$2A \rightleftharpoons B + C, \qquad A \rightleftharpoons B,$$

which has stoichiometric matrix, in reversible and irreversible forms,

$$S_r = \begin{bmatrix} -2 & -1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}, \qquad S_{ir} = \begin{bmatrix} -2 & 2 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 0 & 0 \end{bmatrix}.$$

Here S_r is WSD, but S_{ir} is not. Thus examining S_r alone could give rise to the wrong conclusion that multiple equilibria can be ruled out in the mass action case.

This example also illustrates the importance of reversibility in the mass action case. Consider the above system with one reaction now irreversible:

$$2A \rightleftharpoons B + C, \qquad B \to A.$$

This has stoichiometric matrix, in irreversible form

$$S_{ir} = \begin{bmatrix} -2 & 2 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & 0 \end{bmatrix},$$

which is in fact WSD. Thus with mass action dynamics this system does not admit multiple equilibria when embedded in a CFSTR. It is perhaps surprising that if the reaction $B \to A$ were replaced with $A \to B$, then the system would no longer be WSD, and the conclusion would no longer hold. Instead, for certain choices of the rate constants, the system would cease to be injective, and multiple equilibria, while not guaranteed, can no longer be ruled out by this method.

5.4. Autocatalytic reactions. Consider the reactions in [5] of the form

$$m_1 A + m_2 B \rightleftharpoons (m_1 + m_2)A$$

for some positive integers m_1 and m_2. Recasting these as

$$m_1 A + m_2 B \rightleftharpoons C, \qquad C \rightleftharpoons (m_1 + m_2)A$$

and assuming NAC dynamics gives rise to stoichiometric matrices, in reversible and irreversible forms,

$$S_r = \begin{bmatrix} -m_1 & (m_1+m_2) \\ -m_2 & 0 \\ 1 & -1 \end{bmatrix}, \qquad S_{ir} = \begin{bmatrix} -m_1 & m_1 & (m_1+m_2) & -(m_1+m_2) \\ -m_2 & m_2 & 0 & 0 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

Barring the trivial possibilities that $m_1 = 0$ or $m_2 = 0$, S_r is never SSD and S_{ir} is never WSD. Thus multiple equilibria cannot be ruled out in general or for mass action systems. However, for mass action dynamics in the cases $m_1 = 1$, $m_2 = 1$ and $m_1 = 1$, $m_2 = 2$, it is known that multiple equilibria cannot exist [5], illustrating that singularity of the Jacobian is not sufficient to guarantee multiple equilibria. This is

because, although a function in some class may fail to be injective, the class may not allow this failure to occur near its zeroes.

In fact it is easy to show that when a reactant occurs on both sides of a reaction with different stoichiometries, and we rewrite the system as two NAC reactions with an intermediate complex, the system cannot be SSD or WSD. Consider the reaction system

$$nA + \cdots \rightleftharpoons C, \qquad C \rightleftharpoons mA + \cdots,$$

which might result from such a rewriting. Assume for definiteness that $m > n$. Then the irreversible stoichiometric matrix S_{ir} has a 2×2 submatrix of the form

$$T = \begin{bmatrix} -n & m \\ 1 & -1 \end{bmatrix},$$

which is clearly not SNS, not singular, and does not satisfy $\det(T)\det(T_-) \geq 0$ either.

5.5. Computational considerations. Although it is easy to write down algorithms to check whether a given matrix is SSD or WSD, the actual computation involves checking a large number of submatrices, and can be lengthy if the reaction network is large. Since large stoichiometric matrices are in general highly sparse, considerable speed-up can be achieved by using algorithms to identify submatrices which have (identically) zero determinant without actually attempting to compute the determinant. Similarly, intelligent algorithms should avoid recomputation of the determinants of matrices when they occur as submatrices in larger matrices.

Another technique which can speed up the classification of a matrix as SSD or WSD relies on the fact that it is possible to ignore all substrates which occur in only one reaction, as shown in Lemmas A.4–A.6 in Appendix A. This greatly shortens the calculations in many real examples. Consider example (i) in Table 5.1:

$$A + B \rightleftharpoons P, \qquad B + C \rightleftharpoons Q, \qquad C \rightleftharpoons 2A,$$

which has stoichiometric matrix, in reversible and irreversible forms,

$$S_r = \begin{bmatrix} -1 & 0 & 2 \\ -1 & -1 & 0 \\ 0 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad S_{ir} = \begin{bmatrix} -1 & 1 & 0 & 0 & 2 & -2 \\ -1 & 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & -1 & 1 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \end{bmatrix}.$$

Since P and Q each occur only in a single reaction, when checking whether the system is SSD and WSD, respectively, it suffices to check the reduced matrices

$$\tilde{S}_r = \begin{bmatrix} -1 & 0 & 2 \\ -1 & -1 & 0 \\ 0 & -1 & -1 \end{bmatrix}, \qquad \tilde{S}_{ir} = \begin{bmatrix} -1 & 1 & 0 & 0 & 2 & -2 \\ -1 & 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & -1 & 1 \end{bmatrix},$$

which considerably reduces the computational effort.

6. Rate-dependent negative feedback processes. Having seen that NAC reaction systems often give rise to Jacobians which are $P_0^{(-)}$ or $P^{(-)}$ matrices, and how this property is deeply associated with *reaction structure* rather than reaction details, we now examine the process of adding a rate-dependent negative feedback

process to a given system—i.e., adding a scalar quantity ψ (perhaps the concentration of another reactant) which inhibits or activates reactions according to whether it is produced or used up in them. We also allow ψ to be subject to an outflow/degradation process.

The full system becomes

$$\dot{x} = Sv(x,\psi), \qquad \dot{\psi} = C(v(x,\psi)) - L(\psi). \tag{6.1}$$

The function $C(v)$ represents the reaction-rate-dependent creation of ψ, while $L(\psi)$ represents its level-dependent outflow or degradation.

We define the following quantities:

1. $F \equiv \frac{\partial v}{\partial \psi}$. This m vector describes the dependence of the reaction rates on ψ.
2. $P \equiv \left(\frac{\partial C}{\partial v}\right)^T$. This m vector describes the way that the production of ψ depends on the reaction rates.
3. $u \equiv \frac{\partial L}{\partial \psi}$. This scalar describes the rate of decay of ψ.

The most general mathematical meaning of the negative feedback assumption is that the vectors F and P lie in opposite cones generated by some set of m orthogonal vectors. Thus for some orthogonal transformation U,

$$P \in K \equiv U(\mathbb{R}^m_+) \quad \text{and} \quad F \in -K.$$

The case generally encountered in examples is where K is a particular orthant so that U is a so-called signature matrix (a diagonal matrix with diagonal entries ± 1) and we know the signs of the elements of P and F, but not their values. In fact we will initially assume that $U = I$, the identity matrix, i.e., $P \in \mathbb{R}^m_+$ and $F \in \mathbb{R}^m_-$, showing later how the results can be extended to the general case.

The Jacobian of (6.1) is now the key object of interest. It can be written in block form:

$$J = \begin{bmatrix} SV & SF \\ P^T V & P^T F - u \end{bmatrix}. \tag{6.2}$$

We will prove all the results in this section by examining matrices (and submatrices) of the form J above. In order to discuss the negative feedback assumption, we adopt the following standard notation (see [3], for example): Given a vector $y \in \mathbb{R}^n$,

- $y \geq 0$ will mean that $y_i \geq 0$ for all i,
- $y > 0$ will mean that $y \geq 0$ and $y \neq 0$,
- $y \gg 0$ will mean that $y_i > 0$ for all i,
- $y \leq 0$, $y < 0$, and $y \ll 0$ will have analogous meanings.

With this notation, the negative feedback assumption can be rephrased as $F < 0$ and $P > 0$. We will assume that $u > 0$ and make claims about when J is a $P^{(-)}$ matrix—the extension to $u = 0$ and the $P_0^{(-)}$ case will be automatic, using the continuous dependence of determinants on the entries in a matrix.

Having discussed these preliminaries, we now ask the following question: Assuming that SV is a $P^{(-)}$ matrix, under what conditions will the Jacobian J in (6.2) remain a $P^{(-)}$ matrix for all values of $F < 0$, $P > 0$, and $u > 0$? The complete answer to this question is contained in Theorems 6.2 and 6.3.

In what is to follow, we will make use of the following formula for the determinant of a matrix.

LEMMA 6.1. *Let A be any matrix written in block form as follows:*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where A_{11} and A_{22} are square matrices. Assuming that A_{11} is nonsingular, then

$$\det(A) = \det(A_{22} - A_{21}A_{11}^{-1}A_{12})\det(A_{11}).$$

Proof. See, for example, [16, p. 46]. □

Note that if A_{22} is a scalar, the equation becomes

$$\det(A) = (A_{22} - A_{21}A_{11}^{-1}A_{12})\det(A_{11}).$$

We now state the basic theorem about the determinant of Jacobians of the form in (6.2), from which results on $P^{(-)}$ properties will follow easily.

THEOREM 6.2. *Consider a matrix J of the form in* (6.2). *Define $\mathcal{S} = \ker(S)$ and $\mathcal{V} = \ker(V^T)$. Assume that* $\det(SV)$ *has sign* $(-1)^n$. *Define $Z \equiv V(SV)^{-1}S$. Then the following statements are equivalent:*

(A) *Given any vector $w_1 \in \mathcal{S}$ satisfying $w_1 \not\geq 0$, we can find a vector $w_2 \in \mathcal{V}$ satisfying $w_2 > 0$ such that $\langle w_1, w_2 \rangle < 0$.*

(B) $\det(J)$ *has sign $(-1)^{n+1}$ for any choice of $F < 0$, $P > 0$, and $u > 0$.*

(C) *$I - Z$ is a nonnegative matrix.*

Before we begin the proof of the theorem we discuss a couple of the assumptions. The assumption that $\det(SV)$ is always of sign $(-1)^n$ implies that $\ker(V)$ and $\ker(S^T)$ consist only of 0. The condition that $\ker(S^T)$ consists only of 0 in turn means that there are no conserved quantities in the system [8], certainly true in the CFSTR case. In any case, where there are conserved quantities, the system can generally be redefined with some variables being eliminated to remove these.

A situation in which the theorem is trivially satisfied is when S and hence V are square matrices—i.e., there are the same number of substrates and reactions. Then the condition that SV is nonsingular implies that S and V are both nonsingular and hence both $\mathcal{S}$ and $\mathcal{V}$ consist only of zero, and there are no vectors $w_1 \in \mathcal{S}$ satisfying $w_1 \not\geq 0$. In this case the matrix Z is the identity. As mentioned earlier, if S is a nonsingular square matrix, then any equilibria correspond to all reaction rates being zero.

During the proof, we will see that condition (A) of the theorem has the following geometric interpretation: It means that the projection of any nonnegative vector $y > 0$ along $\mathcal{V}^T$ onto $\mathcal{S}$ is nonnegative. We remark that there is an important special case where condition (A) is immediately satisfied—this is when $\mathcal{S}$ is one dimensional, lying entirely in the nonnegative and nonpositive orthants, and $\mathcal{V}$ contains some strictly positive vectors. Then given $w_1 \in \mathcal{S}$, $w_1 \not\geq 0 \Rightarrow w_1 < 0$, and given any $w_2 \in \mathcal{V}$ satisfying $w_2 \gg 0$, we have $\langle w_1, w_2 \rangle < 0$. In general, however, where $\mathcal{S}$ can intersect other orthants, the existence of the vector w_2 will depend on the structures of $\mathcal{S}$ and $\mathcal{V}$.

Proof of Theorem 6.2. We show that both (A) and (B) are equivalent to (C), starting with (B) $\Leftrightarrow$ (C). Using Lemma 6.1, we get that

$$\begin{aligned}\det(J) &= (P^T F - u - P^T V(SV)^{-1}SF)\det(SV)\\ &= (P^T(I - Z)F - u)\det(SV),\end{aligned}$$

where I is the $m \times m$ unit matrix. Since $\det(SV)$ has sign $(-1)^n$, this means immediately that (6.2) will have determinant of sign $(-1)^{n+1}$ as long as $P^T(I-Z)F - u < 0$. This is true for all $u > 0$ iff $P^T(I - Z)F \leq 0$. This in turn is true for all $P > 0$, and $F < 0$ iff $(I - Z)$ is a nonnegative matrix (i.e., it leaves the nonnegative orthant invariant). Otherwise we can choose F and P appropriately so that $P^T(I - Z)F > 0$. Thus (B) $\Leftrightarrow$ (C).

We now show that (A) $\Rightarrow$ (C). It is easy to see that $Z^2 = Z$—i.e., Z is a projection. As SV is nonsingular, $\ker(Z) = \ker(S) = \mathcal{S}$ and $\ker(Z^T) = \ker(V^T) = \mathcal{V}$. Thus Z acts as a projection along $\mathcal{S}$ onto $\mathcal{V}^\perp$, and $I - Z$ projects along $\mathcal{V}^\perp$ onto $\mathcal{S}$.

Consider an arbitrary vector $y > 0$. Write $y = y_1 + y_2$, where $y_1 \equiv (I - Z)y \in \mathcal{S}$ and $y_2 \equiv Zy \in \mathcal{V}^\perp$. Now if $y_1 \not\geq 0$, then by assumption we can choose a vector $z \in \mathcal{V}$ satisfying $z > 0$ and $\langle z, y_1\rangle < 0$. But then $\langle z, y\rangle = \langle z, y_1\rangle < 0$, contradicting the fact that $z > 0$ and $y > 0$. So $y_1 \geq 0$. Thus $(I - Z)$ leaves the nonnegative orthant invariant and is a nonnegative matrix. Thus (A) $\Rightarrow$ (C).

Finally, (C) $\Rightarrow$ (A): If $(I - Z)$ is a nonnegative matrix, we show that, given any $y_1 \in \mathcal{S}$ satisfying $y_1 \not\geq 0$, there is a $z \in \mathcal{V}$ satisfying $z > 0$ such that $\langle z, y_1\rangle < 0$. Note that if $y_1 \not\geq 0$, then there is some vector $r \gg 0$ such that $\langle r, y_1\rangle < 0$. So

$$0 > \langle r, y_1\rangle = \langle r, (I - Z)y_1\rangle = \langle (I - Z^T)r, y_1\rangle.$$

Now note that $(I - Z^T)r > 0$ because $(I - Z^T) = (I - Z)^T$ is a nonnegative matrix and $r \gg 0$. Moreover, $(I - Z^T)r \in \mathcal{V}$ since $(I - Z^T)$ is a projection along $\mathcal{S}^\perp$ onto $\mathcal{V}$. So $z \equiv (I - Z^T)r$ is a positive vector in $\mathcal{V}$ which satisfies $\langle z, y_1\rangle < 0$. ∎

Theorem 6.2 leads immediately to the following.

THEOREM 6.3. *Let J be a matrix of the form defined in* (6.2). *Let α be some subset of $\{1, \ldots, n\}$, S^α be the matrix S with rows belonging to α deleted, and V^α the matrix V with columns belonging to α deleted. Define $\mathcal{S}^\alpha = \ker(S^\alpha)$ and $\mathcal{V}^\alpha = \ker((V^\alpha)^T)$.*

Assume that SV is a $P^{(-)}$ matrix. Then the following statements are equivalent:

(A) *For every $\alpha \subset \{1, \ldots, n\}$, given any vector $w_1 \in \mathcal{S}^\alpha$ satisfying $w_1 \not\geq 0$, we can find a vector $w_2 \in \mathcal{V}^\alpha$ satisfying $w_2 > 0$ such that $\langle w_1, w_2\rangle < 0$.*

(B) *J is a $P^{(-)}$ matrix for any choice of $F < 0$, $P > 0$, and $u > 0$.*

Proof. Since SV is a $P^{(-)}$ matrix, to prove that J is a $P^{(-)}$ matrix it suffices to treat all the principal submatrices of J obtained by deleting a set of rows/columns *not* including the final row and column. We show that for any $\alpha \subset \{1, \ldots, n\}$ the principal minor corresponding to the deletion of rows and columns from α has sign $(-1)^{n+1-|\alpha|}$.

In the trivial case where $\alpha = \{1, \ldots, n\}$, the principal submatrix corresponding to the removal of rows and columns from α is simply the scalar $P^T F - u$, which we know to be negative. In the case where $\alpha = \emptyset$, the principal submatrix is J itself. In general the principal submatrix corresponding to the removal of rows and columns from α is

$$J^\alpha \equiv \begin{bmatrix} S^\alpha V^\alpha & S^\alpha F \\ P^T V^\alpha & P^T F - u \end{bmatrix}.$$

$S^\alpha V^\alpha$ is a principal submatrix of SV, and since SV is a $P^{(-)}$ matrix, its determinant has sign $(-1)^{n-|\alpha|}$. J^α is of the form in (6.2), and to prove that $\det(J^\alpha)$ has sign $(-1)^{n+1-|\alpha|}$ it suffices by Theorem 6.2 that given any vector $w_1 \in \mathcal{S}^\alpha$ satisfying $w_1 \not\geq 0$, we can find a vector $w_2 \in \mathcal{V}^\alpha$ satisfying $w_2 > 0$ such that $\langle w_1, w_2\rangle < 0$.

The converse result follows because of the sufficiency of the condition set out in Theorem 6.2. ∎

Although the results above are about the $P^{(-)}$ case, they extend to the $P_0^{(-)}$ case, as seen in the next result.

COROLLARY 6.4. *Assume that the conditions of Theorem* 6.3 *are fulfilled and hence that the matrix J in* (6.2) *is a $P^{(-)}$ matrix for all values of $F < 0$, $P > 0$, and $u > 0$. Then J is a $P_0^{(-)}$ matrix for all values of $F \leq 0$, $P \geq 0$, and $u \geq 0$.*

Proof. Given any particular $\tilde{F} \leq 0$, $\tilde{P} \geq 0$, and $\tilde{u} \geq 0$ and J constructed using these, we can construct a sequence of $P^{(-)}$ matrices $\{J_i\}$ converging to J by choosing sequences $\{F_i\} < 0$, $\{P_i\} > 0$, and $\{u_i\} > 0$ converging to $\tilde{F}$, $\tilde{P}$, and $\tilde{u}$, respectively. Thus J lies in the closure of the $P^{(-)}$ matrices and must be a $P_0^{(-)}$ matrix. ∎

6.1. Extension to the general case. We now show briefly how the arguments in Theorems 6.2 and 6.3 extend to the general case where U is some orthogonal transformation, $K = U(\mathbb{R}^m_+)$, $P \in K$, and $F \in -K$. For arbitrary orthogonal U, the statement of Theorem 6.2 modifies to the following.

THEOREM 6.5. *Consider a matrix of the form J in* (6.2)*, and let U be any $m \times m$ orthogonal matrix. Define $K = U(\mathbb{R}^m_+)$, $\mathcal{S} = \ker(SU)$, and $\mathcal{V} = \ker(V^T U)$. Assume that* $\det(SV)$ *has sign* $(-1)^n$*. Define $Z \equiv V(SV)^{-1}S$. Then the following statements are equivalent:*

(A) *Given any vector $w_1 \in \mathcal{S}$ satisfying $w_1 \not\geq 0$, we can find a $w_2 \in \mathcal{V}$ satisfying $w_2 > 0$ such that $\langle w_1, w_2 \rangle < 0$.*

(B) $\det(J)$ *has sign $(-1)^{n+1}$ for any choice of $F \in -K$, $P \in K$, and $u > 0$.*

(C) *$I - U^T Z U$ is a nonnegative matrix*

Proof. The proof is identical to that of Theorem 6.2. In following through the steps the only things to note are that $U^T P \in \mathbb{R}^m_+$ and $U^T F \in \mathbb{R}^m_-$. Further,

$$\begin{aligned} P^T(I - Z)F &= P^T U U^T (I - Z) U U^T F \\ &= P^T U (I - U^T Z U) U^T F, \end{aligned}$$

and $U^T Z U$ is a projection, now projecting along $\ker(SU)$ onto $\ker(V^T U)$. ∎

Similarly, Theorem 6.3 extends to the following.

THEOREM 6.6. *Let J be a matrix of the form defined in* (6.2) *and U be any orthogonal matrix. Define $K = U(\mathbb{R}^m_+)$. Let α be some subset of $\{1, \ldots, n\}$, S^α be the matrix S with rows belonging to α removed, and V^α the matrix V with columns belonging to α removed. Define $\mathcal{S}^\alpha = \ker(S^\alpha U)$ and $\mathcal{V}^\alpha = \ker((V^\alpha)^T U)$.*

Assume that SV is a $P^{(-)}$ matrix. Then the following statements are equivalent:

(A) *For every α, given any vector $w_1 \in \mathcal{S}^\alpha$ satisfying $w_1 \not\geq 0$, we can find a vector $w_2 \in \mathcal{V}^\alpha$ satisfying $w_2 > 0$ such that $\langle w_1, w_2 \rangle < 0$.*

(B) *J is a $P^{(-)}$ matrix for any choice of $P \in K$, $F \in -K$, and $u > 0$.*

Proof. The proof follows directly from that of Theorem 6.5. ∎

6.2. Two examples. The way the above theorems can be used is seen in the two examples to follow. The first is a rather trivial example for illustrative purposes; the second is considerably harder and arises from a real biological system.

It is appropriate to mention that because for NAC reaction systems (see section 3) S and V^T have opposite sign structures, there are certain natural relationships between $\ker(S^\alpha)$ and $\ker((V^T)^\alpha)$. However, this fact alone does not imply that condition (A) of Theorem 6.2 is automatically fulfilled. Our first example illustrates this.

Example 1. Consider an open reaction system in which a single substrate x is involved in three processes, one of which produces a molecule of x, one of which produces two molecules of x, and one of which degrades a molecule of x, so that $S = [1, 2, -1]$. Let $V^T = [-a, -b, c]$, where $a, b, c \geq 0$ and not all are equal to zero. Now $SV = -(a + 2b + c) < 0$, so the Jacobian of the basic system is a negative scalar and hence a $P^{(-)}$ matrix. $\mathcal{S}$ is the plane in $\mathbb{R}^3$ satisfying $x_1 + 2x_2 - x_3 = 0$, while $\mathcal{V}$ is the plane satisfying $ax_1 + bx_2 - cx_3 = 0$.

Now assume that there is a rate-dependent feedback process such that the three reactions all create and are inhibited by some quantity. With P_i and F_i taking their usual meanings, we can check that the Jacobian of the system is

$$J = \begin{bmatrix} -(a+2b+c) & F_1 + 2F_2 - F_3 \\ -aP_1 - bP_2 + cP_3 & P_1F_1 + P_2F_2 + P_3F_3 - u \end{bmatrix}.$$

It is not immediately obvious by inspection that there are indeed choices of $P_i \geq 0$ and $F_i \leq 0$ for which J is not a $P^{(-)}$ matrix, but an easy calculation gives

$$I - V(SV)^{-1}S = \frac{1}{(a+2b+c)} \begin{bmatrix} 2b+c & -2a & a \\ -b & a+c & b \\ c & 2c & a+2b \end{bmatrix},$$

which is clearly not a nonnegative matrix unless a and b are both zero. So, for small $u > 0$ and some choices of P and F, J is indeed not a $P^{(-)}$ matrix, and can in fact be singular. Thus the $P^{(-)}$ matrix property can be destroyed by a rate-dependent negative feedback process for this NAC reaction system.

Example 2. In [1] a model of mitochondrial metabolism was presented consisting of a system of coupled redox reactions, some of which interacted with the proton gradient across the mitochondrial membrane. Without this gradient the Jacobian of the system could be written as the product of a matrix

$$S = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$

and a matrix V:

$$V = \begin{bmatrix} f_{11} & 0 & 0 & \cdots & 0 & 0 \\ -F_{21} & f_{22} & 0 & \cdots & 0 & 0 \\ 0 & -F_{32} & f_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & f_{n-1,n-1} & 0 \\ 0 & 0 & 0 & \cdots & -F_{n-1,n} & f_{nn} \\ 0 & 0 & 0 & \cdots & 0 & -F_{n+1,n} \end{bmatrix}.$$

All quantities of the form F_{ij} and f_{ii} are strictly positive. Note that V is a rectangular $(n+1) \times (n)$ matrix (i.e., $V : \mathbb{R}^n \to \mathbb{R}^{n+1}$), while S is a rectangular $(n) \times (n+1)$ matrix (i.e., $S : \mathbb{R}^{n+1} \to \mathbb{R}^n$), and they have opposite sign structures.

The Jacobian J of the full system with the potential included is an $(n+1)\times(n+1)$ matrix of the form:

$$\begin{bmatrix} -f_{11} - F_{21} & f_{22} & \cdots & 0 & F_2 - F_1 \\ F_{21} & -f_{22} - F_{32} & \cdots & 0 & F_3 - F_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -f_{nn} - F_{n+1,n} & F_{n+1} - F_n \\ P_1f_{11} - P_2F_{21} & P_2f_{22} - P_3F_{32} & \cdots & P_nf_{nn} - P_{n+1}F_{n+1,n} & -u + \sum_{i=1}^{n+1}(P_iF_i) \end{bmatrix}.$$

Elementary physical assumptions imply that $u > 0$, $P \equiv [P_1, P_2, \ldots P_{n+1}]^T > 0$, and $F \equiv [F_1, F_2, \ldots, F_{n+1}]^T < 0$. (Note that in the notation of [1], $P_i = p_i$ and $F_i = -F_{i\psi}$.)

We wish to use Theorem 6.2 to show that J is a $P^{(-)}$ matrix for all $P > 0$, $F < 0$, and $u > 0$. Incidentally this is hard to show by any direct method but becomes almost immediate by Theorem 6.2. Note first that SV is of the form discussed in Appendix B, and thus SV is a nonsingular $M^{(-)}$ matrix and hence a $P^{(-)}$ matrix.

We start by showing that the sign of $\det(J)$ is $(-1)^{n+1}$. Since $\det(SV)$ has sign $(-1)^n$, it suffices to examine $\ker(S)$ and $\ker(V^T)$. We can see that the $\ker(S)$ consists only of multiples of the vector $[1, 1, \ldots, 1]^T$. On the other hand, by inspection or induction, the strictly positive vector defined by

$$y_1 = 1, \tag{6.3}$$

$$y_{i+1} = \frac{f_{i,i}}{F_{i+1,i}} y_i, \quad i = 1, \ldots, n, \tag{6.4}$$

spans $\ker(V^T)$. Thus this situation corresponds to the special case where $\ker(S)$ lies entirely in the nonnegative and nonpositive orthants of $\mathbb{R}^{n+1}$ and $\ker(V^T)$ contains a strictly positive vector, confirming that $\det(J)$ has sign $(-1)^{n+1}$.

We now show that J is a $P^{(-)}$ matrix, using Theorem 6.3. For any $\alpha \subset \{1, \ldots, n\}$, the coordinates of a vector $x \in \ker(S^\alpha)$ are defined by the equations

$$x_{i+1} = x_i, \quad i = 1, \ldots, n, \quad i \notin \alpha.$$

On the other hand, vectors $y \in \ker((V^\alpha)^T)$ satisfy

$$y_{i+1} = \frac{f_{i,i}}{F_{i+1,i}} y_i, \quad i = 1, \ldots, n, \quad i \notin \alpha.$$

Let x be an arbitrary vector in $\ker(S^\alpha)$ with some coordinates $x_j, \ldots, x_{j+k}$ negative. Then, regardless of the sizes of f_{ii} and F_{ij}, we can choose a positive vector $y \in \ker((V^\alpha)^T)$ with $y_j, \ldots, y_{j+k}$ much larger in magnitude than the other coordinates of y so that $\langle x, y \rangle < 0$. Thus the conditions of Theorem 6.3 are satisfied, and the Jacobian is a $P^{(-)}$ matrix.

7. Conclusions and extensions. We have shown that the structure of chemical reaction systems alone can determine whether their Jacobians are $P^{(-)}$ matrices. The property of the stoichiometric matrix being SSD for general reaction systems, and WSD for mass action reactions, has been shown to be fundamentally linked to whether these systems can admit multiple equilibria. A technique has been presented to study when the $P^{(-)}$ matrix property is preserved under rate-dependent negative feedback.

There are several possible extensions to this work. In the discussion on rate-dependent negative feedback processes, an arbitrary row with particular sign structure was added to the stoichiometric matrix S, and a column with opposite sign structure to the matrix V. However, in the case where the extra row in S corresponds to a chemical reactant this row is a constant, and only the column added to V can vary. This situation clearly gives rise to less restrictive conditions on S and V, which would preserve the $P^{(-)}$ property of the Jacobian under the feedback. A related question is to find a *geometric* (rather than a combinatorial) characterization of when adding a row (column) to a given SSD matrix preserves the SSD property. Finding such a

characterization would be helpful in explaining why many real reaction systems have the SSD property.

The reader might have noted that during this paper we have nowhere used the law of atomic balance [7]. Intuitively we know that the two reactions $A \rightleftharpoons B$ and $A \rightleftharpoons 2B$ cannot both be true reactions (with no inflow or outflow involved). Mathematically this corresponds to the fact that the stoichiometric matrix S of the true reactions should have at least one (often more) nonnegative left eigenvectors of zero, corresponding to conserved quantities. This endows S with additional structure, and it would be of interest to examine how this extra structure affects the likelihood of a given stoichiometric matrix being SSD.

We discussed the fact that certain additional assumptions can mean that $P^{(-)}$ matrices are actually Hurwitz. One of the most interesting of these is sign-symmetry, which can be implied by certain physical assumptions. We hope, in future work, to expand on these ideas, as they form an interesting extension to the results in this paper.

On the same theme, when reaction systems have Jacobians with more structure than simply being a $P^{(-)}$ matrix, and can be shown to be Hurwitz, it may sometimes be possible to write down sufficient conditions which guarantee that the system with feedback remains Hurwitz, for example, if the Jacobian of the full system is an H matrix [13], or is similar to one by a transformation preserving the $M^{(-)}$ structure of the original system.

Appendix A. Properties of SSD and WSD matrices. We collect a few easy results on SSD and WSD matrices which are needed for the arguments in this text. Note that, by definition, any submatrix of an SSD (WSD) matrix is SSD (WSD). Note also that swapping rows/columns of a matrix does not alter whether it is SSD (WSD).

The first result is a trivial consequence of the definitions and the properties of determinants.

LEMMA A.1. *Let S be any square matrix. Multiply some column or row in S by a scalar constant to get a new matrix $\tilde{S}$. Then if S is SNS or singular, so is $\tilde{S}$.*

The next result states that it is possible to augment matrices in certain simple ways and preserve the SSD property.

LEMMA A.2. *Let S be an SSD matrix. Augment S with a single column (row) which is a scalar multiple of some column (row) of S to get a new matrix S_{aug}. Then S_{aug} is SSD.*

Proof. Any square submatrix of S_{aug} either

1. occurs in S, in which case it is SNS or singular because S is SSD,
2. is a square submatrix of S with one column/row multiplied by a scalar, in which case it is SNS or singular by Lemma A.1,
3. contains some subset of both the original column/row and its multiple and hence is singular. ∎

Incidentally the above result would not hold if we replaced SSD by WSD: Although by definition any submatrix of a WSD matrix is WSD, the augmented versions of WSD matrices are not necessarily WSD. Thus in the statements of the theorems on mass action systems in section 4 it is essential that the systems be written as sets of irreversible reactions.

The next result shows that the set of WSD matrices contains the set of SSD matrices. (This is also a corollary of Theorem 1.2.5. in [4].)

LEMMA A.3. *Let S be an SNS or singular matrix and S_- be the matrix S with all positive entries set to zero. Then* $\det(S)\det(S_-) \geq 0$.

Proof. If S is singular, the result is trivial, so assume that S is SNS. Consider the family of matrices $S_p = (1-p)S + pS_-$ with $p \in [0,1]$. By the definition of SNS, if S is SNS, then S_p is in the same qualitative class as S for $p \in [0,1)$. By continuity of the determinant, S_- either has the same sign as S or is singular. □

From the previous two results it follows that if the stoichiometric matrix of a system of reactions is SSD, then it is also SSD when written as a system of irreversible reactions, in which case it is WSD when written this way.

The next few results are useful from an algorithmic point of view—they can considerably reduce the computational effort involved in calculating whether a matrix is SSD/WSD or not.

LEMMA A.4. *Let S be an SSD matrix. Let S_{aug} be the matrix S with a row/column containing at most one nonzero element added. Then S_{aug} is SSD.*

Proof. Any square submatrix of S_{aug} is either

1. a submatrix of S, and hence SNS or singular,
2. a single element and hence trivially SNS or singular, or
3. a submatrix of S augmented with an extra row/column containing at most one nonzero element. In this case the determinant is either zero or the product of a nonzero element and the determinant of a submatrix of S which is itself SNS or singular. □

From this the next result follows immediately.

LEMMA A.5. *Let S be a matrix which is not SSD. Let S_{dim} be the matrix S with some rows/columns containing no more than one nonzero element removed. Then S_{dim} is not SSD.*

Proof. Suppose that S_{dim} is SSD. A square submatrix of S is either

1. diagonal,
2. a submatrix of S_{dim},
3. a submatrix of S_{dim} augmented with rows/columns containing no more than one nonzero element.

In the first two cases it is trivial that the square submatrix is SNS or singular. In the third case the result follows from repeated application of Lemma A.4. □

Since any submatrix of an SSD matrix is SSD by definition, it follows that the full stoichiometric matrix of a CFSTR system (termed S_f in the text) is SSD iff the stoichiometric matrix S of the true reactions is SSD. In other words, columns containing a single element (corresponding to inflow/outflow processes) can be ignored when checking whether a matrix is SSD. Lemma A.5 also often considerably reduces the computational effort involved in checking whether a matrix is SSD, by allowing one to ignore rows in S containing a single element (i.e., to ignore reactants which participate in only one reaction).

The above result also extends to WSD matrices and thus reduces the computational effort involved in checking whether a matrix which has been shown not to be SSD is actually WSD. The next lemma, while tedious to state, is actually very useful in practice.

LEMMA A.6. *Let S_r refer to the stoichiometric matrix of a system of reactions, and S_{ir} refer to the stoichiometric matrix of the system written as a set of irreversible reactions. Let α be the set of rows in S_r containing a single element, and γ be the set of columns in S_r containing a single element. Let S_{dim} be the matrix S_{ir} with rows from α and columns from γ deleted. Then S_{ir} is WSD iff S_{dim} is.*

Proof. The "only if" part is trivial, as S_{dim} is a submatrix of S_{ir}. Suppose that S_{ir} is not WSD, and consider a square submatrix T of S_{ir} which does not satisfy $\det(T)\det(T_-) \geq 0$. Any elements of T not in S_{dim} must lie in rows/columns of T containing a single nonzero element, because if they lie in rows containing a single positive and a single negative element, then two columns of T will be multiples of each other and hence T will be singular. Further, any nonzero elements of T not in S_{dim} must be negative, since otherwise T_- would contain a row of zeroes and thus be singular. The only way that $\det(T)$ can be nonzero is if it takes the form of the product of these negative elements with the determinant of a submatrix $\tilde{T}$ of S_{dim}. Similarly $\det(T_-)$ must take the form of the product of these negative elements with the determinant of $\tilde{T}_-$. Thus $\det(T)\det(T_-)$ is a positive multiple of $\det(\tilde{T})\det(\tilde{T}_-)$, implying that $\det(\tilde{T})\det(\tilde{T}_-) < 0$. Thus S_{dim} is not WSD. □

This final lemma means that in checking whether a non-SSD matrix is actually WSD one can first remove rows corresponding to reactants which occur only in one (perhaps reversible) reaction from the stoichiometric matrix before checking the matrix.

Appendix B. Binary reaction systems. We present an important class of systems which give rise to $P^{(-)}$ matrix Jacobians. Consider a set of n reactants A_i, $i = 1, \ldots, n$. Assume that the only reactions taking place are (reversible or irreversible) interconversions between the reactants along with some inflow and outflow processes. It is reasonable to assume that the rates depend on the substrates in a monotone way, so such systems are nonautocatalytic. They have been discussed in some detail in [2], where global stability of a unique equilibrium was shown using techniques connected with logarithmic norms. Each column of the stoichiometric matrix S contains either a $+1$ and a -1 (interconversion) or a single negative entry (outflow). It can be shown inductively that this structure implies that S is SSD and hence that the Jacobian is a $P_0^{(-)}$ matrix.

Here we show that, subject to a weak assumption on the inflow and outflow processes, the Jacobian J is a $P^{(-)}$ matrix and in fact a nonsingular $M^{(-)}$ matrix.

Associated with any such interconversion network is a directed graph $\mathcal{G}$ on $n+1$ nodes. Nodes $i = 1, \ldots, n$ correspond to the n substrates, while the extra node which we term node 0 corresponds to the zero complex, i.e., to the outside of the system. For $i, j \geq 1$, there is an edge from node i to node j ($i \neq j$) iff $J_{ji} > 0$—i.e., A_i can be converted to A_j, or alternatively the rate of conversion of A_j to A_i is inhibited by the concentration of A_i. On the other hand, for $i \geq 1$, there is an edge from node i to node 0 iff A_i is subject either to an outflow process or to an inflow process whose rate is inhibited by an increase in the concentration of i. This has the consequence of ensuring that J is strictly dominant in the ith column.

Our assumption is that *there is a directed path in $\mathcal{G}$ from any node to node* 0. This has the physical interpretation that the concentration of any substrate is affected by the "outside," a considerably weaker condition than insisting on a CFSTR. We refer the reader to [2] for the details, but the above assumptions together imply that the following hold:

1. J has negative diagonal entries.
2. J has nonnegative off-diagonal entries.
3. There is a constant coordinate transformation T such that TJT^{-1} still satisfies conditions 1 and 2 and is also strictly diagonally dominant in every column; hence J is Hurwitz.

These three facts combine to ensure that J is a nonsingular $M^{(-)}$ matrix [3].

Nonsingular $M^{(-)}$ matrices are a subset of $P^{(-)}$ matrices. In fact, all trajectories converge to a unique equilibrium [2].

Note that the basic system of coupled redox reactions presented in [1] gives rise to a Jacobian of the above form, even though in this case electrons are being transferred rather than reactants interconverted.

Acknowledgment. Thanks are due to the reviewer of this manuscript who helped us understand several connections between this work and the work in [5].

REFERENCES

[1] M. Banaji, *A generic model of electron transport in mitochondria*, J. Theoret. Biol., 243 (2006), pp. 501–516.
[2] M. Banaji and S. Baigent, *Electron transfer networks*, J. Math. Chem., (2007), to appear; online version available with DOI 10.1007/s10910-007-9257-3.
[3] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Classics in Appl. Math. 9, SIAM, Philadelphia, 1979.
[4] R. A. Brualdi and B. L. Shader, *Matrices of Sign-Solvable Linear Systems*, Cambridge Tracts in Math. 116, Cambridge University Press, Cambridge, UK, 1995.
[5] G. Craciun and M. Feinberg, *Multiple equilibria in complex chemical reaction networks:* I. *The injectivity property*, SIAM J. Appl. Math., 65 (2005), pp. 1526–1546.
[6] G. Craciun and M. Feinberg, *Multiple equilibria in complex chemical reaction networks:* II. *The species-reaction graph*, SIAM J. Appl. Math., 66 (2006), pp. 1321–1338.
[7] P. Érdi and J. Tóth, *Mathematical Models of Chemical Reactions*, Nonlinear Science: Theory and Applications, Manchester University Press, Manchester, UK, 1989.
[8] I. Famili and B. O. Palsson, *The convex basis of the left null space of the stoichiometric matrix leads to the definition of metabolically meaningful pools*, Biophys. J., 85 (2003), pp. 16–26.
[9] A. Fernandes, C. Gutierrez, and R. Rabanal, *On local diffeomorphisms of* $\mathbb{R}^n$ *that are injective*, Qual. Theory Dyn. Syst., 4 (2003), pp. 255–262.
[10] D. Gale and H. Nikaido, *The Jacobian matrix and global univalence of mappings*, Math. Ann., 159 (1965), pp. 81–93.
[11] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959.
[12] J. Hadamard, *Sur les transformations ponctuelles*, Bull. Soc. Math. France, 34 (1906), pp. 71–84.
[13] D. Hershkowitz, *Recent directions in matrix stability*, Linear Algebra Appl., 171 (1992), pp. 161–186.
[14] D. Hershkowitz and N. Keller, *Positivity of principal minors, sign symmetry and stability*, Linear Algebra Appl., 364 (2003), pp. 105–124.
[15] R. B. Kellogg, *On complex eigenvalues of* M *and* P *matrices*, Numer. Math., 19 (1972), pp. 70–175.
[16] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Computer Science and Applied Mathematics, Academic Press, New York, 1985.
[17] H. Nikaido, *Convex Structures and Economic Theory*, Academic Press, New York, 1968.
[18] S. Pinchuk, *A counterexample to the real Jacobian conjecture*, Math. Z., 217 (1994), pp. 1–4.
[19] S. M. Rump, *On P-matrices*, Linear Algebra Appl., 363 (2003), pp. 237–250.
[20] B. Smyth and F. Xavier, *Injectivity of local diffeomorphisms from nearly spectral conditions*, J. Differential Equations, 130 (1996), pp. 406–414.

SIAM J. APPL. MATH.
Vol. 67, No. 6, pp. 1548–1581

WEAK LACUNAE OF ELECTROMAGNETIC WAVES IN DILUTE PLASMA*

S. V. TSYNKOV†

Abstract. The propagation of waves is said to be diffusionless, and the corresponding governing PDE (or system) is said to satisfy Huygens' principle if the waves due to compactly supported sources have sharp aft fronts. The areas of no disturbance behind the aft fronts are called lacunae. Diffusionless propagation of waves is rare, whereas its opposite—diffusive propagation accompanied by aftereffects—is common. Nonetheless, lacunae can still be observed in a number of important applications, including the Maxwell equations in vacuum or in dielectrics with static response. In the framework of these applications, lacunae can be efficiently exploited for the numerical simulation of unsteady waves, and considerable progress has been made toward the development of lacunae-based methods for computational electromagnetism. Maxwell equations in vacuum are Huygens' because they reduce to a set of d'Alembert equations. Besides d'Alembert equations, there are no other scalar Huygens' equations in the standard $3+1$-dimensional Minkowski space-time. In terms of physics, this means that the mechanisms of dissipation and dispersion destroy the lacunae. In fact, all conventional low-frequency electromagnetic models, such as metals with Ohm conductivity, semiconductors, and magnetohydrodynamic media, are diffusive. An important case of the propagation of high-frequency electromagnetic waves in plasma is governed by the Klein–Gordon equation. It does not reduce to the d'Alembert equation either, and therefore the corresponding propagation is diffusive as well. However, one can still identify "weak lacunae" in the solutions of the Klein–Gordon equation, with the aft fronts that can be clearly observed, although they may not be as sharp as in the pure Huygens' case. Moreover, one can show that the "depth" of a weak lacuna is controlled by the dimensionless ratio of the Langmuir frequency to the primary carrying frequency of the waves.

Key words. Huygens' principle, wave diffusion, aftereffects, aft fronts, lacunae, ionospheric propagation, isotropic plasma, Langmuir frequency, cold plasma, transverse waves, Maxwell equations, Klein–Gordon equation, weak dispersion, short waves, external magnetic field, cyclotron frequency, gyrotropy, Faraday rotation

AMS subject classifications. 78A40, 35Q60, 81U30, 65Z05

DOI. 10.1137/060655134

1. Introduction.

1.1. The Huygens' principle.

Consider a three-dimensional Cauchy problem for the wave (d'Alembert) equation:

$$\frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2}-\Delta\varphi=f(\boldsymbol{x},t),\qquad \varphi(\boldsymbol{x},0)=\varphi_t(\boldsymbol{x},0)=0,\tag{1.1}$$
$$\mathbb{R}^3\ni\boldsymbol{x}=(x_1,x_2,x_3).$$

The fundamental solution of the d'Alembert operator is the expanding spherical wave (single layer)

$$\mathcal{E}(\boldsymbol{x},t)=\frac{\Theta(t)}{4\pi}\frac{\delta(|\boldsymbol{x}|-ct)}{t},\tag{1.2}$$

*Received by the editors March 24, 2006; accepted for publication (in revised form) April 24, 2007; published electronically September 12, 2007. This research was supported by US Air Force grant FA9550-04-1-0118.
http://www.siam.org/journals/siap/67-6/65513.html
†Department of Mathematics, North Carolina State University, Box 8205, Raleigh, NC 27695 (tsynkov@math.ncsu.edu, http://www4.ncsu.edu/~stsynkov).

where $\Theta(t)$ is the Heaviside function, and the solution to the Cauchy problem (1.1) is given by the convolution of the fundamental solution (1.2) with the right-hand side $f(\boldsymbol{x},t)$, i.e., by the Kirchhoff integral

$$\varphi(\boldsymbol{x},t)=\mathcal{E}*f=\frac{1}{4\pi}\iiint\limits_{\varrho\le ct}\frac{f\left(\boldsymbol{\xi},t-\varrho/c\right)}{\varrho}d\boldsymbol{\xi}, \tag{1.3}$$

where $\mathbb{R}^3\ni\boldsymbol{\xi}=(\xi_1,\xi_2,\xi_3)$ and $\varrho=|\boldsymbol{x}-\boldsymbol{\xi}|$.

Assume now that the right-hand side $f(\boldsymbol{x},t)$ is compactly supported in space and in time, i.e., that $\operatorname{supp} f\subseteq Q$, where Q is a bounded region in $\mathbb{R}^3\times[0,+\infty)\equiv\{(\boldsymbol{x},t)\,|\,\boldsymbol{x}\in\mathbb{R}^3,\,0\le t<+\infty\}$. Then, the Kirchhoff formula (1.3) immediately implies that

$$\varphi(\boldsymbol{x},t)\equiv 0\ \text{ for }\ (\boldsymbol{x},t)\in\bigcap_{(\boldsymbol{\xi},\tau)\in Q}\left\{(\boldsymbol{x},t)\big||\boldsymbol{x}-\boldsymbol{\xi}|<c(t-\tau),\ t>\tau\right\}. \tag{1.4}$$

The region of space-time defined by formula (1.4) is known as the *lacuna* of the solution $\varphi(\boldsymbol{x},t)$ of problem (1.1), because the solution vanishes there. This region can be interpreted as the intersection of all the characteristic cones of the d'Alembert equation, once the vertex of the cone sweeps the support Q of the right-hand side.

The presence of lacunae (or lacunas) in the solution is equivalent to the existence of the sharp aft fronts of the waves. In other words, the perturbation due to a compactly supported source first reaches a given fixed location of the observer and then ceases completely once a finite interval of time has elapsed. Subsequently, the solution at this location remains identically zero. Lacunae can then be viewed as areas of "quietness" behind the aft fronts, and the latter, reciprocally, serve as boundaries of the lacunae.

Differential equations, for which lacunae can be identified in their solutions, are said to satisfy *the Huygens' principle*. The most well-known classical example is provided by the foregoing d'Alembert equation. The Huygens' principle should not be confused with another concept that bears the same name and that often appears in the context of wave propagation in optics. Namely, according to *the Huygens' construction,* at every given moment of time the front of the propagating wave can be considered a collection of secondary sources that altogether define the wave field at subsequent moments of time [5].

Existence of the lacunae is a rare and fragile property. Its opposite is known as *the diffusion of waves* and is considered common. The diffusion manifests itself by aftereffects that accompany the propagation of waves governed by non-Huygens' equations. In this case, there are no sharp aft fronts, and once the perturbation has reached a given observation point it will never cease but only decay in amplitude.

A key constraint that distinguishes between the diffusionless and diffusive propagation is that lacunae may exist only *if the number of space dimensions is odd.* In particular, the propagation of waves governed by the d'Alembert equation on the plane ($\mathbb{R}^2$, as opposed to $\mathbb{R}^3$) is already characterized by aftereffects.

Another important consideration is that studying the wave phenomena in the time domain is essential for the analysis and interpretation of the Huygens' principle. Indeed, a standard frequency-domain model is the Helmholtz equation

$$\Delta\hat{\varphi}+k^2\hat{\varphi}=\hat{f}, \tag{1.5}$$

which is obtained from the d'Alembert equation by applying the Fourier transform in time. In (1.5), $k^2=\omega^2/c^2$, and $\hat{\varphi}$ denotes the complex amplitude of the time-harmonic

wave at the frequency ω (i.e., the ω Fourier coefficient). Solutions of the Helmholtz equation (1.5) are known to be analytic in the areas of homogeneity; therefore, they may not turn into zero only on a subdomain.

A review of the facts and publications in the literature pertaining to the Huygens' principle can be found in [3]; see also [10, 11]. The question of describing the hyperbolic differential equations and systems that admit the diffusionless propagation of waves was first formulated by Hadamard [12, 13, 14]. He did not know any other examples besides the classical d'Alembert equation. The notion of lacunae was introduced and studied by Petrowsky in [23]; see also [7, Chapter VI]. He obtained general conditions for the coefficients of hyperbolic equations/systems that guaranteed the presence of lacunae. Subsequent work in this direction was done by Atiyah, Bott, and Gȧrding in [1, 2]. However, no other constructive examples of lacunae in the solutions have been found besides solutions of the wave equation and its equivalents. In fact, Matthisson [20] has shown that in the standard $3+1$-dimensional Minkowski space-time the only scalar hyperbolic equation that satisfies the Huygens' principle is the wave equation. From the standpoint of applications, this result provides one of the most convenient and useful criteria. Namely, the equation may be Huygens' only if it is equivalent to the d'Alembert equation. We will employ this criterion for the analysis in the current paper. In this regard, we also emphasize that the aforementioned equivalence does not have to be global; a given equation may only locally reduce to the d'Alembert equation. An interesting illustration of this fact is provided by Lax and Phillips in [19]—they analyze the waves that propagate on an n-dimensional sphere, where n is odd, and prove that the propagation is diffusionless. The first examples of nontrivial scalar equations (i.e., nonequivalent to the d'Alembert equation) that satisfy the Huygens' principle were built by Stellmacher (see [28, 16, 29]) in the spaces $\mathbb{R}^n$ for odd $n \geq 5$. His examples have the form $c^{-2}\varphi_{tt} - \Delta\varphi + H(\boldsymbol{x}, t)\varphi = 0$, where the function $H(\boldsymbol{x}, t)$ is specially chosen to guarantee the diffusionless propagation, in which case it is called the Huygens' potential [3]. There are also examples of nontrivial diffusionless (i.e., Huygens') systems (as opposed to scalar equations) in the standard Minkowski $3+1$ space-time [26, 3, 10], as well as examples of nontrivial scalar Huygens' equations in a $3+1$-dimensional space-time but equipped with an alternative metric (the so-called plane wave metric); see [3, 10, 9].

1.2. Applications of lacunae. Lacunae of a given differential equation or system can be efficiently exploited for designing advanced numerical integration techniques. Lacunae-based methods have been developed previously for solving the scalar wave equation [25, 24], as well as for the problems of computational acoustics [31] and computational electromagnetism [32, 33, 34]. For the simplest possible setup that involves the radiation of waves by a known source, these methods guarantee that the grid convergence of a given discrete approximation will be *uniform in time.* For a more general setting that involves a sophisticated or potentially unknown mechanism of wave generation confined to a bounded region, lacunae-based methods facilitate construction of highly accurate unsteady artificial boundary conditions (ABCs) with only fixed and limited extent of temporal nonlocality in time. Note that overcoming the nonlocality of the exact unsteady ABCs in time has long been regarded as a challenging numerical issue [30]. From this perspective it is important to emphasize that the bound on temporal nonlocality obtained through the use of lacunae does not come at the expense of any approximation and/or simplification of the model; it is rather an implication of the fundamental properties of the corresponding solutions.

In addition to having the aforementioned computational benefits, lacunae can

also be instrumental in performing a number of tasks other than numerical ones. For example, explicit knowledge of their shape can help in planning of electromagnetic measurements and subsequent interpretation of the results.

In the current paper, we are not going to concentrate on numerical issues, except in section 3.5. Instead, we will focus on the phenomenon of lacunae itself. In particular, we will see that in the context of electromagnetism, only the simplest models that involve the propagation of waves in vacuum or in dielectrics with static response admit lacunae in the classical sense of the word. Many other traditional low-frequency models, such as different types of dielectrics, metals, semiconductors, magnetohydrodynamic media (MHD), turn out to be diffusive. However, for the important case of the propagation of high-frequency electromagnetic waves in dilute plasma, lacunae can still be identified in the solutions of the Maxwell equations in some approximate sense. Moreover, one can show that the quality, or "depth," of these weak lacunae is controlled by the ratio of the Langmuir frequency, which is a key parameter that characterizes temporal responses of the plasma to the primary carrying frequency of the incident wave.

2. Traditional electromagnetic models.

2.1. The Maxwell system of equations. Lacunae in vacuum. The evolution of electromagnetic field in vacuum is governed by the classical Maxwell equations

$$\begin{aligned} &\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} + \operatorname{curl}\boldsymbol{E} = \boldsymbol{0}, && \operatorname{div}\boldsymbol{B} = 0, \\ &\frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t} - \operatorname{curl}\boldsymbol{B} = -\frac{4\pi}{c}\boldsymbol{j}_{\text{ext}}, && \operatorname{div}\boldsymbol{E} = 4\pi\rho_{\text{ext}}. \end{aligned} \tag{2.1}$$

In system (2.1), $\boldsymbol{E}$ and $\boldsymbol{B}$ are intensities of the electric and magnetic field, respectively, c is the speed of light, $\boldsymbol{j}_{\text{ext}}$ is the density of the extraneous current, and ρ_{ext} is the density of the extraneous electric charge [17]. *A necessary solvability condition* for system (2.1) is continuity of the charges and currents:

$$\frac{\partial \rho_{\text{ext}}}{\partial t} + \operatorname{div}\boldsymbol{j}_{\text{ext}} = 0. \tag{2.2}$$

Equation (2.2) is obtained by taking divergence of the second unsteady equation of (2.1) and then substituting the second steady-state equation of (2.1). From the standpoint of physics, continuity (2.2) implies the conservation of electric charge. The rate of change of the total charge contained in any given region of space is equal to the flux of the charge, i.e., the total current, through the boundary of this region.

By differentiating each unsteady equation of (2.1) with respect to time, taking curl of the other unsteady equation, substituting $\operatorname{curlcurl}[\,\cdot\,] = \operatorname{graddiv}[\,\cdot\,] - \Delta[\,\cdot\,]$, and employing the corresponding steady-state equation of (2.1), we arrive at the following individual equations for the field intensities $\boldsymbol{B}$ and $\boldsymbol{E}$:

$$\begin{aligned} &\frac{1}{c^2}\frac{\partial^2 \boldsymbol{B}}{\partial t^2} - \Delta\boldsymbol{B} = \frac{4\pi}{c}\operatorname{curl}\boldsymbol{j}_{\text{ext}}, \\ &\frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} - \Delta\boldsymbol{E} = -4\pi\left[\frac{1}{c^2}\frac{\partial \boldsymbol{j}_{\text{ext}}}{\partial t} + \operatorname{grad}\rho_{\text{ext}}\right]. \end{aligned} \tag{2.3}$$

Equations (2.3) are vector d'Alembert equations with the propagation speed c. Each equation of (2.3) is Huygens' in $\mathbb{R}^3$, and hence system (2.1) is also Huygens'. If the

charges ρ_{ext} and currents $\boldsymbol{j}_{\text{ext}}$ are compactly supported, then the solution of (2.1) will have a lacuna of the same structure as determined by the Kirchhoff integral (1.3). Hence, the three-dimensional propagation of electromagnetic waves in vacuum is diffusionless.

Equations (2.1) will also remain a valid model for describing the electromagnetic field in various materials, but *only on the microscopic level.* The macroscopic equations are obtained by averaging; see [18]. In doing so, the impinging electromagnetic field may give rise to the induced charges and currents (see section 2.2), which, in turn, may affect the fields themselves. This range of phenomena is described by introducing the electric induction (or displacement) $\boldsymbol{D}$ and the magnetic field $\boldsymbol{H}$, whereas the "old" quantity $\boldsymbol{B}$ is referred to as the magnetic induction. The macroscopic Maxwell equations in the medium then become

$$\begin{aligned} &\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} + \operatorname{curl}\boldsymbol{E} = \boldsymbol{0}, && \operatorname{div}\boldsymbol{B} = 0, \\ &\frac{1}{c}\frac{\partial \boldsymbol{D}}{\partial t} - \operatorname{curl}\boldsymbol{H} = -\frac{4\pi}{c}\boldsymbol{j}_{\text{ext}}, && \operatorname{div}\boldsymbol{D} = 4\pi\rho_{\text{ext}}. \end{aligned} \tag{2.4}$$

Note that once $\boldsymbol{B}$ is referred to as the induction, and $\boldsymbol{H}$ as the magnetic field, system (2.4) looks mathematically more symmetric. However, as far as the physics is concerned, the true intensity of the magnetic field[1] is $\boldsymbol{B}$ rather than $\boldsymbol{H}$. As for the right-hand sides' $\boldsymbol{j}_{\text{ext}}$ and ρ_{ext} of system (2.4), they may be interpreted differently for different types of media and may sometimes be treated only as formal mathematical source terms.

System (2.4) is underdetermined unless additional relations are specified between the electric quantities $\boldsymbol{E}$ and $\boldsymbol{D}$ and the magnetic quantities $\boldsymbol{H}$ and $\boldsymbol{B}$. These relations are determined by the medium, across which the electromagnetic waves propagate. They are called *the responses.* The responses may vary drastically for different types of media and different regimes of propagation. The simplest response is static.

2.2. Dielectric media with static response. Lacunae. A dielectric medium may not support a constant (i.e., steady-state) electric current. Responses of a dielectric medium can be characterized in terms of the electric polarization $\boldsymbol{P}$, which is the induced electric dipole moment per unit volume, and magnetization $\boldsymbol{M}$, which is the induced magnetic dipole moment per unit volume. Then, by definition,

$$\boldsymbol{D} = \boldsymbol{E} + 4\pi\boldsymbol{P} \quad \text{and} \quad \boldsymbol{B} = \boldsymbol{H} + 4\pi\boldsymbol{M}. \tag{2.5}$$

In an isotropic dielectric with static response, the electric induction $\boldsymbol{D}$ is assumed directly proportional to the electric field $\boldsymbol{E}$, and the magnetic induction $\boldsymbol{B}$ is assumed directly proportional to the magnetic field $\boldsymbol{H}$:

$$\boldsymbol{D} = \epsilon\boldsymbol{E} \quad \text{and} \quad \boldsymbol{B} = \mu\boldsymbol{H}, \tag{2.6}$$

where the dielectric permittivity $\epsilon = \text{const}$ and the magnetic permeability $\mu = \text{const}$. In vacuum, we have $\epsilon = \mu = 1$, so that (2.4), (2.6) transform back to (2.1). In dielectric media other than vacuum, the assumptions of $\epsilon = \text{const}$ and $\mu = \text{const}$ may hold only for static incident fields. They can be used in the case of unsteady fields as well, but only as approximations and provided that *the incident frequencies*

[1] A quantitative characteristic of the field that determines how it affects the moving charged particles.

are low,[2] i.e., considerably lower than the typical frequencies of the molecular or electronic oscillations that are responsible for the onset of electric polarization and/or magnetization of the medium.

Under the assumption of a static response (2.6), the Maxwell equations (2.4) reduce to

$$\begin{aligned} &\frac{\mu}{c}\frac{\partial \boldsymbol{H}}{\partial t} + \operatorname{curl}\boldsymbol{E} = \boldsymbol{0}, && \operatorname{div}\boldsymbol{H} = 0, \\ &\frac{\epsilon}{c}\frac{\partial \boldsymbol{E}}{\partial t} - \operatorname{curl}\boldsymbol{H} = -\frac{4\pi}{c}\boldsymbol{j}_{\text{ext}}, && \operatorname{div}\boldsymbol{E} = \frac{4\pi}{\epsilon}\rho_{\text{ext}}. \end{aligned} \tag{2.7}$$

Then, a procedure identical to the one used when deriving equations (2.3) from (2.1) yields

$$\begin{aligned} &\frac{\epsilon\mu}{c^2}\frac{\partial^2 \boldsymbol{H}}{\partial t^2} - \Delta \boldsymbol{H} = \frac{4\pi}{c}\operatorname{curl}\boldsymbol{j}_{\text{ext}}, \\ &\frac{\epsilon\mu}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} - \Delta \boldsymbol{E} = -4\pi\left[\frac{\mu}{c^2}\frac{\partial \boldsymbol{j}_{\text{ext}}}{\partial t} + \frac{1}{\epsilon}\operatorname{grad}\rho_{\text{ext}}\right]. \end{aligned} \tag{2.8}$$

Thus, equations (2.8) that individually govern the fields $\boldsymbol{H}$ and $\boldsymbol{E}$ in $\mathbb{R}^3$ are Huygens'. As such, so is system (2.7). The corresponding wave speed $c/\sqrt{\epsilon\mu}$ is slower than the speed of light c.

Unfortunately, the propagation in vacuum or in dielectrics with static response is practically the only case of electromagnetic propagation with no aftereffects. In sections 2.3, 2.4, and 2.5, we will see that many conventional electrodynamic models appear diffusive even before the onset of *dispersion,* i.e., for *low frequencies,* when static relations between $\boldsymbol{D}$, $\boldsymbol{B}$ and $\boldsymbol{E}$, $\boldsymbol{H}$ can still be employed for unsteady fields. The propagation remains diffusive in the case of higher incident frequencies as well.[3]

Note also that the description of the responses in terms of the polarization $\boldsymbol{P}$ and magnetization $\boldsymbol{M}$ (see (2.5)) naturally brings along the definition of the induced charge ρ_{ind} and the induced current $\boldsymbol{j}_{\text{ind}}$:

$$\rho_{\text{ind}} = -\operatorname{div}\boldsymbol{P}, \qquad \boldsymbol{j}_{\text{ind}} = \frac{\partial \boldsymbol{P}}{\partial t} + c\operatorname{curl}\boldsymbol{M}. \tag{2.9}$$

Substitution of (2.5) and (2.9) into the Maxwell equations (2.4) yields

$$\begin{aligned} &\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} + \operatorname{curl}\boldsymbol{E} = \boldsymbol{0}, && \operatorname{div}\boldsymbol{B} = 0, \\ &\frac{1}{c}\frac{\partial \boldsymbol{E}}{\partial t} - \operatorname{curl}\boldsymbol{B} = -\frac{4\pi}{c}(\boldsymbol{j}_{\text{ext}} + \boldsymbol{j}_{\text{ind}}), && \operatorname{div}\boldsymbol{E} = 4\pi(\rho_{\text{ext}} + \rho_{\text{ind}}). \end{aligned} \tag{2.10}$$

System (2.10) is identical to (2.1), except that on its right-hand side we have the full current $\boldsymbol{j} = \boldsymbol{j}_{\text{ext}} + \boldsymbol{j}_{\text{ind}}$ and the full charge $\rho = \rho_{\text{ext}} + \rho_{\text{ind}}$ instead of only the extraneous quantities. This is an alternative way of representing the electromagnetic field inside a material—by looking at the actual intensities $\boldsymbol{B}$ and $\boldsymbol{E}$ only, but driven by the induced sources added to the original extraneous sources.[4]

[2]The notion of incident frequency is to be interpreted broadly here as frequency of any external excitation to the field inside the material, whether it be the frequency of the actual impinging wave or the frequency of the extraneous sources.

[3]Incident frequencies on the order of, or higher than, the characteristic microscopic frequencies for a given medium.

[4]Extraneous sources may or may not be present in every particular case.

2.3. Ohm conductivity in metals. In contradistinction to dielectrics, conducting materials can support a constant electric current. The steady-state model of a conductor can be obtained by dropping the displacement current $\frac{\partial \boldsymbol{D}}{\partial t} = \epsilon \frac{\partial \boldsymbol{E}}{\partial t}$ from the second unsteady Maxwell equation (2.4) or (2.7), which yields

$$\operatorname{curl} \boldsymbol{H} = \frac{4\pi}{c} \boldsymbol{j}_{\mathrm{c}}. \tag{2.11}$$

The quantity $\boldsymbol{j}_{\mathrm{c}}$ on the right-hand side of (2.11) is called the conductivity current. In the pure static case it is assumed given, and then (2.11) is solved along with $\operatorname{div} \boldsymbol{B} = 0$ to determine the magnetic field. Note that according to formula (2.11) the conductivity current is solenoidal, $\operatorname{div} \boldsymbol{j}_{\mathrm{c}} = 0$, which is a manifestation of the conservation of charge in this case.

The foregoing static model for conducting materials such as metals can also be applied to the analysis of slowly varying electromagnetic fields. In this case, however, the conductivity current $\boldsymbol{j}_{\mathrm{c}}$ shall no longer be treated as given. It rather becomes an unsteady current induced by the electric field that, in turn, is due to the variation in the magnetic field. Then, one also needs to add the first unsteady equation of the Maxwell system (2.4) or (2.7) to (2.11) and $\operatorname{div} \boldsymbol{B} = 0$. In doing so, the displacement current may still remain omitted from (2.11). The justification for not including it into the unsteady analysis is outlined in section 2.4, where a more comprehensive model is considered that includes semiconductors.

The key relation that one still needs in order to complete the unsteady model is a connection between the conductivity current and the electric field. Often, this connection is provided by the same classical Ohm law of electrostatics that establishes the direct proportionality between $\boldsymbol{j}_{\mathrm{c}}$ and the electric intensity $\boldsymbol{E}$:

$$\boldsymbol{j}_{\mathrm{c}} = \sigma \boldsymbol{E}. \tag{2.12}$$

The quantity σ in formula (2.12) is the electric conductivity; in the isotropic case it is a scalar. The conductivity σ can be assumed constant, and accordingly, static relations (2.11), (2.12) can be used for the unsteady fields in metals, under conditions similar to those discussed in section 2.2. Namely, the frequency of the incident field must be much lower than the characteristic frequencies of the microscopic mechanism of conductivity, which is due to the collisions between the conductivity electrons and atoms of the crystal lattice. Therefore, the incident frequency must be much lower than the collision frequency $\mathcal{O}(v_{\mathrm{e}}/\delta)$, where v_{e} is the electron thermal speed and δ is the mean free path.

By combining the first two equations of (2.4) with relations (2.11) and (2.12), we obtain the following system of equations that governs the unsteady electromagnetic field in metals:

$$\begin{gathered} \frac{1}{c} \frac{\partial \boldsymbol{B}}{\partial t} + \operatorname{curl} \boldsymbol{E} = \boldsymbol{0}, \qquad \operatorname{div} \boldsymbol{B} = 0, \\ \operatorname{curl} \boldsymbol{H} = \frac{4\pi}{c} \sigma \boldsymbol{E}, \end{gathered} \tag{2.13}$$

where we again assume that $\boldsymbol{B} = \mu \boldsymbol{H}$ with $\mu = \mathrm{const}$. From system (2.13) we easily obtain

$$\frac{1}{c} \frac{\partial \boldsymbol{B}}{\partial t} + \operatorname{curl} \frac{c}{4\pi\sigma} \operatorname{curl} \boldsymbol{H} = \boldsymbol{0},$$

which, along with $\mathrm{div}\boldsymbol{B} = \mu \mathrm{div}\boldsymbol{H} = 0$, yields the following parabolic equation for the magnetic field $\boldsymbol{H}$:

$$\frac{\partial \boldsymbol{H}}{\partial t} - \frac{c^2}{4\pi\sigma\mu}\Delta \boldsymbol{H} = \boldsymbol{0}. \tag{2.14}$$

Once (2.14) is solved, the electric field $\boldsymbol{E}$ is determined by the magnetic field through the last equation of (2.13). Equation (2.14) is not equivalent to the d'Alembert equation. Hence, according to the Matthisson criterion [20], it is not Huygens', and there may be no lacunae in its solutions.

We should also notice that (2.14) is homogeneous and therefore may only be driven by the initial and/or boundary conditions, whereas previously we have analyzed lacunae in the solutions due to the compactly supported right-hand sides. Thus, let us see how a source term for (2.14) can be generated.

Let us introduce a nonphysical artificial current $\boldsymbol{j}_{\mathrm{a}}$ that will be included on the right-hand side of (2.11) and as such will be affecting the magnetic field $\boldsymbol{H}$,

$$\mathrm{curl}\boldsymbol{H} = \frac{4\pi}{c}\boldsymbol{j}_{\mathrm{c}} + \frac{4\pi}{c}\boldsymbol{j}_{\mathrm{a}}, \tag{2.15}$$

but will not itself be driven by the induced electric field $\boldsymbol{E}$ through the Ohm law (2.12). Then, we use (2.15) instead of (2.11) and obtain a modified form of system (2.13):

$$\begin{gathered} \frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} + \mathrm{curl}\boldsymbol{E} = \boldsymbol{0}, \qquad \mathrm{div}\boldsymbol{B} = 0, \\ \mathrm{curl}\boldsymbol{H} = \frac{4\pi}{c}\sigma\boldsymbol{E} + \frac{4\pi}{c}\boldsymbol{j}_{\mathrm{a}}. \end{gathered} \tag{2.16}$$

The conservation of charge in the case is expressed as the total current being solenoidal: $\mathrm{div}(\boldsymbol{j}_{\mathrm{c}} + \boldsymbol{j}_{\mathrm{a}}) = 0$. For simplicity, and with no substantial loss of generality (see Theorem 1 in [34]), we can also assume that the artificial current $\boldsymbol{j}_{\mathrm{a}}$ itself is divergence-free, $\mathrm{div}\boldsymbol{j}_{\mathrm{a}} = 0$. In this case, the electric field will remain solenoidal as in system (2.13): $\mathrm{div}\boldsymbol{E} = 0$. From (2.16) we obtain the inhomogeneous counterpart of (2.14):

$$\frac{\partial \boldsymbol{H}}{\partial t} - \frac{c^2}{4\pi\sigma\mu}\Delta \boldsymbol{H} = \frac{1}{\sigma}\mathrm{curl}\boldsymbol{j}_{\mathrm{a}}. \tag{2.17}$$

Solutions of (2.17) do not have lacunae even if $\boldsymbol{j}_{\mathrm{a}}$ is compactly supported. We can therefore conclude that the propagation of electromagnetic waves in the media with Ohm conductivity is diffusive.

2.4. Semiconductors. Let us now look more thoroughly into how one shall actually treat the displacement current for conducting materials. Keeping the unsteady term $\frac{1}{c}\frac{\partial \boldsymbol{D}}{\partial t} = \frac{\epsilon}{c}\frac{\partial \boldsymbol{E}}{\partial t}$, i.e., considering

$$\mathrm{curl}\boldsymbol{H} = \frac{4\pi}{c}\sigma\boldsymbol{E} - \frac{\epsilon}{c}\frac{\partial \boldsymbol{E}}{\partial t} \tag{2.18}$$

instead of (2.11), (2.12), can make sense only under the special circumstances when the second term on the right-hand side of (2.18) is of the same order of magnitude as the first term, or at least not negligibly small compared to the first term. If the field is time-harmonic, then the ratio of these two terms is $\mathcal{O}\left(\frac{\epsilon\omega}{4\pi\sigma}\right)$. In metals, we typically have $\frac{\omega}{\sigma} \ll 1$ for the entire range of frequencies, for which the conductivity σ can still be considered constant [18]. Therefore, (2.18) in metals indeed reduces to (2.13).

In semiconductors, however, because of the low concentration of conductivity electrons, the value of σ could be very small, so that for all those frequencies, for which σ and ϵ can still be regarded as constants, we may already have $\frac{\epsilon\omega}{4\pi\sigma} = \mathcal{O}(1)$. Then, the Maxwell equations become (cf. formulae (2.13) and (2.7))

$$\begin{aligned} &\frac{\mu}{c}\frac{\partial \boldsymbol{H}}{\partial t} + \mathrm{curl}\boldsymbol{E} = \boldsymbol{0}, \quad \mathrm{div}\boldsymbol{H} = 0, \\ &\frac{\epsilon}{c}\frac{\partial \boldsymbol{E}}{\partial t} + \frac{4\pi}{c}\sigma\boldsymbol{E} - \mathrm{curl}\boldsymbol{H} = \boldsymbol{0}, \quad \mathrm{div}\boldsymbol{E} = 0. \end{aligned} \tag{2.19}$$

By differentiating the second unsteady equation of (2.19) with respect to t, taking curl of the first unsteady equation, and then substituting $\mathrm{div}\boldsymbol{E} = 0$, we arrive at the telegrapher's equation for the electric field:

$$\frac{\epsilon\mu}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} + \frac{4\pi\mu\sigma}{c^2}\frac{\partial \boldsymbol{E}}{\partial t} - \Delta\boldsymbol{E} = \boldsymbol{0}. \tag{2.20}$$

A right-hand side for (2.20) can be built similarly to how it was done in section 2.3 for (2.14). Namely, if we were to formally add the artificial source terms $-\frac{4\pi}{c}\boldsymbol{j}_{\mathrm{a}}$ and $\frac{4\pi}{\epsilon}\rho_{\mathrm{a}}$ to the second pair of the Maxwell equations (2.19), then we would have obtained the following equation instead of (2.20):

$$\frac{\epsilon\mu}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} + \frac{4\pi\mu\sigma}{c^2}\frac{\partial \boldsymbol{E}}{\partial t} - \Delta\boldsymbol{E} = -4\pi\left[\frac{\mu}{c^2}\frac{\partial \boldsymbol{j}_{\mathrm{a}}}{\partial t} + \frac{1}{\epsilon}\mathrm{grad}\rho_{\mathrm{a}}\right]. \tag{2.21}$$

The operator on the left-hand side of (2.21) is not equivalent to the d'Alembert operator. Therefore, the Huygens' principle will not hold, and there will be no lacunae. Note also that the larger the ratio $\frac{\epsilon\omega}{4\pi\sigma}$, the more of a standard dielectric behavior will be displayed by the medium governed by (2.18).

2.5. Magnetohydrodynamics. The case of a conducting medium in motion is not very different from the stationary conducting medium analyzed in section 2.3. Instead of the Ohm law (2.12) we now have

$$\boldsymbol{j}_{\mathrm{c}} = \sigma\left(\boldsymbol{E} + \frac{1}{c}\boldsymbol{u}\times\boldsymbol{B}\right), \tag{2.22}$$

where $\boldsymbol{u}$ denotes the velocity of the conducting fluid. The second term on the right-hand side of (2.22) is the so-called Lorentz correction that helps obtain the electric field in the frame of reference that moves with the velocity $\boldsymbol{u}$, provided that $|\boldsymbol{u}| \ll c$; see [17]. Accordingly, instead of system (2.13) we obtain

$$\begin{aligned} &\frac{1}{c}\frac{\partial \boldsymbol{B}}{\partial t} + \mathrm{curl}\boldsymbol{E} = \boldsymbol{0}, \qquad \mathrm{div}\boldsymbol{B} = 0, \\ &\mathrm{curl}\boldsymbol{H} = \frac{4\pi}{c}\sigma\left(\boldsymbol{E} + \frac{1}{c}\boldsymbol{u}\times\boldsymbol{B}\right), \end{aligned}$$

and instead of (2.14) we have

$$\frac{\partial \boldsymbol{H}}{\partial t} - \mathrm{curl}(\boldsymbol{u}\times\boldsymbol{H}) - \frac{c^2}{4\pi\sigma\mu}\Delta\boldsymbol{H} = \boldsymbol{0}. \tag{2.23}$$

As before, (2.23) is to be solved under the condition that the magnetic field is solenoidal: $\mathrm{div}\boldsymbol{H} = 0$.

Unlike in section 2.3, in magnetohydrodynamics the electromagnetic equations are not independent. They are coupled to the equations of the fluid flow through the quantity $\boldsymbol{u}$ in (2.23). Moreover, the ponderomotive force $\frac{1}{c}\boldsymbol{j}_{\mathrm{c}} \times \boldsymbol{H}$ is added to the right-hand side of the momentum equation of the fluid, and the Joule heat $\boldsymbol{j}_{\mathrm{c}}^2/\sigma$ is added to the right-hand side of the energy equation of the fluid. Therefore, we cannot directly apply the Matthisson criterion to (2.23); this can only be done if we consider the velocity field $\boldsymbol{u}$ as given. Then, the answer is still negative—(2.23) is not Huygens'.

Of particular interest may be the case of very large (theoretically, infinite) conductivities σ, when the dissipative term $\sim \Delta \boldsymbol{H}$ can be dropped from (2.23). Let us then consider the equations of inviscid compressible flow coupled with (2.23) for the magnetic field with no magnetic viscosity, $\frac{c^2}{4\pi\sigma\mu} = 0$:

$$\begin{aligned} \frac{d\varrho}{dt} + \varrho \mathrm{div}\boldsymbol{u} &= 0, \\ \varrho \frac{d\boldsymbol{u}}{dt} + \mathrm{grad}p &= \frac{1}{4\pi}\mathrm{curl}\boldsymbol{H} \times \boldsymbol{H}, \\ \frac{\partial \boldsymbol{H}}{\partial t} &= \mathrm{curl}(\boldsymbol{u} \times \boldsymbol{H}). \end{aligned} \tag{2.24}$$

In system (2.24), ϱ, p, and T are the density, pressure, and temperature of the fluid, respectively, and $\frac{d}{dt} = \frac{\partial}{\partial t} + (\boldsymbol{u} \cdot \mathrm{grad})$. It is easy to show (see, e.g., [27, Vol. 1]) that infinite conductivity also implies the adiabatic nature of the flow, because the Joule heat $\frac{1}{\sigma}\boldsymbol{j}_{\mathrm{c}}^2$ must be disregarded. Then, instead of the energy equation, system (2.24) can be supplemented by the Poisson adiabatic relation between the pressure and density of a thermodynamically ideal fluid: $p = \mathrm{const} \cdot \varrho^{\gamma}$, where $\gamma = \frac{c_p}{c_v}$ is the ratio of specific heats.

Let us linearize equations (2.24) at the background of an ambient conducting fluid immersed into a constant magnetic field, i.e., at the background of a constant solution: $\varrho = \varrho_0$, $p = p_0$, $\boldsymbol{u} = \boldsymbol{u}_0 = \boldsymbol{0}$, and $\boldsymbol{H} = \boldsymbol{H}_0$. Let $\varrho = \varrho_0 + \tilde{\varrho}$, $p = p_0 + \tilde{p}$, $\boldsymbol{u} = \tilde{\boldsymbol{u}}$, and $\boldsymbol{H} = \boldsymbol{H}_0 + \tilde{\boldsymbol{H}}$, where all the quantities with the tilde are small perturbations. Retaining only the first order terms with respect to these perturbations, we obtain

$$\begin{aligned} \frac{\partial \tilde{\varrho}}{\partial t} + \varrho_0 \mathrm{div}\tilde{\boldsymbol{u}} &= 0, \\ \varrho_0 \frac{\partial \tilde{\boldsymbol{u}}}{\partial t} + \mathrm{grad}\tilde{p} &= \frac{1}{4\pi}\mathrm{curl}\tilde{\boldsymbol{H}} \times \boldsymbol{H}_0, \\ \frac{\partial \tilde{\boldsymbol{H}}}{\partial t} &= \mathrm{curl}(\tilde{\boldsymbol{u}} \times \boldsymbol{H}_0), \\ \tilde{p} &= \frac{\gamma p_0}{\varrho_0}\tilde{\varrho}. \end{aligned}$$

Then, introducing the displacement vector $\boldsymbol{x}$ as $\frac{\partial \boldsymbol{x}}{\partial t} = \tilde{\boldsymbol{u}}$, we can derive the following equation (see [15]):

$$\frac{\partial^2 \boldsymbol{x}}{\partial t^2} = c_{\mathrm{s}}^2 \mathrm{graddiv}\boldsymbol{x} + c_{\mathrm{A}}^2 \mathrm{grad}_{\perp}\mathrm{div}\boldsymbol{x}_{\perp} + c_{\mathrm{A}}^2 \frac{\partial^2 \boldsymbol{x}_{\perp}}{\partial z^2}, \tag{2.25}$$

where $c_{\mathrm{s}} = \sqrt{\gamma p_0/\varrho_0}$ is the conventional speed of sound, $c_{\mathrm{A}} = |\boldsymbol{H}_0|/\sqrt{4\pi\varrho_0}$ is the Alfvén speed, and $\boldsymbol{x}_{\perp}$ and $\mathrm{grad}_{\perp}$ are the components of $\boldsymbol{x}$ and the gradient, respectively, orthogonal to the magnetic field $\boldsymbol{H}_0$.

For a particular class of transverse displacements, $\boldsymbol{x} = \boldsymbol{x}_\perp$ and $\mathrm{div}\boldsymbol{x}_\perp = 0$, we obtain from (2.25)

$$\frac{\partial^2 \boldsymbol{x}_\perp}{\partial t^2} = c_{\mathrm{A}}^2 \frac{\partial^2 \boldsymbol{x}_\perp}{\partial z^2}. \tag{2.26}$$

This is a one-dimensional d'Alembert equation that describes the propagation of the so-called Alfvén waves along the magnetic field with the speed c_{A}. Even though the number of space dimensions in (2.26) is odd, the one-dimensional case is special. Solutions to (2.26) may display the Huygens' behavior only if the equation is driven by some particular classes of initial data, whereas for the general RHS there is wave diffusion.

If the component of $\boldsymbol{x}$ along the magnetic field $\boldsymbol{H}_0$ is not zero, i.e., $x_3 \neq 0$, then (2.25) yields

$$\frac{\partial^2 x_3}{\partial t^2} = c_{\mathrm{s}}^2 \frac{\partial^2 x_3}{\partial z^2} + c_{\mathrm{s}}^2 \frac{\partial}{\partial z} \mathrm{div}\boldsymbol{x}_\perp. \tag{2.27}$$

For $\mathrm{div}\boldsymbol{x}_\perp = 0$, (2.27) governs the propagation of the so-called ion sound along the magnetic field with the speed c_{s}. As in the previous case of the Alfvén waves, the propagation of ion sound is diffusive.

To supplement (2.27), a second equation can be derived from (2.25) that would govern $\mathrm{div}\boldsymbol{x}_\perp$:

$$\begin{aligned} \frac{\partial^2 \mathrm{div}\boldsymbol{x}_\perp}{\partial t^2} &= c_{\mathrm{s}}^2 \mathrm{divgrad}_\perp \underbrace{\mathrm{div}\boldsymbol{x}}_{\mathrm{div}\boldsymbol{x}_\perp + \frac{\partial x_3}{\partial z}} + c_{\mathrm{A}}^2 \underbrace{\left[\mathrm{divgrad}_\perp \mathrm{div}\boldsymbol{x}_\perp + \frac{\partial^2 \mathrm{div}\boldsymbol{x}_\perp}{\partial z^2} \right]}_{\Delta \mathrm{div}\boldsymbol{x}_\perp} \\ &= c_{\mathrm{s}}^2 \mathrm{divgrad}_\perp \mathrm{div}\boldsymbol{x}_\perp + c_{\mathrm{s}}^2 \mathrm{divgrad}_\perp \frac{\partial x_3}{\partial z} + c_{\mathrm{A}}^2 \Delta \mathrm{div}\boldsymbol{x}_\perp. \end{aligned} \tag{2.28}$$

Equations (2.27) and (2.28) form a system with the unknowns x_3 and $\mathrm{div}\boldsymbol{x}_\perp$. These equations decouple only when $c_{\mathrm{s}} \ll c_{\mathrm{A}}$. In this case, the terms $\sim c_{\mathrm{s}}^2$ on the right-hand side of (2.28) can be disregarded, which yields

$$\frac{\partial^2 \mathrm{div}\boldsymbol{x}_\perp}{\partial t^2} = c_{\mathrm{A}}^2 \Delta \mathrm{div}\boldsymbol{x}_\perp. \tag{2.29}$$

Equation (2.29) governs the so-called magnetoacoustic waves that propagate with the Alfvén speed c_{A}. It is a true three-dimensional d'Alembert equation and as such, is Huygens'. The assumption of the speed of sound c_{s} being much slower than the Alfvén speed c_{A} holds when the thermodynamic pressure p_0 is much lower than the quantity $\boldsymbol{H}_0^2/8\pi$, which can be interpreted as pressure of the magnetic field [15].

Hence, lacunae can potentially exist in the solutions for the transverse quantity $\mathrm{div}\boldsymbol{x}_\perp$. However, $\mathrm{div}\boldsymbol{x}_\perp$ is then substituted into (2.27) to find the longitudinal displacement x_3 in magnetoacoustic waves, and the spatially one-dimensional solution for x_3 will, generally speaking, be diffusive. Altogether, the propagation of waves governed by (2.29), (2.27) will be only partially diffusionless.

2.6. Summary on low-frequency models. Having analyzed a number of conventional low-frequency electromagnetic models, we conclude that for most of them the propagation of waves is diffusive; i.e., the Huygens' principle does not hold. The driving frequency in these models is assumed lower than the characteristic microscopic frequencies of the medium, so that the material coefficients can be taken as constants

(permittivity, permeability, and conductivity). The mechanism that destroys the lacunae in all these cases is typically of a dissipative nature, related to the electric conductivity. An exception, for which the Huygens' principle holds, is pure dielectric materials with static response. Another partial exception is magnetoacoustic waves in the medium with infinite conductivity.

When the incident (driving) frequency becomes higher, the material coefficients can no longer be assumed constant. Instead, they become frequency-dependent, and while relations (2.6) can still keep their form, all the quantities involved have to be considered in the frequency domain rather than in the time domain. In other words, relations (2.6) transform into the corresponding relations between the Fourier coefficients of the fields and of material "constants," while in the physical space the medium responses typically appear nonlocal in time (given by convolution-type integrals); see [18]. It is also known that the discrepancy between $\boldsymbol{H}$ and $\boldsymbol{B}$ becomes unimportant/negligible even for relatively low frequencies. Hence, for higher frequencies only the discrepancy between $\boldsymbol{D}$ and $\boldsymbol{E}$ matters.

Hereafter, we will depart from the low-frequency framework and analyze the propagation of high-frequency electromagnetic waves in the dilute ionospheric plasma. We will see that in this case the key mechanism that can destroy the lacunae is of a dispersive nature. We will also see that under certain assumptions lacunae can still be identified in this dispersive medium, but in an approximate sense.

3. High-frequency electromagnetic waves in dilute plasma.

3.1. Characteristics of the medium. Our ultimate goal will be to work out an approximate interpretation of the Huygens' principle as it applies to the propagation of electromagnetic waves through the Earth's ionosphere. The ionosphere is a layer of dilute plasma (weakly ionized rarefied gas which is electrically neutral as a whole) surrounding the Earth at heights roughly between 60 km and 400 km from the surface. The primary source of ionization in the ionosphere is solar radiation. The negatively charged particles in the ionosphere are electrons with the charge of $e = -4.803 \cdot 10^{-10}$ Gaussian units and the mass of $m_{\mathrm{e}} = 9.1 \cdot 10^{-28} g$, and the positively charged particles are ions that are much heavier: $m_{\mathrm{i}}/m_{\mathrm{e}} \gtrsim 2.93 \cdot 10^4$. The ionosphere is, in fact, layered, and its local parameters strongly depend on the altitude; this dependence for key characteristics, such as the concentrations of charged particles, may be nonmonotonic. The parameters of the ionosphere also change between day and night and winter and summer, and depend on the level of solar activity; more detail can be found, e.g., in [8, 6]. In our subsequent considerations, we will be quoting the parameters typical for the so-called F-layer (that starts at about 130 km above the Earth's surface) during the periods of low solar activity. The concentrations of the negatively and positively charged particles are equal, and we will mostly use the electron concentration: $n_{\mathrm{e}} \approx 10^6 cm^{-3}$. Note that the concentration of neutral atoms and molecules in the F-layer could be as high as $n_{\mathrm{m}} = 10^{10}\mathrm{cm}^{-3}$. A typical value of the electron temperature in the F-layer is $T_{\mathrm{e}} \approx 2000K$; the ions are a few times colder.

Several key quantities that depend on the foregoing parameters characterize the properties of the ionospheric plasma. The plasma electron frequency, also known as the Langmuir frequency, is defined as $\omega_{\mathrm{pe}} = \sqrt{\frac{4\pi e^2 n_{\mathrm{e}}}{m_{\mathrm{e}}}}$; it provides a fundamental temporal scale. For the specific parameters of the plasma given above we obtain $\omega_{\mathrm{pe}} \approx 5.64 \cdot 10^7 rad/s \approx 9$ MHz; in the literature, one can find the range of values for the Langmuir frequency in the ionosphere between 3 MHz and 15 MHz. The thermal speed of the electrons, $v_{\mathrm{e}} = \sqrt{3\kappa T_{\mathrm{e}}/2m_{\mathrm{e}}} \approx 3 \cdot 10^7$ cm/s, provides a characteristic

velocity, where $\kappa = 1.38 \cdot 10^{-16} erg/K$ is the Boltzmann constant; the speed v_e is roughly three orders of magnitude slower than the speed of light in vacuum, $c = 3 \cdot 10^{10}$ cm/s. The speed of the waves that propagate through the plasma will subsequently need to be compared to the characteristic velocity v_e. The Debye shielding length, $d = \sqrt{\frac{\kappa T_e}{8\pi e^2 n_e}} \approx 0.22$ cm, provides a characteristic spatial scale for the shielding of a point charge immersed into the plasma by other charges; shielding effectively results in multiplication of the classical Coulomb electrostatic potential by the rapidly decaying function $e^{-r/d}$.

Another important parameter yet to be included in the consideration is the magnetic field of the Earth, $\boldsymbol{B}_0$, $|\boldsymbol{B}_0| \approx 0.3G$. It brings along another characteristic frequency known as the electron cyclotron frequency, $\Omega_e = \frac{e|\boldsymbol{B}_0|}{c \cdot m_e} \approx 0.8$ MHz, which is about an order of magnitude lower than the Langmuir frequency. The presence of $\boldsymbol{B}_0$ implies anisotropy of the plasma and transforms it into a gyrotropic medium; see [18]. The propagation of electromagnetic waves through such a medium is accompanied by interesting effects, e.g., the Faraday rotation. In the literature, these effects are typically studied in the frequency domain (see [18, Chapter XI]); for our analysis we will use the time domain (see section 3.6).

3.2. Cold plasma. In the Maxwell system of equations (2.10), assume that no extraneous charges or currents are present; then take curl of the first unsteady equation and by substitution eliminate the magnetic field from the second unsteady equation, having differentiated it with respect to time. This yields

$$\frac{\partial^2 \boldsymbol{E}}{\partial t^2} + c^2 \mathrm{curlcurl}\boldsymbol{E} = -4\pi \frac{\partial \boldsymbol{j}_{\mathrm{ind}}}{\partial t}. \tag{3.1}$$

Equation (3.1) is the key governing equation for the electric field. However, it still requires that the time derivative of the induced current on the right-hand side be specified. To do so, we will use the approximation known as cold plasma (see, e.g., [8, 21]); the meaning of the term will be explained later.

To obtain the current, let us write Newton's second law of motion for the electrons:

$$m_e \frac{d\boldsymbol{u}}{dt} + m_e \nu_{\mathrm{eff}} \boldsymbol{u} = -e\boldsymbol{E} - \frac{e}{c}\boldsymbol{u} \times \boldsymbol{B}. \tag{3.2}$$

As the ions are much heavier than the electrons, their motion is not taken into account. In (3.2), $\boldsymbol{u}$ denotes the velocity of the electrons due to the applied electromagnetic field (as opposed to the thermal velocity). Equation (3.2) is nonrelativistic because $\kappa T / m_e c^2 \approx 3.37 \cdot 10^{-7} \ll 1$. The quantity ν_{eff} in (3.2) is the effective frequency of collisions between the electrons and other particles (both charged and neutral). Note that the acceleration term in (3.2) is important in the case of high frequencies, whereas in the low-frequency case it is often omitted. Omitting the acceleration term results in (3.2) being transformed into the (generalized) Ohm law; see [15]. In the high-frequency case we can instead drop the collision term $m_e \nu_{\mathrm{eff}} \boldsymbol{u}$ on the left-hand side of (3.2). This term is responsible for the mechanism of Ohm conductivity in the plasma and is dropped because typical collision frequencies ν_{eff} in the ionosphere are low. A thorough analysis of collisions in dilute plasma requires the calculation of cross-sections using the apparatus of quantum mechanics; it goes beyond the scope of this paper, and we refer the reader to [8]. Here we only mention that for the collisions of electrons with either positive ions or neutral molecules in the F-layer we have $\nu_{\mathrm{eff}} \sim 10^2 s^{-1} \ll \omega_{\mathrm{pe}}$, and as we are predominantly interested in high incident frequencies, $\omega \gg \omega_{\mathrm{pe}}$, we can indeed disregard the collisions term in (3.2).

In the isotropic case, when the constant magnetic field $\boldsymbol{B}_0$ is not taken into account, the Lorentz term on the right-hand side of (3.2) can also be neglected. The reason is that unlike, for example, the case of MHD (section 2.5), when plasma is immersed into the magnetic field and the electric field is induced, here we are assuming that both the electric field and the magnetic field have roughly the same magnitude in the impinging wave. Then, the term $-\frac{e}{c}\boldsymbol{u}\times\boldsymbol{B}$ becomes a small relativistic correction, because $|\boldsymbol{u}| \ll c$. The latter relation always holds, because even when the plasma is not at thermal equilibrium, i.e., when the velocity distribution function is not Maxwellian, the speed of systematic motion $|\boldsymbol{u}|$ is still much slower than the average particle speed $\sqrt{2K/m_e e}$ (K is the kinetic energy), which, in turn, is much slower than the speed of light. Altogether, (3.2) then reduces to

$$m_{\mathrm{e}}\frac{d\boldsymbol{u}}{dt} = -e\boldsymbol{E}. \tag{3.3}$$

Next, by expressing the induced current as $\boldsymbol{j}_{\mathrm{ind}} = -en_{\mathrm{e}}\boldsymbol{u}$, we transform (3.3) into

$$\frac{\partial \boldsymbol{j}_{\mathrm{ind}}}{\partial t} = -en_{\mathrm{e}}\frac{\partial \boldsymbol{u}}{\partial t} = \frac{e^2 n_{\mathrm{e}}}{m_{\mathrm{e}}}\boldsymbol{E}. \tag{3.4}$$

In doing so we note that the foregoing expression $\boldsymbol{j}_{\mathrm{ind}} = -en_{\mathrm{e}}\boldsymbol{u}$ corresponds to a simplified framework, whereas, strictly speaking, we should have written $\boldsymbol{j}_{\mathrm{ind}} = -e\int \boldsymbol{v}f(\boldsymbol{v})d\boldsymbol{v}$, where $f(\boldsymbol{v})$ is the probability distribution function for electron velocities. In this paper, however, we employ the elementary approach rather than the full-fledged kinetic considerations.

We would also like to emphasize that the relation (3.4) between the induced current and the electric field is local in space, because (3.3) is an ordinary differential equation. In the frequency domain, when all the variables are interpreted as Fourier components, we immediately have

$$\boldsymbol{j}_{\mathrm{ind}}(\omega) = \frac{\omega_{\mathrm{pe}}}{4\pi}\frac{1}{i\omega}\boldsymbol{E}(\omega),$$

and since $\frac{\partial \boldsymbol{D}}{\partial t} = \frac{\partial \boldsymbol{E}}{\partial t} + 4\pi\frac{\partial \boldsymbol{P}}{\partial t} = \frac{\partial \boldsymbol{E}}{\partial t} + 4\pi\boldsymbol{j}_{\mathrm{ind}}$ (assuming $\mu = 1$; see (2.5), (2.9)), we obtain

$$\boldsymbol{D}(\omega) = \boldsymbol{E}(\omega) - \frac{\omega_{\mathrm{pe}}^2}{\omega^2}\boldsymbol{E}(\omega) \stackrel{\mathrm{def}}{=} \varepsilon\boldsymbol{E}(\omega) \quad \Rightarrow \quad \varepsilon = 1 - \frac{\omega_{\mathrm{pe}}^2}{\omega^2}. \tag{3.5}$$

In other words, the electric permittivity ε depends only on the incident frequency ω and does not depend on the wavenumber $\boldsymbol{k}$. This is equivalent to neglecting the phenomenon of spatial dispersion in the plasma. It can indeed be neglected if $a \ll \lambda$, where a is a characteristic length and λ is the wavelength in the plasma. For the characteristic length we are taking the distance traveled by the electron during one period of fast oscillation, $a = 2\pi v_{\mathrm{e}}/\omega$, and $\lambda = 2\pi v_{\mathrm{ph}}/\omega = 2\pi/k$, where $k = |\boldsymbol{k}|$ and v_{ph} is the phase speed of the waves. Hence, we need to require that the phase speed be much faster than the thermal speed of the electrons:

$$v_{\mathrm{ph}} = \frac{\omega}{k} \gg v_{\mathrm{e}} = \sqrt{\frac{3\kappa T}{2m_{\mathrm{e}}}}, \tag{3.6}$$

which is also equivalent to requiring that $kd \ll \omega/\omega_{\mathrm{pe}}$, where d is the Debye shielding length. The meaning of the term *cold plasma* can be explained with the help of

relation (3.6). Namely, the temperature should be sufficiently low so that the thermal speed is much slower that the phase speed of the waves.

Finally, by substituting expression (3.4) into the right-hand side of (3.1), we obtain

$$\frac{\partial^2 \boldsymbol{E}}{\partial t^2} + c^2 \mathrm{curlcurl}\boldsymbol{E} + \omega_{\mathrm{pe}}^2 \boldsymbol{E} = \boldsymbol{0}. \tag{3.7}$$

Equation (3.7) is a self-contained governing equation for the electric field $\boldsymbol{E}$. It no longer includes any other unknown quantities that need to be determined through additional considerations. Equation (3.7) admits different types of propagating waves that we are going to analyze.

3.3. Longitudinal and transverse waves. According to the Helmholtz theorem (see [22, section 1.5]), any vector field has a unique representation as a sum of its irrotational (longitudinal) and solenoidal (transverse) components. In other words, we can write

$$\boldsymbol{E} = \boldsymbol{E}_{\shortparallel} + \boldsymbol{E}_{\perp}, \quad \text{where} \quad \mathrm{curl}\boldsymbol{E}_{\shortparallel} = \boldsymbol{0} \quad \text{and} \quad \mathrm{div}\boldsymbol{E}_{\perp} = 0. \tag{3.8}$$

Note that calling the curl-free and divergence-free parts of the field by their alternative names—the longitudinal and transverse components, respectively—has a clear physical interpretation. Namely, in the frequency domain a plane wave propagating in an isotropic medium has the form $\boldsymbol{E} \sim e^{i\omega t + i\boldsymbol{k}\cdot\boldsymbol{r}}$, where $\boldsymbol{r}$ is the radius vector. Then, clearly, $\mathrm{curl}\boldsymbol{E} \sim \boldsymbol{k} \times \boldsymbol{E}$ and $\mathrm{div}\boldsymbol{E} \sim \boldsymbol{k} \cdot \boldsymbol{E}$. As such, $\mathrm{curl}\boldsymbol{E}_{\shortparallel} = \boldsymbol{0}$ would mean that $\boldsymbol{k} \times \boldsymbol{E}_{\shortparallel} = \boldsymbol{0}$, or in other words, that $\boldsymbol{E}_{\shortparallel}$ is parallel to the wave vector $\boldsymbol{k}$, which justifies its name of the longitudinal component. Similarly, $\mathrm{div}\boldsymbol{E}_{\perp} = 0$ would imply that $\boldsymbol{k} \cdot \boldsymbol{E}_{\perp} = 0$, or in other words, that $\boldsymbol{E}_{\perp}$ is perpendicular to the wave vector $\boldsymbol{k}$, which justifies its name of the transverse component.

Let us consider the longitudinal waves first. In this case, (3.7) reduces to

$$\frac{\partial^2 \boldsymbol{E}_{\shortparallel}}{\partial t^2} + \omega_{\mathrm{pe}}^2 \boldsymbol{E}_{\shortparallel} = \boldsymbol{0}. \tag{3.9}$$

Equation (3.9) governs the so-called Langmuir waves in plasma. As there is no spatial differentiation in (3.9), the Langmuir waves can basically be interpreted as high-frequency oscillations of the entire volume of plasma. The dispersion relation for the Langmuir waves is straightforward: $\omega^2 = \omega_{\mathrm{pe}}^2$, which means that the oscillations always occur with one and the same frequency $\omega_{\mathrm{pe}} = \sqrt{\frac{4\pi e^2 n_{\mathrm{e}}}{m_{\mathrm{e}}}}$. Accordingly, the group velocity of these waves is zero: $v_{\mathrm{gr}} \overset{\mathrm{def}}{=} \frac{\partial \omega}{\partial k} = 0$, which means that no energy transport is associated with the Langmuir waves.

On the other hand, propagation of the Langmuir waves is accompanied by perturbations of the local electric neutrality of the plasma. Indeed, according to the second steady-state Maxwell equation (2.10), when there are no extraneous charges we have

$$\rho_{\mathrm{ind}} = \frac{1}{4\pi}\mathrm{div}\boldsymbol{E} = \frac{1}{4\pi}\mathrm{div}\boldsymbol{E}_{\shortparallel},$$

and, consequently, the density of the induced charge ρ_{ind} undergoes oscillations with the frequency ω_{pe}, because it is governed by the same differential equation as (3.9):

$$\frac{\partial^2 \rho_{\mathrm{ind}}}{\partial t^2} + \omega_{\mathrm{pe}}^2 \rho_{\mathrm{ind}} = 0.$$

Let us reemphasize that the foregoing considerations are valid only when the phase velocity of the waves is large; see (3.6). By substituting $\omega = \omega_{\mathrm{pe}}$ we obtain $v_{\mathrm{ph}} = \omega_{\mathrm{pe}}/k \gg \omega_{\mathrm{pe}} d = v_{\mathrm{e}}$, which means $kd \ll 1$, or in other words, the wavelength must be much greater than the Debye shielding length: $\lambda \gg d$. If this constraint does not hold, i.e., if $kd \sim 1$, then $v_{\mathrm{ph}} \sim v_{\mathrm{e}}$, and the assumption of cold plasma breaks down. In this case, the dispersion relation of the plasma can be obtained only by solving the kinetic equation. As shown, e.g., in [21, Chapter 13], the Langmuir waves become dispersive for slower phase speeds: $\omega^2 = \omega_{\mathrm{pe}}^2 + 3k^2 v_{\mathrm{e}}^2$. Going even further down in the phase speed, i.e., allowing for $v_{\mathrm{ph}} \ll v_{\mathrm{e}}$, would necessitate taking the ions' motion into account; this leads to the ion sound that has been briefly discussed in section 2.5.

Having provided this very concise account of longitudinal oscillations, we will next turn to the primary subject of our discussion, the transverse high-frequency waves.

3.4. Transverse waves. To study the evolution of the transverse component $\boldsymbol{E}_\perp$ of the electric field, we first notice that $\mathrm{div}\boldsymbol{E}_\perp = 0$ implies $\mathrm{curlcurl}\boldsymbol{E}_\perp = -\Delta \boldsymbol{E}_\perp$, and, consequently, (3.7) transforms into the well-known Klein–Gordon equation

$$\frac{\partial^2 \boldsymbol{E}_\perp}{\partial t^2} - c^2 \Delta \boldsymbol{E}_\perp + \omega_{\mathrm{pe}}^2 \boldsymbol{E}_\perp = \boldsymbol{0}. \tag{3.10}$$

The dispersion relation for the Klein–Gordon equation (3.10) is easy to obtain. It reads

$$\omega^2 = \omega_{\mathrm{pe}}^2 + c^2 k^2, \tag{3.11}$$

which, in particular, means that similarly to the previous longitudinal case (see section 3.3), only high-frequency transverse waves can propagate in the plasma governed by (3.10). The range of allowable frequencies that corresponds to (3.11) is defined as $\omega > \omega_{\mathrm{pe}}$.

From relation (3.11), one can easily obtain the phase speed and the group speeds of the waves:

$$v_{\mathrm{ph}} = c\left(1 + \omega_{\mathrm{pe}}^2/c^2k^2\right)^{\frac{1}{2}} > c, \tag{3.12}$$

$$v_{\mathrm{gr}} = c\left(1 + \omega_{\mathrm{pe}}^2/c^2k^2\right)^{-\frac{1}{2}} < c. \tag{3.13}$$

Unlike in the longitudinal case of section 3.3, the propagation of transverse waves preserves the local electric neutrality of the plasma, because $\mathrm{div}\boldsymbol{E}_\perp = 0$. Moreover, it is possible to show (see [21, Chapter 13]), that even if one employs kinetic considerations for the analysis of transverse waves with a slow phase speed, $\omega/k \ll v_{\mathrm{e}}$, there will, in fact, be no such waves. In other words, there are no thermal transverse modes analogous to the thermal longitudinal modes.

The dispersion properties of high-frequency transverse waves are of particular interest. Let us first assume that $\frac{\omega_{\mathrm{pe}}}{ck} = \frac{v_{\mathrm{e}}}{ckd} \ll 1$, which implies that $kd \gg \frac{v_{\mathrm{e}}}{c} \approx 10^{-3}$, or in other words, that the waves are short: $\lambda \ll 10^3 d$, with the wavelength much shorter than a thousand times the Debye shielding length. These waves exhibit

a weakly dispersive behavior, as substitution of $\frac{\omega_{\rm pe}}{ck} \ll 1$ into (3.12) and (3.13) immediately yields

$$v_{\rm ph} \approx c\left(1+\frac{\omega_{\rm pe}^2}{2c^2k^2}\right), \tag{3.14}$$

$$v_{\rm gr} \approx c\left(1-\frac{\omega_{\rm pe}^2}{2c^2k^2}\right). \tag{3.15}$$

We indeed see that both the phase speed $v_{\rm ph}$ of (3.14) and the group speed $v_{\rm gr}$ of (3.15) are close to the speed of light c, with the former being slightly faster than c and the latter being slightly slower than c. The frequency in this case, according to (3.11), is approximately equal to the speed of light times the wavenumber ($\omega \approx ck \gg \omega_{\rm pe}$), and is also much higher than the Langmuir frequency. Note that the ultimate case of $v_{\rm ph} = v_{\rm gr} = c$, $\omega = ck$, would correspond to the propagation of waves with no dispersion in the framework of a pure d'Alembert equation rather than the Klein–Gordon equation.

In contradistinction to the short waves, the long transverse waves governed by (3.10) are similar to the longitudinal waves. Indeed, let $\frac{\omega_{\rm pe}}{ck} \gg 1$; it means that $\lambda \gg 10^3 d$ and also that $\omega \gtrsim \omega_{\rm pe}$, i.e., that the waves propagate with the frequencies close to the lowest possible frequency $\omega_{\rm pe}$. In this case, $v_{\rm ph} \approx \omega_{\rm pe}/k$, and $v_{\rm gr} = c\cdot\frac{ck}{\omega_{\rm pe}} \ll c$; i.e., the expression for the phase velocity is basically the same as that in the longitudinal case (see section 3.3), while the group velocity is small (in the pure longitudinal case it is equal to zero). This behavior is not surprising because the longer the wave, the less of a spatial variation per unit length it undergoes, and, consequently, the more the corresponding oscillations should resemble the oscillations of the entire plasma volume as a whole, which are characteristic of the longitudinal case.

In general, we should mention that the foregoing dispersion properties, while not completely unparalled, are, perhaps, still less typical than the inverse situation, when the long waves, rather than the short waves, exhibit a weakly dispersive behavior; see [15]. Our primary goal, however, is to see what can be said about the lacunae and the Huygens' principle for the waves governed by (3.10). From the previous considerations we conclude that *it is for the short waves, which are only weakly dispersive, that one can possibly observe some sort of "lacunae"* in the solutions of (3.10). Indeed, in this case the propagation speeds (3.12) and (3.13) (see also (3.14) and (3.15)) are close to the nondispersive propagation velocity c, and therefore one may expect to see relatively few aftereffects behind what would have been the sharp aft fronts in the genuine Huygens' case. To provide a somewhat more accurate yet still qualitative argument, let us consider the waves propagating from an instantaneous point source located at the origin. Given the distance to the source r and the moment of time t, one can easily see that only those waves that have the group velocity $v_{\rm gr} = r/t$ can reach the location r precisely at the moment t. Using expression (3.13) for the group velocity, we obtain a formula for k as it depends on r and t:

$$k = \frac{\omega_{\rm pe}}{c}\left(\frac{c^2t^2}{r^2}-1\right)^{-\frac{1}{2}}. \tag{3.16}$$

We see that the wavenumbers are defined only inside the light cone $r \le ct$. Formula (3.16) also indicates that for a given moment of time t, the larger the r, the larger the k.

In other words, the closer the value of r to ct, the shorter the wave that reaches this location at time t, and ultimately, for the purely nondispersive propagation $r = ct$ the wavelength $\lambda = 2\pi/k$ defined by (3.16) becomes equal to zero.

Let us now fix some large wavenumber $k_1 \gg \frac{\omega_{\rm pe}}{c}$ and consider a wave packet propagating from the origin with the range of wavenumbers $k \geq k_1$. By noticing that the group velocity $v_{\rm gr}$ of (3.13) is a monotone increasing function of k, we conclude that the range of group velocities for this packet will be

$$c\left(1+\omega_{\rm pe}^2/c^2k_1^2\right)^{-\frac{1}{2}} \leq v_{\rm gr} < c.$$

Therefore, at every given moment of time t we can easily estimate how wide this packet is going to be. The width of the packet can be thought of as the spatial extent of the "tail" behind the aft front $r = ct$:

$$\delta_{\rm tail} = (c - \min_k v_{\rm gr})t = c\left[1 - \left(1 + \frac{\omega_{\rm pe}^2}{c^2k_1^2}\right)^{-\frac{1}{2}}\right] t \approx ct \cdot \frac{\omega_{\rm pe}^2}{2c^2k_1^2}. \tag{3.17}$$

We see that the tail expands linearly with time and shrinks quadratically as the minimum borderline wavenumber k_1 for the packet increases. We also note that the short waves, as they are defined above, $k \gg \frac{\omega_{\rm pe}}{c}$, propagate with high frequencies $\omega \approx ck \gg \omega_{\rm pe}$. Therefore, we can equivalently reformulate our general expectation in terms of the frequency rather than the wavelength. Namely, we hope that lacunae could be approximately observed in the solutions of (3.10) for high frequencies $\omega \gg \omega_{\rm pe}$, whereas the overall range of frequencies allowed by the dispersion relation (3.11) is $\omega > \omega_{\rm pe}$. Using the dispersion relation (3.11), we can also recast estimate (3.17) for the width of the aftereffects region (the tail) as

$$\delta_{\rm tail} \approx ct \cdot \frac{\omega_{\rm pe}^2}{2\omega_1^2}, \tag{3.18}$$

where $\omega_1^2 = \omega_{\rm pe}^2 + c^2k_1^2$ is the minimum borderline frequency for the packet we are considering: $\omega \geq \omega_1 \gg \omega_{\rm pe}$. We also note that the ratio of the Langmuir frequency over the driving frequency of the waves that appears in formula (3.18) is going to play a key role in our subsequent analysis.

Let us emphasize, however, that the entire discussion based on the dispersion relation (3.11) is basically conducted in the frequency domain. On the other hand, we have seen in section 1.1 that the frequency domain is inadequate for the analysis of lacunae and the Huygens' principle. A time-domain analysis is needed in order to see how the Huygens' principle can be interpreted for the weakly dispersive transverse waves governed by (3.10).

Consider a three-dimensional Cauchy problem for the inhomogeneous Klein–Gordon equation (cf. (1.1)):

$$\frac{\partial^2\varphi}{\partial t^2} - c^2\Delta\varphi + \omega_{\rm pe}^2\varphi = f(\boldsymbol{x},t), \qquad \varphi(\boldsymbol{x},0) = \varphi_t(\boldsymbol{x},0) = 0, \tag{3.19}$$
$$\mathbb{R}^3 \ni \boldsymbol{x} = (x_1,x_2,x_3).$$

Compared to the vector equation (3.10), the differential equation in (3.19) is scalar and may govern, e.g., one Cartesian component of the total field. The right-hand side $f(\boldsymbol{x},t)$ may be due to the extraneous current.

The Klein–Gordon equation is obviously not equivalent to the d'Alembert equation, and therefore, according to the Matthisson criterion [20], its solutions must be diffusive and may have no lacunae in the classical sense of the word. The discrepancy between the two equations is accounted for by the term $\omega_{\rm pe}^2\varphi$ in (3.10). This term is responsible for the onset of dispersion that ruins the lacunae. We still hope, though, that the behavior of solutions to (3.10) will be close to Huygens' when the dispersion is weak. Therefore, while it is clear that the term $\omega_{\rm pe}^2\varphi$ in the Klein–Gordon equation may not be completely disregarded, we would nonetheless like to see when it can be legitimately classified as "small." Note that it is not as straightforward as simply calling the coefficient $\omega_{\rm pe}^2$ small, because this coefficient is is *not dimensionless.* As such, we would rather need to identify special classes of solutions $\varphi=\varphi(x,t)$, for which the entire term $\omega_{\rm pe}^2\varphi$ can be deemed small. The previous frequency-domain considerations suggest that this may be the case when a high driving frequency $\omega\gg\omega_{\rm pe}$ is brought into the time-domain analysis.

The fundamental solution for the Klein–Gordon operator can be obtained in the closed form (see [4]):

$$\mathcal{E}(\boldsymbol{x},t)=\underbrace{\frac{\Theta(t)}{2\pi c}\delta\left(\beta^2\right)}_{\mathcal{E}_1(\boldsymbol{x},t)}-\underbrace{\frac{\omega_{\rm pe}^2}{4\pi c^3}\Theta(t)\Theta\left(\beta^2\right)\frac{J_1(y)}{y}}_{\mathcal{E}_2(\boldsymbol{x},t)}, \tag{3.20}$$

where $\beta^2=c^2t^2-|\boldsymbol{x}|^2$, $y=\frac{\omega_{\rm pe}}{c}\beta$, $J_1(\,\cdot\,)$ is the Bessel function, and $\Theta(\,\cdot\,)$ denotes the Heaviside function, as before. The first term $\mathcal{E}_1(\boldsymbol{x},t)$ on the right-hand side of formula (3.20) is the same as the fundamental solution of the d'Alembert operator; see (1.2). The second term $\mathcal{E}_2(\boldsymbol{x},t)$ can be interpreted as a correction due to the presence of $\omega_{\rm pe}^2\varphi$ in (3.19). Accordingly, solution $\varphi=\varphi(\boldsymbol{x},t)$ of the Cauchy problem (3.19) is given by the convolution

$$\varphi=\mathcal{E}*f=\mathcal{E}_1*f-\mathcal{E}_2*f=\varphi_1-\varphi_2, \tag{3.21}$$

where the first term $\varphi_1=\mathcal{E}_1*f$ on the right-hand side of (3.21) is the Kirchhoff integral (cf. formula (1.3)), while the second term $\varphi_2=\mathcal{E}_2*f$ is basically what "contaminates" the lacuna. We are going to study the properties of exactly this contaminating term for a particular choice of f.

Namely, we will consider the following point excitation for problem (3.19):

$$f(\boldsymbol{x},t)=\begin{cases}M\cdot\delta(\boldsymbol{x})\cdot\sin(\omega t)\equiv\delta(\boldsymbol{x})\tilde{f}(t), & 0\le t\le T,\\ 0, & t<0 \text{ and } t>T,\end{cases} \tag{3.22}$$

where $M>0$ and $T>0$ are two parameters and ω denotes the driving frequency. We will assume that the source (3.22) undergoes sufficiently many oscillations with frequency ω during the interval $0\le t\le T$. At the same time, this interval still remains finite, which allows us to preserve the time-dependent nature of the problem rather than have it transformed into the frequency domain. Choosing the right-hand side $f(\boldsymbol{x},t)$ of (3.19) in the form (3.22) enables us to perform a sufficiently straightforward analysis on one hand, and, on the other hand, it still allows us to illustrate the key phenomena of interest.

According to the definition of the fundamental solution (see (3.20)), we have

(3.23)
$$\varphi_2 = \frac{\omega_{\rm pe}^2}{4\pi c^3}\int_0^t \iiint_{|\boldsymbol{x}-\boldsymbol{\xi}|\le c|t-\tau|} \frac{f(\boldsymbol{\xi},\tau)J_1\left(\omega_{\rm pe}\sqrt{(t-\tau)^2-|\boldsymbol{x}-\boldsymbol{\xi}|^2/c^2}\right)}{\omega_{\rm pe}\sqrt{(t-\tau)^2-|\boldsymbol{x}-\boldsymbol{\xi}|^2/c^2}}d\boldsymbol{\xi}d\tau$$
$$= \frac{\omega_{\rm pe}^2}{4\pi c^3}\int_0^{T_1} \frac{\tilde f(\tau)J_1\left(\omega_{\rm pe}\sqrt{(t-\tau)^2-|\boldsymbol{x}|^2/c^2}\right)}{\omega_{\rm pe}\sqrt{(t-\tau)^2-|\boldsymbol{x}|^2/c^2}}d\tau = \frac{\omega_{\rm pe}^2}{4\pi c^3}\int_0^{T_1}\frac{\tilde f(\tau)J_1(y)}{y}d\tau,$$

where $T_1 = \min\{(t-|\boldsymbol{x}|/c), T\}$, $y = y(\tau,t,\boldsymbol{x}) = \omega_{\rm pe}\sqrt{(t-\tau)^2-|\boldsymbol{x}|^2/c^2}$, and $\tilde f(\tau)$ denotes the temporal dependence of the source term (3.22): $\tilde f(\tau) = M\sin(\omega\tau)$. We will analyze the cases of small and large arguments y of the Bessel function J_1 in formula (3.24). Let us first note that if y is small, or more precisely, if $0 \le y \le \mu_1^{(2)}$, where $\mu_1^{(2)}$ is the first positive root of the Bessel function $J_2(y)$, then the function

$$G(y) \stackrel{\rm def}{=} \frac{J_1(y)}{y}$$

is a monotone decreasing function of the argument y. Indeed, we have

$$G'(y) = \frac{d}{dy}\left[\frac{J_1(y)}{y}\right] = -\frac{J_2(y)}{y} \le 0 \quad \text{if } y\in[0,\mu_1^{(2)}].$$

The inequality $0 \le y \le \mu_1^{(2)}$ implies a constraint on the maximum value of t. In the worst-case scenario—$\tau = 0$ and $|\boldsymbol{x}| = 0$—this constraint reads

$$t \le \mu_1^{(2)}/\omega_{\rm pe} \equiv T_0, \tag{3.24}$$

and from here on we will require that the most conservative sufficient condition (3.24) hold in order to guarantee that the value of y be sufficiently small.

We also notice that the function $y = y(\tau,\,\cdot\,)$ is a monotone decreasing function of its argument τ on the interval $0 \le \tau \le T_1$. Consequently, the composite function $\tilde G(\tau) = G(y(\tau,\,\cdot\,))$ is a monotone increasing function of τ. We can then apply the Bonnet theorem (second mean value theorem) (see [35]), to the last integral from (3.24) and obtain

$$\varphi_2 = \frac{\omega_{\rm pe}^2}{4\pi c^3}\left[\tilde G(0)\int_0^\eta \tilde f(\tau)d\tau + \tilde G(T_1)\int_\eta^{T_1}\tilde f(\tau)d\tau\right], \tag{3.25}$$

where η is some point of the interval $[0,\,T_1]$.

We note that the contaminating part φ_2 of the solution will eventually need to be compared against its regular part φ_1, which, according to (1.3), is given by

$$\varphi_1(\boldsymbol{x},t) = \frac{1}{4\pi c^2}\frac{\tilde f(t-|\boldsymbol{x}|/c)}{|\boldsymbol{x}|}, \tag{3.26}$$

where, again, $\tilde f(t) = M\sin(\omega t)$ for $t\in[0,T]$; see (3.22). The function φ_1 of (3.26) represents a genuine d'Alembert wave packet due to the source (3.22); it may differ from zero only on the region $c(t-T) \le |\boldsymbol{x}| \le ct$. For $|\boldsymbol{x}| > ct$ we have $\varphi_1(\boldsymbol{x},t) = 0$

because the propagation speed is finite, and for $|\boldsymbol{x}| < c(t-T)$ we have $\varphi_1(\boldsymbol{x},t) = 0$ because this is a lacuna of the wave equation.

Similarly, for $|\boldsymbol{x}| > ct$ we also have $\varphi_2(\boldsymbol{x},t) = 0$. However, otherwise $\varphi_2(\boldsymbol{x},t) \neq 0$ either in the wave packet area $c(t-T) \le |\boldsymbol{x}| \le ct$ or in the lacuna area $|\boldsymbol{x}| < c(t-T)$. The wave packet area corresponds to $T_1 = (t - |\boldsymbol{x}|/c) \le T$; then $y(T_1, \cdot\,) = 0$, and, consequently, $\tilde{G}(T_1) = G(0) = 1/2$ in formula (3.25). In contradistinction to that, the area that would have been a lacuna in the nondispersive case corresponds to $T_1 = T < t - |\boldsymbol{x}|/c$, which means $y(T_1, \cdot\,) = y(T, \cdot\,) > 0$ and $\tilde{G}(T_1) < 1/2$. Altogether, the constants in formula (3.25) can be estimated as follows (recall that $\mu_1^{(2)} \approx 5.13562230$):

$$-6.61397437 \cdot 10^{-2} = G(\mu_1^{(2)}) \le \tilde{G}(0) < \tilde{G}(T_1) \le G(0) = \frac{1}{2}. \tag{3.27}$$

By evaluating the integrals in (3.25), we obtain

$$\begin{aligned} \varphi_2 &= \frac{\omega_{\mathrm{pe}} M}{4\pi c^3} \frac{\omega_{\mathrm{pe}}}{\omega} \Big[\tilde{G}(0)\,(1 - \cos(\omega\eta)) + \tilde{G}(T_1)\,(\cos(\omega\eta) - \cos(\omega T_1)) \Big] \\ &= \frac{\omega_{\mathrm{pe}} M}{4\pi c^3} \frac{\omega_{\mathrm{pe}}}{\omega} \Big[\tilde{G}(0) + (\tilde{G}(T_1) - \tilde{G}(0)) \cos(\omega\eta) - \tilde{G}(T_1) \cos(\omega T_1) \Big], \end{aligned} \tag{3.28}$$

and according to estimates (3.27), the absolute value of the quantity in rectangular brackets in formula (3.28) may never exceed $3/2 - G(\mu_1^{(2)})$.

Let us now compare the dispersionless solution φ_1 of (3.26) with the dispersion-induced correction φ_2 of (3.28). Note that φ_1 is defined only inside the wave packet area, $c(t-T) \le |\boldsymbol{x}| \le ct$, including the aft front $|\boldsymbol{x}| = c(t-T)$. We can then recast formula (3.26) as

$$\varphi_1(\boldsymbol{x},t) = \frac{M}{4\pi c^3} \frac{\sin(\omega T_1)}{t - T_1} \tag{3.29}$$

and thus obtain

$$\frac{\sup\limits_{|\boldsymbol{x}| \le ct} |\varphi_2(\boldsymbol{x},t)|}{\sup\limits_{c(t-T) \le |\boldsymbol{x}| \le ct} |\varphi_1(\boldsymbol{x},t)|} = \Big(\tfrac{3}{2} - G(\mu_1^{(2)}) \Big)\, \omega_{\mathrm{pe}} (t - T_1) \frac{\omega_{\mathrm{pe}}}{\omega}. \tag{3.30}$$

In formula (3.29), we can always consider $\omega_{\mathrm{pe}}(t - T_1) < \mu_1^{(2)}$ because of inequality (3.24). As such,

$$\frac{\sup |\varphi_2|}{\sup |\varphi_1|} = \mathcal{O}\left(\frac{\omega_{\mathrm{pe}}}{\omega} \right). \tag{3.31}$$

Estimate (3.31) is important as it quantifies the previously outlined "tentative" consideration that the higher the driving frequency, the more of a lacuna one might be able to observe in the corresponding solution. It is because of this particular estimate (see (3.31)) that we can call the region $|\boldsymbol{x}| < c(t-T)$ for $t \le T_0$ *a weak lacuna* and also refer to the quantity on the left-hand side of (3.31) as its *"depth."* Indeed, the region $|\boldsymbol{x}| < c(t-T)$ corresponds to the genuine lacuna of the d'Alembert equation. In the dispersive case, there is still a residual field inside this region, but its magnitude relative to the magnitude of the field in the packet (the depth of a weak lacuna) is

small and, quantitatively, is proportional to the ratio of the Langmuir frequency over the driving frequency of the waves.

Next, we will consider the opposite case—that of the large argument y of the Bessel function J_1 in formula (3.24). Our goal will be to justify a relation similar to (3.31) for long propagation times.

Let $y \gg 1$. Then, we will use the asymptotic form of the Bessel function $J_1(y)$,

$$J_1(y) = \sqrt{\frac{2}{\pi y}} \cos\left(y - \frac{3\pi}{4}\right) + \mathcal{O}\left(y^{-\frac{3}{2}}\right), \tag{3.32}$$

which means that by disregarding the higher order terms $\mathcal{O}(y^{-\frac{5}{2}})$ in the integral (3.24) we can recast it as

$$\varphi_2 \approx \frac{\omega_{\rm pe}^2}{4\pi c^3} \sqrt{\frac{2}{\pi}} \int_0^{T_1} \tilde{f}(\tau) y^{-\frac{3}{2}} \cos\left(y - \frac{3\pi}{4}\right) d\tau. \tag{3.33}$$

We would like to estimate the magnitude of $\varphi_2(\boldsymbol{x}, t)$ of (3.33) for $|\boldsymbol{x}| < c(t-T)$, i.e., inside the region that would have been a lacuna in the nondispersive case. This means that the upper integration limit in formula (3.33) can be taken as $T_1 = T$.

Let us first analyze the expression for $y = y(\tau, t, \boldsymbol{x}) = \omega_{\rm pe}\sqrt{(t-\tau)^2 - |\boldsymbol{x}|^2/c^2}$ that enters into formulae (3.32) and (3.33) and see under what conditions it can indeed be regarded as large. Obviously, as $\tau \in [0, T]$, then $\min_\tau y(\tau, t, \boldsymbol{x}) = y(T, t, \boldsymbol{x})$, and it will be sufficient to see when $y(T, t, \boldsymbol{x})$ is large. To begin with, we notice that for a given moment of time t, the quantity $y(T, t, \boldsymbol{x})$ cannot be large all across the lacuna, because on the aft front $|\boldsymbol{x}| = c(t-T)$ we have $y(T, t, c(t-T)) = 0$. Consequently, to be able to legitimately use the asymptotics (3.32) we will need to step inside the lacuna.

Then we introduce the distance δ between a given point inside the lacuna and the aft front at the moment of time, t. For $|\boldsymbol{x}| = c(t-T) - \delta$ we have $y = \frac{\omega_{\rm pe}}{c}\sqrt{2c(t-T)\delta - \delta^2}$. We can therefore conclude that if we consider δ as a function of time, $\delta = \delta(t)$, and require that

$$\lim_{t\to\infty} [2c(t-T)\cdot\delta(t) - \delta^2(t)] = \infty,$$

then the quantity $y = y(T, t, \boldsymbol{x})$ will increase with no bound when $t \longrightarrow \infty$ and $|\boldsymbol{x}| \le c(t-T) - \delta(t)$. Clearly, in so doing the "gap width" δ itself may even decrease as t increases, but only more slowly than $(t-T)^{-1}$. On the other hand, δ may also be a constant or an increasing function of the argument t; in the latter case it may not increase faster than linearly because the lacuna itself expands only linearly with respect to time.

To summarize, we can claim that

$$\lim_{t\to\infty} y(T, t, \boldsymbol{x}) = \infty$$

uniformly for all $\boldsymbol{x}$ such that $|\boldsymbol{x}| \le c(t-T) - \delta(t)$, provided that

$$\frac{\text{const}}{(t-T)\zeta(t)} \le \delta(t) \le (c - c_1)(t-T), \tag{3.34}$$

where $c_1 < c$ and $\zeta(t)$ is an auxiliary function such that $\zeta(t) = o(1)$ as $t \longrightarrow \infty$. Clearly, the most conservative strategy for choosing the gap width, $\delta = (c-c_1)(t-T)$,

where $c_1 < c$, will guarantee the fastest growth of y in a narrower cone $|\boldsymbol{x}| < c_1(t-T)$:

$$\begin{aligned} &\forall \boldsymbol{x}: \ |\boldsymbol{x}| \le c_1(t-T), \ c_1 < c \ \ \& \ \ \forall \tau \in [0,T] \ : \\ &y(\tau,t,\boldsymbol{x}) \ge y(T,t,\boldsymbol{x}) \ge \omega_{\rm pe}\sqrt{1-c_1^2/c^2}(t-T). \end{aligned} \tag{3.35}$$

Estimate (3.35) will allow us to use the asymptotic formulae (3.32) and (3.33) for sufficiently large times t.

Next, we notice that $y^{-\frac{3}{2}}$ is a monotone decreasing function of y for $y > 0$, and as $y = y(\tau,t,\boldsymbol{x})$ is, in turn, a monotone decreasing function of τ for $\tau \in [0,T]$, it follows that $y^{-\frac{3}{2}}$ is a monotone increasing function of τ. Consequently, we can apply the Bonnet theorem again, this time to the integral (3.33), and obtain (recall that $T_1 = T$ for the interior of the lacuna)

$$\begin{aligned} \varphi_2 \approx \frac{\omega_{\rm pe}^2}{4\pi c^3}\sqrt{\frac{2}{\pi}}\Bigg[&(y(0,t,\boldsymbol{x}))^{-\frac{3}{2}}\int_0^{\eta}\tilde{f}(\tau)\cos\left(y(\tau,t,\boldsymbol{x})-\frac{3\pi}{4}\right)d\tau \\ &+(y(T,t,\boldsymbol{x}))^{-\frac{3}{2}}\int_{\eta}^{T}\tilde{f}(\tau)\cos\left(y(\tau,t,\boldsymbol{x})-\frac{3\pi}{4}\right)d\tau\Bigg], \end{aligned} \tag{3.36}$$

where $\eta \in [0,T]$. Let us now substitute $\tilde{f}(\tau) = M\sin(\omega\tau)$ into (3.36):

$$\begin{aligned} \varphi_2 \approx \frac{M\omega_{\rm pe}^2}{8\pi c^3}\sqrt{\frac{2}{\pi}}\Bigg[&(y(0,t,\boldsymbol{x}))^{-\frac{3}{2}}\int_0^{\eta}\left\{\sin\left(\omega\tau+y(\tau,t,\boldsymbol{x})-\frac{3\pi}{4}\right)\right. \\ &\left.-\sin\left(\omega\tau-y(\tau,t,\boldsymbol{x})+\frac{3\pi}{4}\right)\right\}d\tau \\ &+(y(T,t,\boldsymbol{x}))^{-\frac{3}{2}}\int_{\eta}^{T}\left\{\sin\left(\omega\tau+y(\tau,t,\boldsymbol{x})-\frac{3\pi}{4}\right)\right. \\ &\left.-\sin\left(\omega\tau-y(\tau,t,\boldsymbol{x})+\frac{3\pi}{4}\right)\right\}d\tau\Bigg]. \end{aligned}$$

The argument $\left(\omega\tau \pm y(\tau,t,\boldsymbol{x}) \mp \frac{3\pi}{4}\right)$ of the sine functions above can be approximated as follows. Denote $\nu = T-\tau$, $0 \le \nu \le T$, and recast y in the form

$$y(\tau,t,\boldsymbol{x}) = \omega_{\rm pe}\sqrt{(t-T)^2-|\boldsymbol{x}|^2/c^2}\sqrt{1+\frac{2(t-T)\nu+\nu^2}{(t-T)^2-|\boldsymbol{x}|^2/c^2}}. \tag{3.37}$$

Notice that if

$$\frac{2(t-T)\nu}{(t-T)^2-\frac{|\boldsymbol{x}|^2}{c^2}} = \frac{2\nu}{(t-T)\left(1-\frac{|\boldsymbol{x}|^2}{(t-T)^2c^2}\right)} \ll 1,$$

then also

$$\begin{aligned} \frac{\nu^2}{(t-T)^2-\frac{|\boldsymbol{x}|^2}{c^2}} &= \frac{\nu^2}{(t-T)^2\left(1-\frac{|\boldsymbol{x}|^2}{(t-T)^2c^2}\right)} \\ &= \left[\frac{\nu}{(t-T)\left(1-\frac{|\boldsymbol{x}|^2}{(t-T)^2c^2}\right)}\right]^2\left(1-\frac{|\boldsymbol{x}|^2}{(t-T)^2c^2}\right) \ll 1. \end{aligned}$$

Consequently, if the linear term with respect to ν under the second square root in formula (3.37) is indeed small, then the quadratic term can be disregarded, which yields

$$y(\tau,t,\boldsymbol{x}) \approx y(T,t,\boldsymbol{x}) + \frac{\omega_{\rm pe}^2(t-T)\nu}{y(T,t,\boldsymbol{x})} = \underbrace{y(T,t,\boldsymbol{x}) + \frac{\omega_{\rm pe}^2(t-T)T}{y(T,t,\boldsymbol{x})}}_{\text{does not depend on } \tau} - \frac{\omega_{\rm pe}^2(t-T)\tau}{y(T,t,\boldsymbol{x})}.$$

Therefore, we can write

$$\begin{aligned}\varphi_2 \approx \frac{M\omega_{\rm pe}^2}{8\pi c^3}\sqrt{\frac{2}{\pi}}\Bigg[&(y(0,t,\boldsymbol{x}))^{-\frac{3}{2}}\int_0^{\eta}\big\{\sin\left((\omega+\gamma\omega_{\rm pe})\tau-\alpha\right)\\ &\qquad -\sin\left((\omega+\gamma\omega_{\rm pe})\tau-\alpha\right)\big\}d\tau\\ &+(y(T,t,\boldsymbol{x}))^{-\frac{3}{2}}\int_{\eta}^{T}\big\{\sin\left((\omega-\gamma\omega_{\rm pe})\tau+\alpha\right)\\ &\qquad -\sin\left((\omega+\gamma\omega_{\rm pe})\tau-\alpha\right)\big\}d\tau\Bigg],\end{aligned}$$

where $\gamma = \frac{\omega_{\rm pe}(t-T)}{y(T,t,\boldsymbol{x})}$ and $\alpha = y(T,t,\boldsymbol{x}) + \frac{\omega_{\rm pe}^2(t-T)T}{y(T,t,\boldsymbol{x})} - \frac{3\pi}{4}$. The integrals can now be explicitly evaluated:

$$\begin{aligned}\varphi_2 \approx \frac{M\omega_{\rm pe}^2}{8\pi c^3}\sqrt{\frac{2}{\pi}}\Bigg[&(y(0,t,\boldsymbol{x}))^{-\frac{3}{2}}\left\{\frac{\cos\alpha-\cos((\omega-\gamma\omega_{\rm pe})\eta+\alpha)}{\omega-\gamma\omega_{\rm pe}}\right.\\ &\qquad\left.-\frac{\cos\alpha-\cos((\omega+\gamma\omega_{\rm pe})\eta-\alpha)}{\omega+\gamma\omega_{\rm pe}}\right\}\\ &+(y(T,t,\boldsymbol{x}))^{-\frac{3}{2}}\left\{\frac{\cos((\omega-\gamma\omega_{\rm pe})\eta+\alpha)-\cos((\omega-\gamma\omega_{\rm pe})T+\alpha)}{\omega-\gamma\omega_{\rm pe}}\right.\\ &\qquad\left.-\frac{\cos((\omega+\gamma\omega_{\rm pe})\eta-\alpha)-\cos((\omega+\gamma\omega_{\rm pe})T-\alpha)}{\omega+\gamma\omega_{\rm pe}}\right\}\Bigg],\end{aligned}$$

and using (3.35) we obtain

$$\begin{aligned}|\varphi_2(\boldsymbol{x},t)| &\le \frac{M\omega_{\rm pe}^2}{4\pi c^3}\sqrt{\frac{2}{\pi}}\left[(y(0,t,\boldsymbol{x}))^{-\frac{3}{2}}+(y(T,t,\boldsymbol{x}))^{-\frac{3}{2}}\right]\\ &\quad\cdot\left\{\frac{1}{\omega-\gamma\omega_{\rm pe}}+\frac{1}{\omega+\gamma\omega_{\rm pe}}\right\}\\ &\le \frac{M\omega_{\rm pe}^2}{\pi c^3}\sqrt{\frac{2}{\pi}}\omega_{\rm pe}^{-\frac{3}{2}}(t-T)^{-\frac{3}{2}}(1-c_1^2/c^2)^{-\frac{3}{4}}\frac{\omega}{\omega^2-\gamma^2\omega_{\rm pe}^2}.\end{aligned}$$

We also note that, according to (3.35), the quantity γ is bounded: $\gamma = \frac{\omega_{\rm pe}(t-T)}{y(T,t,\boldsymbol{x})} \le \frac{1}{\sqrt{1-c_1^2/c^2}}$. Then, assuming that $\omega \gg \omega_{\rm pe}$, we drop the quadratic term $\mathcal{O}(\frac{\gamma^2\omega_{\rm pe}^2}{\omega^2})$ and get

$$|\varphi_2(\boldsymbol{x},t)| \le \frac{M}{\pi c^3}\sqrt{\frac{2}{\pi}}\omega_{\rm pe}^{-\frac{1}{2}}(t-T)^{-\frac{3}{2}}(1-c_1^2/c^2)^{-\frac{3}{4}}\frac{\omega_{\rm pe}}{\omega}. \tag{3.38}$$

Estimate (3.38) for the correction φ_2 is valid inside the lacuna of the wave equation in a narrower cone $|\boldsymbol{x}| < c(t-T) - \delta(t) = c_1(t-T)$. As before, the magnitude of the correction φ_2 now needs to be compared against the magnitude of the solution φ_1 inside the wave packet. For the purpose of comparison, we will consider φ_1 given by (3.26) on the boundary of the lacuna, i.e., exactly at the aft front $|\boldsymbol{x}| = c(t-T)$:

$$\varphi_1(\boldsymbol{x},t) = \frac{M}{4\pi c^3}\frac{\sin(\omega T)}{t-T}.$$

Using estimate (3.38), we can therefore write (cf. formula (3.31))

$$\frac{\sup|\varphi_2|}{\sup|\varphi_1|} = \mathcal{O}\left(\omega_{\mathrm{pe}}^{-\frac{1}{2}}(t-T)^{-\frac{1}{2}}\frac{\omega_{\mathrm{pe}}}{\omega}\right). \tag{3.39}$$

From estimate (3.39) we see not only that for long propagation times the depth of a weak lacuna is controlled by the ratio $\frac{\omega_{\mathrm{pe}}}{\omega}$ (similar to the case of short times) but that it also decays with the rate proportional to the inverse square root of time. We need to remember, however, that whereas in the previous estimate (3.31) we could use the maximum of the residual field φ_2 all across the lacuna $|\boldsymbol{x}| < c(t-T)$, in estimate (3.39) it can be taken only across a narrower cone $|\boldsymbol{x}| < c(t-T) - \delta(t) = c_1(t-T)$; see formula (3.35).

Let us additionally note that if we were to allow regions wider than the cone $|\boldsymbol{x}| < c_1(t-T)$ when analyzing the rate of growth of y, i.e., if we were to take the gap width $\delta(t)$ increasing more slowly than $(c-c_1)(t-T)$ (see formula (3.34)), then we would still obtain the key quantification of the depth of the weak lacuna by means of $\frac{\omega_{\mathrm{pe}}}{\omega}$, but we could lose the additional decay $\sim \omega_{\mathrm{pe}}^{-\frac{1}{2}}(t-T)^{-\frac{1}{2}}$ for long propagation times. For example, let $\delta(t) = A(t-T)^{\frac{1}{3}}$, where A is an appropriate constant needed to take into account that t is time and δ is distance. Then, for large times t we would obviously have $\delta^2(t) \ll 2c(t-T)\delta(t)$ and, consequently, $y = \frac{\omega_{\mathrm{pe}}}{c}\sqrt{2c(t-T)\delta - \delta^2} \approx \frac{\omega_{\mathrm{pe}}}{c}\sqrt{2cA}(t-T)^{\frac{2}{3}}$. In other words, instead of (3.35) we obtain

$$\begin{gathered}\forall\boldsymbol{x}:\ |\boldsymbol{x}| \le c(t-T) - A(t-T)^{\frac{1}{3}} \ \ \&\ \ \forall\tau\in[0,T]: \\ y(\tau,t,\boldsymbol{x}) \ge y(T,t,\boldsymbol{x}) \gtrsim \frac{\omega_{\mathrm{pe}}}{c}\sqrt{2cA}(t-T)^{\frac{2}{3}}.\end{gathered} \tag{3.40}$$

Accordingly, estimate (3.38) gets replaced by

$$|\varphi_2(\boldsymbol{x},t)| \le \frac{M}{2\pi c^3}\sqrt{\frac{2}{\pi}}\omega_{\mathrm{pe}}^{-\frac{1}{2}}(t-T)^{-1}\left(\frac{2A}{c}\right)^{-\frac{3}{4}}\frac{\omega_{\mathrm{pe}}}{\omega}, \tag{3.41}$$

and instead of (3.39) we obtain a simpler relation (cf. formula (3.31)):

$$\frac{\sup|\varphi_2|}{\sup|\varphi_1|} = \mathcal{O}\left(\frac{\omega_{\mathrm{pe}}}{\omega}\right). \tag{3.42}$$

Clearly, estimate (3.39) guarantees a deeper lacuna for large times t than estimate (3.42) does. However, estimate (3.42) is valid on the region $|\boldsymbol{x}| < c(t-T) - A(t-T)^{\frac{1}{3}}$, which is wider than the cone $|\boldsymbol{x}| < c_1(t-T)$, $c_1 < c$, on which estimate (3.39) holds.

We should reemphasize, however, that both estimates (3.31) and (3.39) (as well as (3.42)) are only asymptotic results, for the small and large values, respectively, of the argument y of the Bessel function J_1 in formula (3.24). To corroborate and further expand the scope of these results, we will evaluate the convolution (3.24) numerically.

TABLE 3.1
The depth of the weak lacuna for different moments of time.

$\frac{\omega_{\mathrm{pe}}}{\omega}$	$\frac{1}{10}$	$\frac{1}{20}$	$\frac{1}{40}$	$\frac{1}{80}$
$t = 0.8$	$1.16 \cdot 10^{-3}$	$4.57 \cdot 10^{-4}$	$1.95 \cdot 10^{-4}$	$1.01 \cdot 10^{-4}$
$t = 4$	$7.64 \cdot 10^{-2}$	$3.96 \cdot 10^{-2}$	$2.0 \cdot 10^{-2}$	$1.04 \cdot 10^{-2}$
$t = 10$	$3.87 \cdot 10^{-1}$	$1.98 \cdot 10^{-1}$	$1.04 \cdot 10^{-1}$	$5.43 \cdot 10^{-2}$
$t = 20$	$1.04 \cdot 10^{0}$	$4.88 \cdot 10^{-1}$	$2.47 \cdot 10^{-1}$	$1.35 \cdot 10^{-1}$

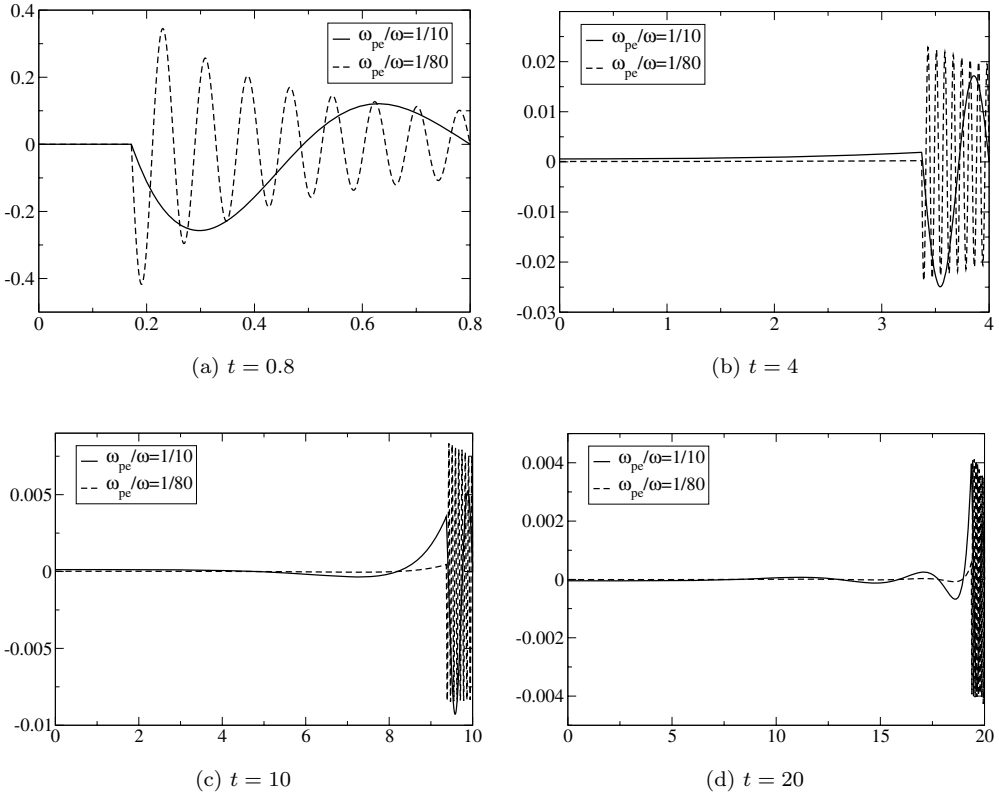
This is done using the Simpson rule on a very fine grid of the argument τ in order to guarantee that the level of the truncation error is far below the magnitude of either φ_1 or φ_2. To provide a most transparent interpretation of the numerical results, we also adopt a slightly different notion of the depth of a weak lacuna, namely, $\frac{\max|\varphi_{\mathrm{lacuna}}|}{\max|\varphi_{\mathrm{packet}}|}$, where $\varphi_{\mathrm{lacuna}} = \varphi_2$ and $\varphi_{\mathrm{packet}} = \varphi_1 + \varphi_2$. This new definition immediately provides a quantitative measure of how big the residual field inside the lacuna is compared to the total field inside the wave packet. For computations, we select $\omega_{\mathrm{pe}} = 1$, $T = 2\pi/10$, and in Table 3.1 present the depth of the weak lacuna for different values of $\omega_{\mathrm{pe}}/\omega$ and different moments of time t.

From Table 3.1, one can clearly see that for all moments of time—small, intermediate (not covered by the asymptotics), and large—the depth of the weak lacuna is indeed proportional to the quantity $\omega_{\mathrm{pe}}/\omega$. However, the maximum of the contaminating field φ_2 is taken in Table 3.1 across the entire lacuna $|\boldsymbol{x}| < c(t-T)$. Therefore, as expected, we do not observe any decay of the depth as the time increases; we rather observe the increase. In fact, this increase is due to the "tail" of the residual field that decays toward the center of the lacuna, as shown in Figure 3.1.

On the other hand, if we were to take a region narrower than the cone $|\boldsymbol{x}| < c(t-T)$ to evaluate the depth of the weak lacuna, then we would be able to actually see its decrease in time, as prescribed previously by the asymptotic estimates. In Table 3.2, we present the same quantity as in Table 3.1, except that $\max|\varphi_{\mathrm{lacuna}}| = \max|\varphi_2|$ is evaluated on a narrower cone $|\boldsymbol{x}| < c_1(t-T)$, where $c_1 = 0.75c$; see formula (3.35). The time range in Table 3.2 covers only intermediate to large intervals. From Table 3.2, one can clearly see not only that the depth of the weak lacuna is inversely proportional to $\omega_{\mathrm{pe}}/\omega$ for every particular moment of time, but that it also decays roughly as the inverse square root of time for every particular value of $\omega_{\mathrm{pe}}/\omega$; see formula (3.39).

An intermediate conclusion that we can draw, based on the combined use of asymptotic arguments and numerical quadratures, is that for high-frequency transverse electromagnetic waves that propagate in a dilute isotropic plasma (with particular pointwise excitation) one can still observe lacunae in the solutions but only in an approximate sense. The depth of these approximate, or weak, lacunae is proportional to the ratio of the Langmuir frequency of the plasma over the primary carrier frequency of the waves.

3.5. Numerical tests. In this section, we report on some results enabled by exploiting the weak lacunae in the computational context. As of yet, these results do not amount to a systematic numerical study. They rather provide a proof-of-concept illustration, whereas a broader and more coherent account of numerical simulations will be reported later.



(a) $t = 0.8$ (b) $t = 4$

(c) $t = 10$ (d) $t = 20$

FIG. 3.1. *Solution of the Klein–Gordon equation inside the lacuna and inside the wave packet.*

TABLE 3.2
The depth of a weak lacuna for $c_1 = 0.75c$ and different moments of time.

$\frac{\omega_{\rm pe}}{\omega}$	$\frac{1}{10}$	$\frac{1}{20}$	$\frac{1}{40}$	$\frac{1}{80}$
$t = 4$	$4.80 \cdot 10^{-2}$	$2.48 \cdot 10^{-2}$	$1.26 \cdot 10^{-2}$	$6.52 \cdot 10^{-3}$
$t = 10$	$3.64 \cdot 10^{-2}$	$1.81 \cdot 10^{-2}$	$9.47 \cdot 10^{-3}$	$4.93 \cdot 10^{-3}$
$t = 20$	$3.14 \cdot 10^{-2}$	$1.33 \cdot 10^{-2}$	$6.56 \cdot 10^{-3}$	$3.56 \cdot 10^{-3}$

We apply the lacunae-based algorithm of [24] to the Klein–Gordon equation (3.19). The algorithm of [24] was originally developed for the d'Alembert equation. It yields nonlocal ABCs that enable the computation of an unsteady wave field on a given finite region of interest. The rest of the space beyond this finite computational region is truncated, and the ABCs provide the required closure at the external artificial boundary so that the outgoing waves can propagate without any nonphysical reflections. Our objective hereafter is to demonstrate that the weak lacunae of section 3.4 can sometimes substitute for the actual lacunae in the numerical framework.

Lacunae-based ABCs for the genuine diffusionless case are constructed in two stages. Below we provide only a very brief description of the method and refer the reader to [24, 25] for details. A key initial assumption is that the overall infinite-domain problem has a unique solution and that (at least) outside of the aforementioned

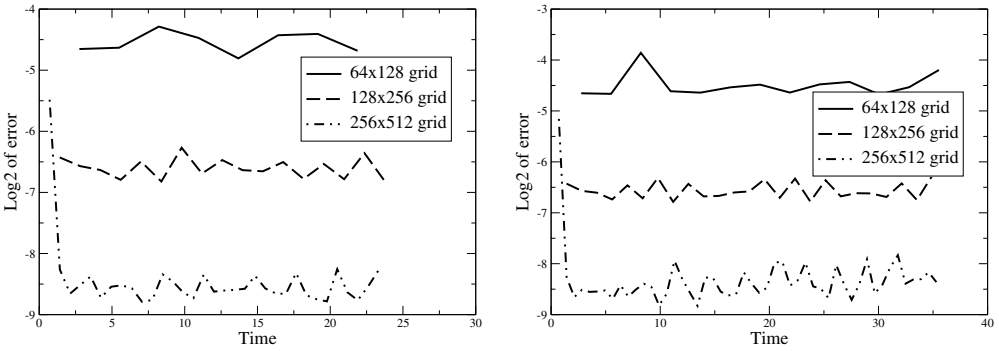
finite region of interest this solution is governed by a linear homogeneous equation, such as the d'Alembert equation. At the first stage, the original problem is decomposed into two subproblems that depend on one another. The interior subproblem is formulated on the bounded computational domain. It inherits all the structure and properties of the original problem on this domain. As the computational domain is obtained by truncation, the interior subproblem obviously requires a closure, i.e., the ABCs, at the outer boundary. The ABCs are to be provided by the solution of the exterior subproblem. The latter, in turn, is formulated on the entire space and is driven by the special auxiliary sources that depend on the solution of the interior problem. The governing equation for the the exterior subproblem on the entire space is the same linear homogeneous equation that governs the solution of the original problem outside the region of interest.

At the second stage, the two problems are integrated concurrently. In doing so, the algorithm for integrating the exterior problem is built around the presence of lacunae. The continuously operating auxiliary sources are partitioned in time into finite fragments. The solution due to each fragment has a lacuna, and the entire domain of interest falls inside this lacuna after a predetermined interval of time. Once this happens, the computation for this particular fragment does not need to be continued any further. Moreover, no wave can travel more than a certain fixed distance away from the source during this interval of time, which implies that the computations can always be conducted on a bounded auxiliary domain of a fixed nonincreasing size. This is the mechanism of transition from an infinite-domain formulation to a finite-domain one. Altogether, one can show that at any given moment of time only a finite fixed number of fragments contribute to the solution of the exterior problem, and each contribution needs to be computed only over a fixed time interval. This yields the exact unsteady ABCs with only fixed and limited extent of nonlocality in time. The performance of these ABCs does not deteriorate when integrating over long time intervals [24].

Replacing genuine lacunae by weak lacunae in the framework of the ABC algorithm basically means that the interior problem is still integrated in its entirety, whereas the dispersive effects for the exterior problem, i.e., for the boundary conditions, are artificially "cut short." Indeed, for each element of the source partition the solution to the exterior problem is computed only until the region of interest falls inside the lacuna. The effect of the corresponding mismatch on the overall numerical performance will be throughly studied in the future. In the meantime, we simply provide some computational examples.

We are solving a model problem of radiation of waves by a known source. The exact solution for this problem is available; it is obtained by reverse engineering, i.e., by picking a function, substituting it under the differential operator, and deriving the right-hand side. For actual computations, we choose the Yee scheme [36], which is a well-known staggered central-difference scheme that has second order accuracy. We also set $\omega_{\mathrm{pe}}/\omega = 1/100$ and select other parameters (grids, geometry, etc.) as in [24]. Namely, the computations are conducted on a uniform grid in the cylindrical coordinates. The methodology does not require that the grid be fitted to the shape of the domain of interest, and we choose the latter spherical. Note that the parameter c_1 (see formulae (3.34) and (3.35)), is not specified explicitly as input for the computational procedure, but other parameters are specified so as to effectively make it $c_1/c \approx 0.9$.

In Figure 3.2, we present the results of the grid convergence study (binary logarithm of the maximum norm of the error as a function of time) for two different values of the diameter of the sphere. The grid dimensions shown in Figure 3.2 pertain to the



(a) Diameter 1.2 (b) Diameter 1.8

FIG. 3.2. *Numerical performance of the ABCs based on weak lacunae.*

auxiliary domain of cylindrical coordinates that has radius π and length 2π. The propagation speed is taken equal to one, and the computations are conducted over the time interval equivalent to 20 times the time required for the waves to travel across the sphere. At least for the particular setup selected, the plots in Figure 3.2 experimentally corroborate the design convergence rate of the scheme (second order) equipped with the ABCs based on the weak lacunae.

3.6. Anisotropic case. As has been mentioned, the primary source of anisotropy in the ionospheric plasma is the magnetic field of the Earth. It may play an important role for the propagation of electromagnetic waves. In particular, it may affect the structure and depth of the weak lacunae. In this section, we outline an approach to analyzing the weak lacunae in the presence of a constant external magnetic field.

Let $\boldsymbol{B}_0 = \text{const}$ be the magnetic field of the Earth. Then, the Lorentz term is to be kept on the right-hand side of (3.2), and instead of (3.3) we obtain

$$m_{\text{e}}\frac{d\boldsymbol{u}}{dt} = -e\boldsymbol{E} - \frac{e}{c}\boldsymbol{u}\times\boldsymbol{B}_0. \tag{3.43}$$

We now need to find the first time derivative of the induced current that provides the excitation for the electric field on the right-hand side of the governing equation (3.1). Substituting $\boldsymbol{j}_{\text{ind}} = -en_{\text{e}}\boldsymbol{u}$ into (3.43), we obtain

$$\boldsymbol{j}'_{\text{ind}} = \frac{\omega_{\text{pe}}^2}{4\pi}\boldsymbol{E} - \Omega_{\text{e}}\boldsymbol{j}_{\text{ind}}\times\frac{\boldsymbol{B}_0}{|\boldsymbol{B}_0|}. \tag{3.44}$$

Equation (3.44) is a first order ordinary differential equation with respect to the unknown current $\boldsymbol{j}_{\text{ind}}$, which is a function of time. It needs to be solved along with (3.1). It is clear that in doing so the dependence of $\boldsymbol{j}'_{\text{ind}}$ on $\boldsymbol{E}$ will be given by a convolution, which means that the responses of the anisotropic medium (3.43) will, generally speaking, be nonlocal in time. Later in the section we will see, however, that under certain assumptions the effect of anisotropy can still be regarded as small.

We begin with providing an elementary frequency-domain analysis. The use of the variable $\boldsymbol{P}$ (polarization), where $\boldsymbol{j}_{\text{ind}} = \frac{\partial\boldsymbol{P}}{\partial t}$, will be more convenient on some occasions, because it has the same dimension as the field $\boldsymbol{E}$. In the frequency domain,

(3.44) can be transformed into

$$\omega^2 \boldsymbol{P}(\omega) + i\omega\Omega_{\mathrm{e}} \boldsymbol{P}(\omega) \times \frac{\boldsymbol{B}_0}{|\boldsymbol{B}_0|} = -\frac{\omega_{\mathrm{pe}}^2}{4\pi} \boldsymbol{E}(\omega). \tag{3.45}$$

Assuming with no loss of generality that the magnetic field $\boldsymbol{B}_0$ is aligned with the Cartesian coordinate z, we solve (3.45) with respect to $\boldsymbol{P}(\omega)$ and obtain

$$\begin{aligned} \boldsymbol{P}(\omega) = -\frac{\omega_{\mathrm{pe}}^2}{4\pi\omega^2} \boldsymbol{E}(\omega) + \frac{\omega_{\mathrm{pe}}^2}{4\pi\omega^2} \frac{i\omega\Omega_{\mathrm{e}}}{\omega^2 - \Omega_{\mathrm{e}}^2} \boldsymbol{E}(\omega) \times \frac{\boldsymbol{B}_0}{|\boldsymbol{B}_0|} \\ - \frac{\omega_{\mathrm{pe}}^2}{4\pi\omega^2} \frac{\Omega_{\mathrm{e}}^2}{\omega^2 - \Omega_{\mathrm{e}}^2} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{E}(\omega). \end{aligned} \tag{3.46}$$

Note that the first term on the right-hand side of (3.46) is exactly the same as we obtained in the isotropic case; see formula (3.5). The second and third terms on the right-hand side of (3.46) are due to the presence of the magnetic field $\boldsymbol{B}_0$. These terms, which are proportional to the first and second power of the cyclotron frequency Ω_{e}, respectively, are known to be responsible for the effects of gyrotropy and Faraday rotation that accompany the propagation of electromagnetic waves in the anisotropic plasma; see [18, 21].

The case of particular interest for us is that of the high-frequency propagation. If $\omega \gg \omega_{\mathrm{pe}}$, then also $\omega \gg \Omega_{\mathrm{e}}$, because according to section 3.1, Ω_{e} is about an order of magnitude lower than ω_{pe} for the typical range of parameters that characterize the ionospheric plasma. Consequently, instead of (3.46) we can write

$$\boldsymbol{P}(\omega) \approx -\frac{\omega_{pe}^2}{4\pi\omega^2} \boldsymbol{E}(\omega) + \frac{\omega_{\mathrm{pe}}^2}{4\pi\omega^2} \frac{i\Omega_{\mathrm{e}}}{\omega} \boldsymbol{E}(\omega) \times \frac{\boldsymbol{B}_0}{|\boldsymbol{B}_0|}. \tag{3.47}$$

Note that $\boldsymbol{B}_0/|\boldsymbol{B}_0|$ on the right-hand side of (3.47) is a dimensionless unit vector in the direction of the magnetic field $\boldsymbol{B}_0$. Then, by comparing the two terms on the right-hand side of (3.47) and by recalling that the effect of the first term on lacunae back in the time domain is $\mathcal{O}(\frac{\omega_{\mathrm{pe}}}{\omega})$ (see estimates (3.31) and (3.39)), we can qualitatively conjecture that the additional effect of anisotropy on lacunae is likely to be $\mathcal{O}(\frac{\omega_{\mathrm{pe}}}{\omega} \cdot \sqrt{\frac{\Omega_{\mathrm{e}}}{\omega}})$. It is expected to be much smaller than the $\mathcal{O}(\frac{\omega_{\mathrm{pe}}}{\omega})$ attributed to the "primary" dispersion, because the extra factor contained in the second term on the right-hand side of (3.47) is $\Omega_{\mathrm{e}}/\omega \ll 1$.

To conduct the analysis in the time domain, we employ the Laplace transform instead of the Fourier transform and, assuming homogeneous initial conditions for the polarization, obtain (cf. formula (3.45))

$$s^2 \boldsymbol{P}(s) + s\Omega_{\mathrm{e}} \boldsymbol{P}(s) \times \frac{\boldsymbol{B}_0}{|\boldsymbol{B}_0|} = \frac{\omega_{\mathrm{pe}}^2}{4\pi} \boldsymbol{E}(s). \tag{3.48}$$

The primary quantity of interest for us is $s^2\boldsymbol{P}$, because $\boldsymbol{j}'_{\mathrm{ind}} = \boldsymbol{P}''$, and we find

$$s^2 \boldsymbol{P}(s) = \frac{\omega_{\mathrm{pe}}^2}{4\pi} \boldsymbol{E}(s) + \frac{\omega_{\mathrm{pe}}^2}{4\pi} \begin{bmatrix} -\frac{\Omega_{\mathrm{e}}^2}{s^2+\Omega_{\mathrm{e}}^2} & -\frac{\Omega_{\mathrm{e}} s}{s^2+\Omega_{\mathrm{e}}^2} & 0 \\ \frac{\Omega_{\mathrm{e}} s}{s^2+\Omega_{\mathrm{e}}^2} & -\frac{\Omega_{\mathrm{e}}^2}{s^2+\Omega_{\mathrm{e}}^2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{E}(s).$$

Consequently,

$$(j'_{\rm ind})_x = P''_x = \frac{\omega_{\rm pe}^2}{4\pi}E_x + \frac{\omega_{\rm pe}^2\Omega_{\rm e}}{4\pi}\left[-\sin(\Omega_{\rm e}t) * E_x(t) - \cos(\Omega_{\rm e}t) * E_y(t)\right],$$

$$(j'_{\rm ind})_y = P''_y = \frac{\omega_{\rm pe}^2}{4\pi}E_y + \frac{\omega_{\rm pe}^2\Omega_{\rm e}}{4\pi}\left[\cos(\Omega_{\rm e}t) * E_x(t) - \sin(\Omega_{\rm e}t) * E_y(t)\right],$$

$$(j'_{\rm ind})_z = P''_z = \frac{\omega_{\rm pe}^2}{4\pi}E_z.$$

From the previous expressions we see that electromagnetic responses of the anisotropic plasma involve off-diagonal terms, i.e., relate different components of the field and current vectors (as opposed to only respective components). Therefore, we will employ diagonalization by means of the transformation $\mathcal{T}$:

$$\mathcal{T} = \begin{bmatrix} i & -i & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathcal{T}^{-1} = \begin{bmatrix} -i/2 & 1/2 & 0 \\ i/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Let $\boldsymbol{E} = \mathcal{T}\boldsymbol{G}$ and $\boldsymbol{P} = \mathcal{T}\boldsymbol{Q}$. Then, (3.48) transforms into

$$s^2\boldsymbol{Q}(s) = \frac{\omega_{\rm pe}^2}{4\pi}\boldsymbol{G}(s) + \frac{\omega_{\rm pe}^2}{4\pi}\frac{\Omega_{\rm e}}{s^2+\Omega_{\rm e}^2}\begin{bmatrix} -\Omega_{\rm e}+is & 0 & 0 \\ 0 & -\Omega_{\rm e}-is & 0 \\ 0 & 0 & 0 \end{bmatrix}\boldsymbol{G}(s). \tag{3.49}$$

If we also define $\boldsymbol{j}_{\rm ind} = \mathcal{T}\boldsymbol{q}$, then $s\boldsymbol{q}(s) = s^2\boldsymbol{Q}(s)$, and from (3.49) we find

$$\begin{aligned} q'_x(t) &= \frac{\omega_{\rm pe}^2}{4\pi}G_x(t) + i\frac{\omega_{\rm pe}^2\Omega_{\rm e}}{4\pi}\left[e^{i\Omega_{\rm e}t} * G_x(t)\right], \\ q'_y(t) &= \frac{\omega_{\rm pe}^2}{4\pi}G_y(t) - i\frac{\omega_{\rm pe}^2\Omega_{\rm e}}{4\pi}\left[e^{-i\Omega_{\rm e}t} * G_y(t)\right], \\ q'_z(t) &= \frac{\omega_{\rm pe}^2}{4\pi}G_z(t). \end{aligned} \tag{3.50}$$

To quantify the effect of anisotropy, we will need to analyze the convolutions on the right-hand side of the first two equations (3.50):

$$e^{\pm i\Omega_{\rm e}t} * G_{x,\,y}(t) = \int_0^t e^{\pm i\Omega_{\rm e}(t-v)}G_{x,\,y}(v)dv = e^{\pm i\Omega_{\rm e}t}\int_0^t e^{\mp i\Omega_{\rm e}v}G_{x,\,y}(v)dv.$$

Consider, for example, the component G_x and introduce the following ansatz: $G_x(v) = e^{i\omega t}\tilde{G}_x(v)$, where $\tilde{G}_x(v)$ is assumed to be a slowly varying function. Then, we integrate by parts twice and obtain

$$\begin{aligned} \int_0^t e^{-i\Omega_{\rm e}v}G_x(v)dv &= \frac{1}{i(\omega-\Omega_{\rm e})}\left[G_x(t)e^{-i\Omega_{\rm e}t} - \int_0^t e^{i(\omega-\Omega_{\rm e})v}\tilde{G}'_x(v)dv\right] \\ &= \frac{1}{i(\omega-\Omega_{\rm e})}G_x(t)e^{-i\Omega_{\rm e}t} + \frac{\tilde{G}'_x(t)e^{i(\omega-\Omega_{\rm e})t} - \tilde{G}'_x(0)}{(\omega-\Omega_{\rm e})^2} \\ &\quad - \frac{1}{(\omega-\Omega_{\rm e})^2}\int_0^t e^{i(\omega-\Omega_{\rm e})v}\tilde{G}''_x(v)dv. \end{aligned}$$

Consequently,

$$\begin{aligned}\frac{\partial q_x}{\partial t} &= \frac{\omega_{\rm pe}^2}{4\pi} G_x(t) + \frac{\omega_{\rm pe}^2}{4\pi}\,\frac{\Omega_{\rm e}}{\omega-\Omega_{\rm e}} G_x(t)\\ &+ e^{i\Omega_{\rm e}t}\,\frac{\omega_{\rm pe}^2}{4\pi}\,\frac{\Omega_{\rm e}}{\omega-\Omega_{\rm e}}\,\frac{\tilde G_x'(t)e^{i(\omega-\Omega_{\rm e})t} - \tilde G_x'(0)}{\omega-\Omega_{\rm e}}\\ &- e^{i\Omega_{\rm e}t}\,\frac{\omega_{\rm pe}^2}{4\pi}\,\frac{\Omega_{\rm e}}{\omega-\Omega_{\rm e}}\,\frac{1}{\omega-\Omega_{\rm e}}\int_0^t e^{i(\omega-\Omega_{\rm e})v}\tilde G_x''(v)dv.\end{aligned} \tag{3.51}$$

Slow variation of $\tilde G_x(v)$ means that it is slow on the scale of the high-frequency oscillation ω, and in many cases this slowness is a natural assumption about the field. Under this assumption, the third and fourth terms on the right-hand side of equality (3.51) can be neglected. Indeed, the third term is small compared to the second one because

$$\frac{\max|G_x'|}{\omega-\Omega_{\rm e}} \ll \max|G_x|.$$

As for the fourth term on the right-hand side of (3.51), using the Riemann–Lebesgue lemma we can write

$$\int_0^t e^{i(\omega-\Omega_{\rm e})v}\tilde G_x''(v)dv = o\left(\max|G_x''|\right) \quad \text{as} \quad \omega\to\infty.$$

Therefore, for high carrier frequencies it is also small compared to the second term. Consequently,

$$\frac{\partial q_x}{\partial t} \approx \frac{\omega_{\rm pe}^2}{4\pi}\left(1+\frac{\Omega_{\rm e}}{\omega-\Omega_{\rm e}}\right) G_x(t), \tag{3.52}$$

and a similar expression can be obtained for q_y'. Hence, when the field is represented as the product of a rapidly oscillating carrier times a slowly varying envelope, the nonlocal responses due to the anisotropy can be approximated by local expressions of the type (3.52).

Finally, let us revisit the governing equation for the field (3.1). We note that when plasma becomes anisotropic, the notion of longitudinal and transverse waves often changes its meaning, and in the literature one would typically consider the waves that propagate along the magnetic field and those that propagate perpendicular to the magnetic field; see, e.g., [8]. Of course, other propagation angles are also possible, and, in general, the split into the longitudinal and transverse components is not always straightforward. We will, however, still consider the transverse field $\boldsymbol{E}_\perp$ in the previous sense of the word, i.e., the one that satisfies $\operatorname{div}\boldsymbol{E}_\perp = 0$. Let also $\boldsymbol{E}_\perp = \mathcal{T}\boldsymbol{G}$; then from (3.1) we obtain

$$\frac{\partial \mathcal{T}\boldsymbol{G}}{\partial t} - c^2\Delta(\mathcal{T}\boldsymbol{G}) + 4\pi\frac{\partial \mathcal{T}\boldsymbol{q}}{\partial t} = \boldsymbol{0}.$$

Since $\mathcal{T}$ is a constant matrix, and the vector Laplacian in the Cartesian coordinates applies independently to individual components, we can use formulae (3.50), (3.52) and write

$$\frac{\partial \boldsymbol{G}}{\partial t} - c^2\Delta\boldsymbol{G} + \omega_{\rm pe}^2\begin{bmatrix} 1+\frac{\Omega_{\rm e}}{\omega-\Omega_{\rm e}} & 0 & 0\\ 0 & 1+\frac{\Omega_{\rm e}}{\omega-\Omega_{\rm e}} & 0\\ 0 & 0 & 1\end{bmatrix}\boldsymbol{G} = \boldsymbol{0}.$$

This is a vector equation, which is equivalent to three scalar Klein–Gordon equations. For the first two components the dispersive term is $\sim \omega_{\rm pe}^2(1+\frac{\Omega_{\rm e}}{\omega-\Omega_{\rm e}})$ as opposed to simply $\sim \omega_{\rm pe}^2$, which was the case in section 3.4. We therefore conclude that the additional effect of anisotropy on weak lacunae of electromagnetic waves in the dilute ionospheric plasma can be approximately measured as $\mathcal{O}(\frac{\omega_{\rm pe}}{\omega}\sqrt{\frac{\Omega_{\rm e}}{\omega-\Omega_{\rm e}}})$.

4. Discussion. Classical lacunae can be observed in the solutions of the Maxwell equations only when the electromagnetic waves propagate in vacuum or in dielectric media with static response. Otherwise, the propagation is accompanied by aftereffects, and there are no sharp aft fronts and no lacunae in the solutions. For low incident frequencies, the mechanism that destroys the lacunae can largely be attributed to dissipation due to the Ohm conductivity. For high incident frequencies, when the material coefficients can no longer be considered constant, the diffusion of waves is basically caused by the physical dispersion. However, for the propagation of transverse electromagnetic waves in dilute plasma, when the incident frequency is much higher than the Langmuir frequency, lacunae can still be identified in the corresponding solutions of the Maxwell equations, although in an approximate sense. The depth of these weak lacunae, i.e., the magnitude of the residual field relative to the magnitude of the field in the primary wave packet, is proportional to the ratio of the Langmuir frequency over the primary carrying frequency of the waves. In the anisotropic case, when the plasma is immersed into the external magnetic field, there is an additional small factor, approximately equal to the square root of the ratio of the cyclotron frequency over the carrier frequency, that affects the depth of the weak lacunae.

An interesting subject for future study could be analysis of the case when anisotropic responses should remain nonlocal in time, as well as a more careful analysis of the conductivity mechanisms in the ionosphere. On the numerical side, the future direction is the ABC algorithm based on the weak lacunae.

REFERENCES

[1] M. F. Atiyah, R. Bott, and L. Gårding, *Lacunas for hyperbolic differential operators with constant coefficients*. I, Acta Math., 124 (1970), pp. 109–189.

[2] M. F. Atiyah, R. Bott, and L. Gårding, *Lacunas for hyperbolic differential operators with constant coefficients*. II, Acta Math., 131 (1973), pp. 145–206.

[3] M. Belger, R. Schimming, and V. Wünsch, *A survey on Huygens' principle*, Z. Anal. Anwendungen, 16 (1997), pp. 9–36.

[4] N. N. Bogoliubov and D. V. Shirkov, *Introduction to the Theory of Quantized Fields*, Interscience Monographs in Physics and Astronomy, Vol. III, Interscience, New York, London, 1959.

[5] M. Born and E. Wolf, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, 7th ed., Cambridge University Press, Cambridge, UK, 1999.

[6] K. G. Budden, *The Propagation of Radio Waves: The Theory of Radio Waves of Low Power in the Ionosphere and Magnetosphere*, Cambridge University Press, Cambridge, UK, 1985.

[7] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. II, Wiley, New York, 1962.

[8] V. L. Ginzburg, *The Propagation of Electromagnetic Waves in Plasmas*, Internat. Ser. Monogr. Electromagnetic Waves 7, Pergamon Press, Oxford, UK, 1964.

[9] P. Günther, *Ein Beispiel einer nichttrivialen Huygensschen Differentialgleichung mit vier unabhängigen Variablen*, Arch. Rational Mech. Anal., 18 (1965), pp. 103–106.

[10] P. Günther, *Huygens' Principle and Hyperbolic Equations*, Perspect. Math. 5, Academic Press, Boston, 1988.

[11] P. Günther, *Huygens' principle and Hadamard's conjecture*, Math. Intelligencer, 13 (1991), pp. 56–63.

[12] J. Hadamard, *Lectures on Cauchy's Problem in Linear Partial Differential Equations*, Yale University Press, New Haven, CT, 1923.

[13] J. Hadamard, *Problème de Cauchy*, Hermann, Paris, 1932.
[14] J. Hadamard, *The problem of diffusion of waves*, Ann. of Math. (2), 43 (1942), pp. 510–522.
[15] B. B. Kadomtsev, *Collective Phenomena in Plasma* [*Kollektivnye yavleniya v plazme*], 2nd ed., Nauka, Moscow, 1988 (in Russian).
[16] J. E. Lagnese and K. L. Stellmacher, *A method of generating classes of Huygens' operators*, J. Math. Mech., 17 (1967), pp. 461–472.
[17] L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics, Vol.* 2*, The Classical Theory of Fields*, 4th ed., Pergamon Press, Oxford, UK, 1975.
[18] L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics, Vol.* 8*, Electrodynamics of Continuous Media*, Pergamon International Library of Science, Technology, Engineering and Social Studies, Pergamon Press, Oxford, 1984.
[19] P. D. Lax and R. S. Phillips, *An example of Huygens' principle*, Comm. Pure Appl. Math., 31 (1978), pp. 415–421.
[20] M. Matthisson, *Le probléme de Hadamard relatif á la diffusion des ondes*, Acta Math., 71 (1939), pp. 249–282.
[21] D. B. Melrose and R. C. McPhedran, *Electromagnetic Processes in Dispersive Media. A Treatment Based on the Dielectric Tensor*, Cambridge University Press, Cambridge, UK, 1991.
[22] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics* (2 volumes), Internat. Ser. Pure Appl. Phys., McGraw–Hill, New York, 1953.
[23] I. Petrowsky, *On the diffusion of waves and the lacunas for hyperbolic equations*, Rec. Math. [Mat. Sbornik] N.S. 17 (1945), pp. 289–370.
[24] V. S. Ryaben'kii, S. V. Tsynkov, and V. I. Turchaninov, *Global discrete artificial boundary conditions for time-dependent wave propagation*, J. Comput. Phys., 174 (2001), pp. 712–758.
[25] V. S. Ryaben'kii, S. V. Tsynkov, and V. I. Turchaninov, *Long-time numerical computation of wave-type solutions driven by moving sources*, Appl. Numer. Math., 38 (2001), pp. 187–222.
[26] R. Schimming, *A review of Huygens' principle for linear hyperbolic differential equations*, in Proceedings of the IMU Symposium Group-Theoretical Methods in Mechanics, Novosibirsk, USSR, 1978, USSR Academy of Science, Siberian Branch, Novosibirsk, Russia, 1978.
[27] L. I. Sedov, *Mechanics of Continuous Media*, Vols. 1 and 2, Ser. Theoret. Appl. Mech. 4, World Scientific, River Edge, NJ, 1997.
[28] K. L. Stellmacher, *Ein Beispiel einer Huyghensschen Differentialgleichung*, Nachr. Akad. Wiss. Göttingen Math.-Phys. Kl., Math.-Phys.-Chem. Abt., 1953 (1953), pp. 133–138.
[29] K. L. Stellmacher, *Eine Klasse huyghenscher Differentialgleichungen und ihre Integration*, Math. Ann., 130 (1955), pp. 219–233.
[30] S. V. Tsynkov, *Numerical solution of problems on unbounded domains. A review*, Appl. Numer. Math., 27 (1998), pp. 465–532.
[31] S. V. Tsynkov, *Artificial boundary conditions for the numerical simulation of unsteady acoustic waves*, J. Comput. Phys., 189 (2003), pp. 626–650.
[32] S. V. Tsynkov, *Artificial Boundary Conditions for the Numerical Simulation of Unsteady Electromagnetic Waves*, Tech. report CRSC–TR03–19, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, 2003.
[33] S. V. Tsynkov, *Lacunae-based artificial boundary conditions for the numerical simulation of unsteady waves governed by vector models*, in Mathematical and Numerical Aspects of Wave Propagation—WAVES 2003 (Jyväskylä, Finland, 2003), G. C. Cohen, E. Heikkola, P. Joly, and P. Neittaanmäki, eds., Springer, Berlin, 2003, pp. 103–108.
[34] S. V. Tsynkov, *On the application of lacunae-based methods to Maxwell's equations*, J. Comput. Phys., 199 (2004), pp. 126–149.
[35] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis. An Introduction to the General Theory of Infinite Processes and of Analytic Functions: With an Account of the Principal Transcendental Functions*, 4th ed., Cambridge University Press, New York, 1962.
[36] K. S. Yee, *Numerical solution of initial boundary value problem involving Maxwell's equations in isotropic media*, IEEE Trans. Antennas and Propagation, 14 (1966), pp. 302–307.

SIAM J. APPL. MATH.
Vol. 67, No. 6, pp. 1582–1609

ACTIVE CONTROL OF SOUND FOR COMPOSITE REGIONS*

A. W. PETERSON† AND S. V. TSYNKOV‡

Abstract. We present a methodology for the active control of time-harmonic wave fields, e.g., acoustic disturbances, in composite regions. This methodology extends our previous approach developed for the case of arcwise connected regions. The overall objective is to eliminate the effect of all outside field sources on a given domain of interest, i.e., to shield this domain. In this context, active shielding means introducing additional field sources, called active controls, that generate the annihilating signal and cancel out the unwanted component of the field. As such, the problem of active shielding can be interpreted as a special inverse source problem for the governing differential equation or system. For a composite domain, not only do the controls prevent interference from all exterior sources, but they can also enforce a predetermined communication pattern between the individual subdomains (as many as desired). In other words, they either allow the subdomains to communicate freely with one another or otherwise have them shielded from their peers. In the paper, we obtain a general solution for the composite active shielding problem and show that it reduces to solving a collection of auxiliary problems for arcwise connected domains. The general solution is constructed in two stages. Namely, if a particular subdomain is not allowed to hear another subdomain, then the supplementary controls are employed first. They communicate the required data prior to building the final set of controls. The general solution can be obtained with only the knowledge of the acoustic signals propagating through the boundaries of the subdomains. No knowledge of the field sources is required, nor is any knowledge of the properties of the medium needed.

Key words. active shielding, noise control, inverse source problem, time-harmonic acoustic fields, composite domain, communication pattern, the Helmholtz equation, generalized Calderon's potentials, exact volumetric cancellation, general solution, incoming and outgoing waves, wave split

AMS subject classifications. 35J05, 35C15, 31C99, 47G30, 35B37

DOI. 10.1137/060662368

1. Introduction.

Active shielding and control of noise is a very rich field with a variety of applications. In the most general terms, exercising active control means introducing additional sources of sound, called controls, to facilitate a specific change in the overall acoustic field. In particular, the desired change may imply canceling all or part of the field on a given region. Referring the reader to other, more detailed sources for a comprehensive review (see [17, 8, 24]), we mention several representative publications in the area. Research by Elliott, Stothers, and Nelson [7] focused on the minimization of noise at pointwise locations. Wright and Vuksanovic expanded the field to include directional noise cancellation in [30, 31]. A large portion of the research done today has been motivated by the airline industry and its desire to control unwanted engine noise in the cabin during flight. There are various methods of dealing with in-flight noise. Damping structural vibrations is one approach to attenuating low frequencies. This is done by placing actuators and sensors throughout the cabin at optimized locations. Kincaid, Laba, and Padula worked extensively on this problem [12, 11]. A comprehensive account of the area, along with many additional references, can be found in [4]. Another method involves placing a series of microphones and

*Received by the editors June 7, 2006; accepted for publication (in revised form) April 26, 2007; published electronically September 12, 2007.
http://www.siam.org/journals/siap/67-6/66236.html
†Mathematics Department, North Carolina State University, Box 8205, Raleigh, NC 27695 (awpeters@unity.ncsu.edu).
‡Corresponding author. Mathematics Department, North Carolina State University, Box 8205, Raleigh, NC 27695 (tsynkov@math.ncsu.edu, http://www.math.ncsu.edu/∼stsynkov).

speakers throughout the cabin and uses acoustic excitation to cancel unwanted noise. Passive techniques such as sound insulation are more effective in dealing with high frequencies. Van der Auweraer et al. tackled the problem of aircraft noise in [26] by using a combination of both methods.

In [9], Fuller and von Flotow present an overview of current common practices in active noise control. One of the most popular algorithms used today in the control of noise is based on a least mean squares (LMS) method. It is employed to tune the control filter to reduce unwanted noise near the sensors and was first introduced by Burgess in [1] and by Widrow, Shur, and Shaffer in [29]. This algorithm was later improved upon by Cabell and Fuller in [2]. While LMS methods offer good results near the sensors in small-scale applications such as mobile phones, they do not allow for the exact volumetric cancellation of noise desired in an airline cabin.

In the current paper, we introduce and study a new formulation of active noise control problem. Namely, the overall region of space to be protected from noise is assumed to be composed of a number of simple, i.e., arcwise connected, (sub)domains. The standard part of the formulation involves shielding the overall domain, i.e., the union of all subdomains, from the unwanted noise. In addition, the individual subdomains are selectively allowed to either communicate freely with one another according to a predetermined pattern or else be shielded from their peers. In doing so, no reciprocity is assumed; i.e., for a given pair of subdomains one may be allowed to hear the other, but not vice versa.

The method of analysis used in this paper builds upon the previous research done by Lončarić, Ryaben'kii, and Tsynkov in [13] and by Tsynkov in [25] for the case of a single arcwise connected domain, and subsequently extended in [14, 15, 16] by investigating various optimization formulations. The approach of [13] allows for the exact volumetric cancellation of time-harmonic noise in a given region. In other words, this region is shielded from the unwanted sound that comes from the outside. The shielding is achieved by first splitting the total acoustic field into the incoming and outgoing components. This can be done unambiguously using only the knowledge of the field and its normal derivative measured at the boundary of the region to be protected. Subsequently, the unwanted incoming component of the field is canceled by additional sources that are insensitive to the outgoing component. Other methods, such as those employed by Nelson and Elliott in [17], require that the noise to be canceled be measured at the boundary by itself, and be distinguished from other components of the acoustic field ahead of time. This restriction does not exist in the methodology presented herein. Moreover, our methodology requires knowledge of neither the volumetric properties of the medium nor the location and strength of the noise sources. Decomposition of the overall sound field into incoming and outgoing components, as well as design of the antinoise sources, are accomplished by applying Calderon's potentials and projections [3]; see also [23]. This is a very convenient and powerful apparatus that allows one to describe all appropriate control sources in closed form. In the simplest case of constant coefficients, the Calderon operators can be obtained using boundary integrals of classical potential theory.

There are two types of control sources that can be explored, volumetric and surface. In [13], it is determined that the general solution for volumetric controls $g = g(x)$ is given by

$$g(x) = -Lw$$

outside of the region to be shielded, where w is a special auxiliary function which must satisfy the Sommerfeld radiation condition at infinity, as well as coincide with

the acoustic field u and its normal derivative $\frac{\partial u}{\partial n}$ at the boundary. Here $L = \Delta + k^2 I$ denotes the Helmholtz operator. Since these are fairly loose restrictions, volumetric controls define a very broad class of solutions to the problem.

Surface controls are concentrated at the boundary. They are given by

$$g^{(surf)} = -\left[\frac{\partial w}{\partial n} - \frac{\partial u}{\partial n}\right]_\Gamma \delta(\Gamma) - \frac{\partial}{\partial n}([w-u]_\Gamma \delta(\Gamma)),$$

where the auxiliary function w is additionally required to satisfy the homogeneous Helmholtz equation, $Lw = 0$, outside the boundary but is no longer required to satisfy any boundary conditions at Γ. The general solution to the surface control problem is discussed in [25]. It is to be noted that surface controls have the same fundamental properties as volumetric controls. A universal framework for both volumetric and surface controls is built by Ryaben'kii and Utyuzhnikov in the recent paper [22]; it treats the governing equation for the field in an operator form.

We should also emphasize that the continuous formulation is not practical for implementation. Any realistic implementation would consist only of a finite number of sensors (microphones) and actuators (speakers). This will lead to a discretization of the problem on a grid. Discrete active shielding problems were analyzed, and the corresponding solutions obtained in [14, 15, 16, 25], as well as more recently in [22]. The finite-difference analysis of [14, 15, 16, 22, 25] uses the constructs developed previously in the works by Ryaben'kii [18] and by Veizman and Ryaben'kii [27, 28].

Specific objective of the current paper. We will extend the methodology of [13] to the case of composite regions. This will allow two or more separate subregions to be fully protected from the influence of outside sources. Moreover, according to a predetermined communication pattern, each individual subregion may or may not be allowed to hear any other subregion. As in [13], only the total acoustic field and its normal derivative specified at the boundaries will be needed for the exact volumetric cancellation of the outside noise, as well as for the realization of a given communication pattern. It will not be necessary to distinguish the "adverse," i.e., unwanted, part of the acoustic field from its "friendly," i.e., wanted, part as this is done automatically by the control system. The methodology will provide a closed form general solution for the controls, including the case of an inhomogeneous medium.

2. Two regions. In this section, the formulation for two separate domains will be discussed. In other words, we will distinguish between the two given disjoint regions and the rest of the space. Let it be noted, however, that the forthcoming methodology could be presented in a more general framework. The rest of the space outside of the two given regions can be treated as a third region on equal terms with the first two. This formulation lends itself more naturally to surface controls separating the three regions. A rigorous analysis of this approach for the finite-difference setting can be found in the recent paper [21], which, in turn, builds upon [18]. We, however, choose a simpler form of presentation in order to make it more accessible for applications. Accordingly, the focus will be on volumetric controls, which will allow for more flexibility in their construction.

2.1. Formulation. Let Ω_1 and Ω_2 be given, where $\Omega_i \subseteq \mathbb{R}^2$ or $\mathbb{R}^3$ is either bounded or unbounded. For simplicity we will first assume that Ω_1 and Ω_2 are two separate bounded regions of $\mathbb{R}^n$ (see Figure 2.1), and such that $\operatorname{dist}(\Omega_1, \Omega_2) \geq \epsilon > 0$. Let Γ_1 and Γ_2 be the boundaries of Ω_1 and Ω_2, respectively. Consider the time-

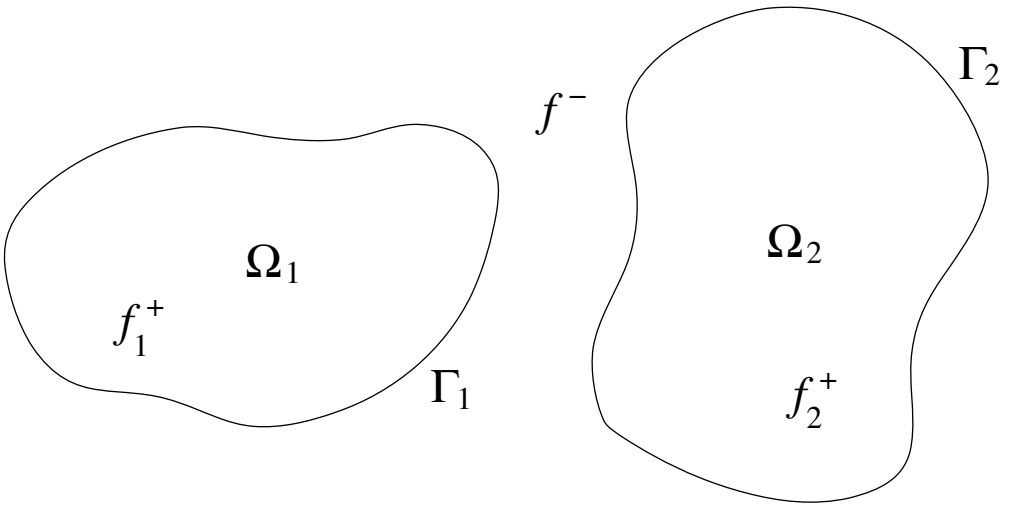


FIG. 2.1. *Two domains.*

harmonic acoustic field $u = u(x)$ governed by the inhomogeneous Helmholtz equation:

$$Lu \equiv \Delta u + k^2 u = f = f_1^+ + f_2^+ + f^-, \tag{2.1}$$

where for the sources we have $\text{supp} f_1^+ \subset \Omega_1$, $\text{supp} f_2^+ \subset \Omega_2$, and $\text{supp} f^- \subset \mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2)$. The overall acoustic field can be represented as

$$u = u_1^+ + u_2^+ + u^-, \tag{2.2a}$$

where

$$Lu_1^+ = f_1^+, \tag{2.2b}$$

$$Lu_2^+ = f_2^+, \tag{2.2c}$$

and

$$Lu^- = f^-. \tag{2.2d}$$

Our first goal is to eliminate all sound from the exterior sources f^- inside Ω_1 and Ω_2 while allowing sound from the sources f_1^+ and f_2^+ to propagate freely between Ω_1 and Ω_2. This is to be achieved by introducing a new control source g. After that, the total acoustic field $\tilde{u}$ will be governed by

$$L\tilde{u} = f_1^+ + f_2^+ + f^- + g.$$

We would like to choose the controls g to guarantee

$$\tilde{u}|_{x\in\Omega_1} = (u_1^+ + u_2^+)|_{x\in\Omega_1}$$

and

$$\tilde{u}|_{x\in\Omega_2} = (u_1^+ + u_2^+)|_{x\in\Omega_2}.$$

In other words, the field after the control inside either Ω_1 or Ω_2 should contain no contribution from the sources f^-. Notice that $g = -f^-$ is a solution to the problem, but it can be very difficult to implement and also requires previous knowledge of the sources f^-. Therefore, other, less expensive, solutions that do not require extensive knowledge of the exterior sources are preferable.

Our second goal is to selectively eliminate the sound that propagates between the regions Ω_1 and Ω_2. This is to be done in addition to the cancellation of the common exterior sound. For example, Ω_1 may be allowed to hear Ω_2, but not vice versa.

Note that the problem of active noise control as formulated above is, in fact, a problem of enabling a desired change in the solution of a given differential equation by appropriately modifying its source terms, i.e., by adding new sources. Consequently, it can be interpreted as an inverse source problem for the corresponding differential equation. Inverse source problems have been extensively studied in the literature, both from the standpoint of physics/engineering (see, e.g., [6, 5]), as well as from the standpoint of mathematics (see, e.g., [10]).

2.2. General solution. Let us first recall that in order to guarantee uniqueness of the solution to the Helmholtz equation (2.1) on unbounded regions, we must require that this solution satisfy the Sommerfeld radiation condition at infinity:

$$\frac{\partial v(x)}{\partial |x|} + ikv(x) = o(|x|^{-1/2}), \quad x \in \mathbb{R}^2, \tag{2.3a}$$

or

$$\frac{\partial v(x)}{\partial |x|} + ikv(x) = o(|x|^{-1}), \quad x \in \mathbb{R}^3. \tag{2.3b}$$

In particular, for any sufficiently smooth function $v = v(x)$ that satisfies the Sommerfeld condition we get

$$v(x) = \int_{\mathbb{R}^n} G(x-y) Lv(y) dy, \tag{2.4}$$

where

$$Lv = \Delta v + k^2 v$$

is the Helmholtz operator and $G = G(x)$ is its fundamental solution on $\mathbb{R}^n$. For $\mathbb{R}^2$, the fundamental solution is given by

$$G(x) = -\frac{1}{4i} H_0^{(2)}(k|x|), \tag{2.5}$$

where $H_0^{(2)}(z)$ is the Hankel function of the second kind defined by means of the Bessel functions $J_0(z)$ and $Y_0(z)$ as $H_0^{(2)}(z) = J_0(z) - iY_0(z)$. For $\mathbb{R}^3$, we have

$$G(x) = -\frac{1}{4\pi} \frac{e^{-ik|x|}}{|x|}. \tag{2.6}$$

Note that the fundamental solutions (2.5) and (2.6) satisfy the Sommerfeld radiation condition at infinity (2.3a) and (2.3b), respectively.

2.2.1. Straightforward cancellation. Let $u = u(x)$ be the overall acoustic field (see (2.1)) and n be the exterior normal to the boundary, and introduce an auxiliary function $w = w(x)$ such that

$$w\big|_{\Gamma_1\cup\Gamma_2} = u\big|_{\Gamma_1\cup\Gamma_2}$$

and

$$\frac{\partial w}{\partial n}\Big|_{\Gamma_1\cup\Gamma_2} = \frac{\partial u}{\partial n}\Big|_{\Gamma_1\cup\Gamma_2}$$

(recall that $\operatorname{dist}(\Gamma_1,\Gamma_2) \geq \epsilon > 0$). We also require that $w(x)$ satisfies the Sommerfeld condition (2.3a) or (2.3b). Next, we define the control sources as follows:

$$g(x) = \begin{cases} -Lw, & x \in \{\mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2)\}, \\ 0, & x \in (\Omega_1 \cup \Omega_2). \end{cases} \tag{2.7}$$

To analyze properties of the controls (2.7), we must determine their output $v = v(x)$ for $x \in \mathbb{R}^n$. Using (2.4), we get[1]

$$\begin{aligned} v(x) &= \int_{\mathbb{R}^n} G g dy = -\int_{\mathbb{R}^n\backslash(\Omega_1\cup\Omega_2)} GLw dy \\ &= -\left(w(x) - \int_{\Omega_1} GLw dy - \int_{\Omega_2} GLw dy \right), \end{aligned}$$

where the individual integrals on the right-hand side are computed by integrating over Ω_1 and Ω_2 and are completely independent. Yet we emphasize that even though the computation of $v(x)$ can be reduced to integration over Ω_1 and Ω_2, the shape of $w(x)$ inside these two domains will not affect the output $v(x)$ since the original controls g are defined outside of $\Omega_1 \cup \Omega_2$; see formula (2.7).

Let us examine the individual terms. By Green's theorem, for $x \in \Omega_1$ we obtain

$$\begin{aligned} w(x) - \int_{\Omega_1} GLw dy &= \int_{\Gamma_1} \left(w \frac{\partial G}{\partial n} - \frac{\partial w}{\partial n} G \right) ds_y \\ &= \int_{\Gamma_1} \left(u \frac{\partial G}{\partial n} - \frac{\partial u}{\partial n} G \right) ds_y \\ &= u^-(x) + u_2^+(x), \qquad x \in \Omega_1, \end{aligned}$$

where n is the normal exterior to Γ_1 and $u^-(x)$ and $u_2^+(x)$ are defined by (2.2d) and (2.2c), respectively. This expression yields the entire incoming component of the field for the domain Ω_1. Next, we need to see what the contribution of $-\int_{\Omega_1} GLw dy$ will be outside of Ω_1. Introduce a smooth auxiliary function $w_1(x)$ such that $w_1(x) = w(x)$ on Ω_1 and $w_1(x)$ is compactly supported on a small neighborhood of Ω_1. Then, for

[1] All integrals hereafter are of the convolution type, as in formula (2.4).

$x \in \mathbb{R}^n \backslash \Omega_1$ we have

$$\begin{aligned} -\int_{\Omega_1} GLwdy &= -\int_{\Omega_1} GLw_1 dy \\ &= -\int_{\Omega_1} GLw_1 dy + w_1 - w_1 \\ &= \int_{\mathbb{R}^n \backslash \Omega_1} GLw_1 dy - w_1 \\ &= \int_{\Gamma_1} \left(w_1 \frac{\partial G}{\partial n} - \frac{\partial w_1}{\partial n} G \right) ds_y \\ &= -u_1^+(x), \qquad x \in \mathbb{R}^n \backslash \Omega_1, \end{aligned}$$

where the third equality in the chain is obtained with the help of formula (2.4) applied to $w_1(x)$. Therefore, we can write

$$-\int_{\Omega_1} GLwdy = \begin{cases} -u_1^+, & x \in \mathbb{R}^n \backslash \Omega_1, \\ -w + u^- + u_2^+, & x \in \Omega_1. \end{cases}$$

We also have a similar output from Ω_2 given by

$$-\int_{\Omega_2} GLwdy = \begin{cases} -u_2^+, & x \in \mathbb{R}^n \backslash \Omega_2, \\ -w + u^- + u_1^+, & x \in \Omega_2. \end{cases}$$

Altogether, the full output of the controls $g(x)$ of (2.7) is as follows:

$$\begin{aligned} v(x) &= -\left(w - \int_{\Omega_1} GLwdy - \int_{\Omega_2} GLwdy \right) \\ &= \begin{cases} -(u^- + u_2^+) + u_2^+ = -u^-, & x \in \Omega_1, \\ -(u^- + u_1^+) + u_1^+ = -u^-, & x \in \Omega_2, \\ -(w - u_1^+ - u_2^+), & x \in \mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2). \end{cases} \end{aligned}$$

Consequently, these controls enable the cancellation of sound due to the exterior sources f^- on the domains Ω_1 and Ω_2 regardless of the specific choice of the auxiliary function w. The output of the controls outside $\Omega_1 \cup \Omega_2$ is given by $u_1^+ + u_2^+ - w$. It duplicates the acoustic field generated inside the two regions with the correction $-w$.

Let us elaborate a little further on the structure of the control output $v(x)$. Assume that $x \in \Omega_1$. Then,

$$\begin{aligned} v(x) &= -w(x) + \int_{\Omega_1} GLwdy + \int_{\Omega_2} GLwdy \\ &= -w + \underbrace{w - (u^- + u_2^+)}_{\text{contribution due to } \Omega_1} + \underbrace{u_2^+}_{\text{due to } \Omega_2} \\ &= -u^-(x), \ x \in \Omega_1, \end{aligned}$$

where $-(u^- + u_2^+)$ from the second term above renders cancellation of the entire incoming wave for Ω_1, and u_2^+ is the interior sound from Ω_2 duplicated by the controls. The same is true for Ω_2. Hence we conclude that the controls double the output of the sources interior to a region on the way out and then halve it as it comes into the other region. As such, the overall acoustic field after the control is given by

$$\begin{aligned} u &= u_1^+ + u_2^+ + u^- + v \\ &= \begin{cases} u_1^+ + u_2^+, & x \in \Omega_1, \\ u_1^+ + u_2^+, & x \in \Omega_2, \\ -w + u^- + 2u_1^+ + 2u_2^+, & x \in \mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2), \end{cases} \end{aligned} \tag{2.8}$$

allowing the domains Ω_1 and Ω_2 to communicate freely with each other without interference from outside sources.

2.2.2. Selective cancellation. Now suppose that we would like Ω_1 to hear Ω_2 without outside interference, but we do not allow Ω_2 to hear anything from outside its boundary, including Ω_1. To achieve this we must elaborate further on how the split between the incoming and outgoing waves works. Consider just one domain Ω_1 with the boundary Γ_1 and again choose the auxiliary function $w_1 = w_1(x)$. Let

$$w_1\big|_{\Gamma_1} = u\big|_{\Gamma_1} \tag{2.9a}$$

and

$$\frac{\partial w_1}{\partial n}\Big|_{\Gamma_1} = \frac{\partial u}{\partial n}\Big|_{\Gamma_1}, \tag{2.9b}$$

where $u = u_1^- + u_1^+$ is the total acoustic field and $u_1^- = u^- + u_2^+$ is the acoustic field generated outside of Ω_1. Then the surface integral gives us

$$\int_{\Gamma_1} \left(w_1 \frac{\partial G}{\partial n} - \frac{\partial w_1}{\partial n} G \right) ds_y = \begin{cases} u_1^-, & x \in \Omega_1, \\ -u_1^+, & x \in \mathbb{R}^n \backslash \Omega_1. \end{cases}$$

With respect to the domain Ω_1, the field u_1^- is incoming, and u_1^+ is outgoing. Assuming that $w_1(x)$ also satisfies the appropriate Sommerfeld radiation condition (2.3a) or (2.3b), the surface integral can be replaced by the volumetric integral, so that for $x \in \Omega_1$ we have

$$\begin{aligned} \int_{\Gamma_1} \left(w_1 \frac{\partial G}{\partial n} - \frac{\partial w_1}{\partial n} G \right) ds_y &= w_1 - \int_{\Omega_1} GLw_1 dy \Big|_{x \in \Omega_1} \\ &= \int_{\mathbb{R}^n \backslash \Omega_1} G(x-y) Lw_1(y) dy \Big|_{x \in \Omega_1}. \end{aligned}$$

This is precisely why we would choose the controls as $g_1(x) = -Lw_1|_{\mathbb{R}^n \backslash \Omega_1}$ if we were to completely eliminate all of the outside sound on Ω_1—because they produce $-u_1^-$ on Ω_1. At the same time, on the complementary domain $\mathbb{R}^n \backslash \Omega_1$ the output of the

controls $g_1(x)$ is the duplicate of the outgoing field u_1^+ corrected by $-w_1$:

$$\begin{aligned}\int_{\mathbb{R}^n} G(x-y)g_1(y)dy &= -\int_{\mathbb{R}^n\setminus\Omega_1} G(x-y)Lw_1(y)dy \\ &= w_1 - \int_{\mathbb{R}^n\setminus\Omega_1} GLw_1 dy - w_1 \\ &= -\int_{\Gamma_1}\left(w_1\frac{\partial G}{\partial n} - \frac{\partial w_1}{\partial n}G\right)ds_y - w_1 = u_1^+ - w_1.\end{aligned} \tag{2.10}$$

Having described individual controls g_1 for a single domain Ω_1, we are now ready to construct the controls so that Ω_1 will hear Ω_2 without outside interference, but Ω_2 will not hear anything from outside its boundary, including Ω_1. The procedure will consist of two stages. At the first stage, we will use the controls $g_1(x)$ as a supplementary tool. Namely, choose an auxiliary function $w_1(x)$ that satisfies conditions (2.9a) and (2.9b), as well as the Sommerfeld condition at infinity. In addition, require that w_1 be compactly supported near Ω_1, in particular, that $w_1(x) = 0$ near Ω_2. This is clearly possible since the distance between the subdomains Ω_1 and Ω_2 is positive. Then, build the supplementary controls

$$g_1(x) = -Lw_1|_{\mathbb{R}^n\setminus\Omega_1}, \qquad g_1(x) = 0|_{\Omega_1}. \tag{2.11}$$

According to formula (2.10), the output of these controls on $\mathbb{R}^n\setminus\Omega_1$ is

$$v_1 = \int_{\mathbb{R}^n} Gg_1 dy = u_1^+ - w_1, \qquad x \in \mathbb{R}^n\setminus\Omega_1, \tag{2.12}$$

and since w_1 is compactly supported near Ω_1, we have $v_1 = u_1^+$ near Ω_2.

At the second stage of building the controls, we begin as usual with our auxiliary function w. It is still required that w satisfy the Sommerfeld condition at infinity, while on Γ_1 we still impose the same boundary conditions (2.9a) and (2.9b):

$$w\big|_{\Gamma_1} = u\big|_{\Gamma_1}$$

and

$$\frac{\partial w}{\partial n}\Big|_{\Gamma_1} = \frac{\partial u}{\partial n}\Big|_{\Gamma_1},$$

where u is the given total acoustic field. The difference is in the boundary conditions on Γ_2. Here it is required that

$$w\big|_{\Gamma_2} = (u+v_1)\big|_{\Gamma_2} \equiv (u+u_1^+)\big|_{\Gamma_2}$$

and

$$\frac{\partial w}{\partial n}\Big|_{\Gamma_2} = \frac{\partial (u+v_1)}{\partial n}\Big|_{\Gamma_2} \equiv \frac{\partial (u+u_1^+)}{\partial n}\Big|_{\Gamma_2},$$

where v_1 was obtained at the first stage; see (2.12). Then, defining the controls as

$$g(x) = -Lw|_{\mathbb{R}^n\setminus(\Omega_1\cup\Omega_2)}, \qquad g(x) = 0|_{(\Omega_1\cup\Omega_2)}$$

yields the output

$$v(x) = -\left(w - \int_{\Omega_1} GLwdy - \int_{\Omega_2} GLwdy\right)$$

$$= \begin{cases} -(u^- + u_2^+) + u_2^+ = -u^-, & x \in \Omega_1, \\ -(u^- + 2u_1^+) + u_1^+ = -(u^- + u_1^+), & x \in \Omega_2, \\ -(w - u_1^+ - u_2^+), & x \in \mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2). \end{cases} \tag{2.13}$$

Therefore, we see that Ω_1 hears Ω_2 without outside interference, but Ω_2 does not hear anything from outside its boundary, including Ω_1.

2.2.3. Proofs. We will now prove that what we have obtained is, in fact, a general solution for the controls with the prescribed properties. That is, we will prove that our method of construction gives all possible controls.

THEOREM 2.1. *Suppose that $\Omega_1 \subset \mathbb{R}^n$ and $\Omega_2 \subset \mathbb{R}^n$ are two disjoint regions:* $\operatorname{dist}(\Omega_1, \Omega_2) \geq \epsilon > 0$, *with the boundaries $\partial\Omega_1 = \Gamma_1$ and $\partial\Omega_2 = \Gamma_2$. Assume that the total acoustic field in $\mathbb{R}^n$ is governed by $Lu \equiv \Delta u + k^2 u = f = f_1^+ + f_2^+ + f^-$, where the sources f are located according to* $\operatorname{supp} f_1^+ \subset \Omega_1$, $\operatorname{supp} f_2^+ \subset \Omega_2$, *and* $\operatorname{supp} f^- \subset \mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2)$. *Let the overall acoustic field u be represented as $u = u_1^+ + u_2^+ + u^-$.*

Let a control source $g = g(x)$ be added to the other sources $f(x)$ such that the overall field $\tilde{u}$ governed by $L\tilde{u} = f_1^+ + f_2^+ + f^- + g$ satisfies

$$\tilde{u} = \begin{cases} u_1^+ + u_2^+, & x \in \Omega_1, \\ u_1^+ + u_2^+, & x \in \Omega_2. \end{cases} \tag{2.14}$$

Then the general solution for the desired control is given by

$$g = -Lw|_{\mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2)}, \qquad g = 0|_{(\Omega_1 \cup \Omega_2)}, \tag{2.15}$$

where w satisfies the Sommerfeld condition (2.3a) *or* (2.3b) *at infinity, as well as the interface conditions*

$$w\big|_{\Gamma_1 \cup \Gamma_2} = u\big|_{\Gamma_1 \cup \Gamma_2} \tag{2.16a}$$

and

$$\frac{\partial w}{\partial n}\Big|_{\Gamma_1 \cup \Gamma_2} = \frac{\partial u}{\partial n}\Big|_{\Gamma_1 \cup \Gamma_2}. \tag{2.16b}$$

Proof. We need to prove that any control g given by (2.15) is an appropriate control and, conversely, that any appropriate control g can be obtained by using a suitable auxiliary function w. Suppose we have a function $w(x)$ that satisfies (2.16a), (2.16b), and the Sommerfeld condition at infinity. Then, according to formula (2.8), the corresponding control g given by formula (2.15) yields the desired properties by eliminating u^- on $\Omega_1 \cup \Omega_2$.

Conversely, suppose that a control g achieves the desired cancellation; see formula (2.14). Then, substituting $\tilde{u} = \tilde{u}(x)$ into the equation $L\tilde{u} = f_1^+ + f_2^+ + f^- + g$, we immediately obtain that $g(x) = 0$ for $x \in (\Omega_1 \cup \Omega_2)$. In other words, $\operatorname{supp} g \subset \mathbb{R}^n \setminus (\Omega_1 \cup \Omega_2)$. Consequently, the output v of the control g is as follows:

$$v(x) = \int_{\mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2)} Ggdy = \begin{cases} -u^-, & x \in \Omega_1, \\ -u^-, & x \in \Omega_2. \end{cases}$$

Consider the equation $-Lw = g - f_1^+ - f_2^+$, where f_1^+ and f_2^+ are the sound sources from Ω_1 and Ω_2, respectively. Its solution, subject to the Sommerfeld condition at infinity (2.3a) or (2.3b), satisfies

$$w = -v + u_1^+ + u_2^+$$
$$= \begin{cases} u^- + u_1^+ + u_2^+ = u, & x \in \Omega_1, \\ u^- + u_1^+ + u_2^+ = u, & x \in \Omega_2. \end{cases}$$

Since $w(x)$ is at least C^1 smooth on $\mathbb{R}^n$, we can claim that it satisfies relations (2.16a) and (2.16b). Therefore, the control $g(x)$ can be obtained by formula (2.15), since $\mathrm{supp} f_1^+ \subset \Omega_1$ and $\mathrm{supp} f_2^+ \subset \Omega_2$. □

THEOREM 2.2. *Suppose that $\Omega_1 \subset \mathbb{R}^n$ and $\Omega_2 \subset \mathbb{R}^n$ are two disjoint regions:* $\mathrm{dist}(\Omega_1, \Omega_2) \geq \epsilon > 0$, *with the boundaries $\partial\Omega_1 = \Gamma_1$ and $\partial\Omega_2 = \Gamma_2$. Assume that the total acoustic field in $\mathbb{R}^n$ is governed by $Lu \equiv \Delta u + k^2 u = f = f_1^+ + f_2^+ + f^-$, where the sources f are located according to* $\mathrm{supp} f_1^+ \subset \Omega_1$, $\mathrm{supp} f_2^+ \subset \Omega_2$, *and* $\mathrm{supp} f^- \subset \mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2)$. *Let the overall acoustic field u be represented as $u = u_1^+ + u_2^+ + u^-$.*

Let a control source $g = g(x)$ be added to the other sources $f(x)$ such that the overall field $\tilde{u}$ governed by $L\tilde{u} = f_1^+ + f_2^+ + f^- + g$ satisfies

$$\tilde{u} = \begin{cases} u_1^+ + u_2^+, & x \in \Omega_1, \\ u_2^+, & x \in \Omega_2. \end{cases} \tag{2.17}$$

Then the general solution for the desired control is given by

$$g = -Lw|_{\mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2)}, \qquad g = 0|_{(\Omega_1 \cup \Omega_2)}, \tag{2.18}$$

where $w = w(x)$ satisfies the Sommerfeld condition (2.3a) *or* (2.3b) *at infinity and the following interface conditions:*

$$w\big|_{\Gamma_1} = u\big|_{\Gamma_1}, \quad w\big|_{\Gamma_2} = (u + u_1^+)\big|_{\Gamma_2} \tag{2.19a}$$

and

$$\frac{\partial w}{\partial n}\Big|_{\Gamma_1} = \frac{\partial u}{\partial n}\Big|_{\Gamma_1}, \quad \frac{\partial w}{\partial n}\Big|_{\Gamma_2} = \frac{\partial (u + u_1^+)}{\partial n}\Big|_{\Gamma_2}. \tag{2.19b}$$

The function u_1^+ on Γ_2 can be obtained as the output v_1 given by formula (2.12) *of the supplementary controls g_1 of* (2.11).

Theorem 2.2 essentially implies that the controls (2.18) are obtained by means of a predictor-corrector procedure. The predictor stage consists of computing v_1 of (2.12) as the output of the control g_1 of (2.11), whereas the corrector stage consists of obtaining the overall composite controls $g(x)$ with the help of the auxiliary function $w(x)$ defined via (2.19).

Proof. We need to prove that any control g given by (2.18) is an appropriate control and, conversely, that any appropriate control g can be obtained by using a suitable auxiliary function w. Suppose we have a function $w(x)$ that satisfies (2.19a), (2.19b), and the Sommerfeld condition at infinity. Then, according to formula (2.13), the corresponding control given by (2.18) provides the desired properties eliminating u^- on $\Omega_1 \cup \Omega_2$ and additionally eliminating u_1^+ on Ω_2.

Conversely, suppose that a control g achieves the desired cancellation; see formula (2.17). Then, $\operatorname{supp} g \in \mathbb{R}^n \setminus (\Omega_1 \cup \Omega_2)$, and the output of the control, v, is as follows:

$$v(x) = \int_{\mathbb{R}^n \setminus (\Omega_1 \cup \Omega_2)} Ggdy = \begin{cases} -u^-, & x \in \Omega_1, \\ -u^- - u_1^+, & x \in \Omega_2. \end{cases}$$

Consider the equation $-Lw = g - f_1^+ - f_2^+$. Its solution, subject to the Sommerfeld condition at infinity (2.3a) or (2.3b), satisfies

$$\begin{aligned} w &= -v + u_1^+ + u_2^+ \\ &= \begin{cases} u^- + u_1^+ + u_2^+ = u, & x \in \Omega_1, \\ u^- + 2u_1^+ + u_2^+ = u + u_1^+, & x \in \Omega_2. \end{cases} \end{aligned}$$

Since $w(x)$ is at least C^1 smooth on $\mathbb{R}^n$, it satisfies relations (2.19a) and (2.19b). Therefore, the control $g(x)$ can be obtained by formula (2.18) since $\operatorname{supp} f_1^+ \subset \Omega_1$ and $\operatorname{supp} f_2^+ \subset \Omega_2$. ∎

3. Multiple regions.

3.1. Formulation. Let $\Omega_1, \Omega_2, \ldots, \Omega_N$ be given, where $\Omega_i \subseteq \mathbb{R}^2$ or $\mathbb{R}^3$ is either bounded or unbounded. For simplicity we will assume that $\Omega_1, \Omega_2, \ldots, \Omega_N$ are separate bounded regions of $\mathbb{R}^n$. Let $\Gamma_1, \Gamma_2, \ldots, \Gamma_N$ be the boundaries of $\Omega_1, \Omega_2, \ldots, \Omega_N$ respectively. Consider the time-harmonic acoustic field u governed by the inhomogeneous Helmholtz equation:

$$Lu \equiv \Delta u + k^2 u = f = f_1^+ + f_2^+ + \cdots + f_N^+ + f^-,$$

where the sources are $\operatorname{supp} f_1^+ \subset \Omega_1$, $\operatorname{supp} f_2^+ \subset \Omega_2, \ldots, \operatorname{supp} f_n^+ \subset \Omega_N$, and $\operatorname{supp} f^- \subset \mathbb{R}^n \setminus (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N)$. Therefore, the overall acoustic field can be represented as

$$u = u_1^+ + u_2^+ + \cdots + u_N^+ + u^-,$$

where

$$\begin{aligned} Lu_1^+ &= f_1^+, \\ Lu_2^+ &= f_2^+, \\ &\cdots \\ Lu_N^+ &= f_N^+, \end{aligned}$$

and

$$Lu^- = f^-.$$

Our goal is to eliminate all sound from the sources f^- inside $\Omega_1, \Omega_2, \ldots, \Omega_N$, while allowing sound from the sources $f_1^+, f_2^+, \ldots, f_N^+$ to propagate between $\Omega_1, \Omega_2, \ldots, \Omega_N$ as we see fit. That is, we wish to selectively eliminate unwanted sound from various regions while leaving other regions free to receive predetermined communications. This is done as before by introducing a new control source g. Therefore, the total acoustic field is now governed by the modified equation

$$L\tilde{u} = f_1^+ + f_2^+ + \cdots + f_N^+ + f^- + g.$$

3.2. General solution.

3.2.1. Straightforward cancellation. We will first demonstrate how to eliminate all sound in $\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N$ that originates from $\mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N)$. As before, we introduce an auxiliary function $w = w(x)$, which satisfies the Sommerfeld condition (2.3a) or (2.3b) at infinity, and is such that

$$w\big|_{\Gamma_i} = u\big|_{\Gamma_i}$$

and

$$\frac{\partial w}{\partial n}\Big|_{\Gamma_i} = \frac{\partial u}{\partial n}\Big|_{\Gamma_i}$$

for all $i = 1, \ldots, N$.

Next, we define the control sources as (cf. formula (2.7))

$$g(x) = \begin{cases} -Lw, & x \in \{\mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N)\}, \\ 0, & x \in (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N), \end{cases}$$

and see that their output $v = v(x)$, $x \in \mathbb{R}^n$, is given by

$$v(x) = \int_{\mathbb{R}^n} G g dy = -\int_{\mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N)} GLw dy$$

$$= -\left(w(x) - \int_{\Omega_1} GLw dy - \int_{\Omega_2} GLw dy - \cdots - \int_{\Omega_N} GLw dy \right),$$

where the individual integrals are computed by integrating over $\Omega_1, \Omega_2, \ldots, \Omega_N$ and are completely independent. Again, all integrals are convolutions, as in section 2.

Let us examine the individual terms. By Green's theorem, for $x \in \Omega_i$, where $i \in \{1, 2, \ldots, N\}$, we obtain

$$w(x) - \int_{\Omega_i} GLw dy = \int_{\Gamma_i} \left(w \frac{\partial G}{\partial n} - \frac{\partial w}{\partial n} G \right) ds_y$$

$$= \int_{\Gamma_i} \left(u \frac{\partial G}{\partial n} - \frac{\partial u}{\partial n} G \right) ds_y$$

$$= u^-(x) + \sum_{\substack{j=1,2,\ldots,N \\ j \neq i}} u_j^+(x), \qquad x \in \Omega_i.$$

This is the entire incoming component for the domain Ω_i. Now the effect of the integral $-\int_{\Omega_i} GLw dy$ outside of Ω_i must be examined. To do this, we introduce a smooth auxiliary function $w_i(x)$ such that $w_i(x) = w(x)$ on Ω_i and $w_i(x)$ is compactly

supported on a small neighborhood of Ω_i. Consequently, for $x \in \mathbb{R}^n \backslash \Omega_i$ we have

$$\begin{aligned}
-\int_{\Omega_i} GLwdy &= -\int_{\Omega_i} GLw_i dy \\
&= -\int_{\Omega_i} GLw_i dy + w_i - w_i \\
&= \int_{\mathbb{R}^n \backslash \Omega_i} GLw_i dy - w_i \\
&= \int_{\Gamma_i} \left(w_i \frac{\partial G}{\partial n} - \frac{\partial w_i}{\partial n} G \right) ds_y \\
&= -u_i^+(x), \quad x \in \mathbb{R}^n \backslash \Omega_i.
\end{aligned}$$

Therefore we can write

$$-\int_{\Omega_i} GLwdy = \begin{cases} -u_i^+, & x \in \mathbb{R}^n \backslash \Omega_i, \\ -w + u^- + \displaystyle\sum_{\substack{j=1,2,\ldots,N, \\ j \neq i}} u_j^+, & x \in \Omega_i. \end{cases}$$

Altogether, the full output of the controls $g(x)$ is as follows:

$$\begin{aligned}
v(x) &= -\left(w - \int_{\Omega_i} GLwdy - \sum_{j \neq i} \int_{\Omega_j} GLwdy \right) \\
&= \begin{cases} -u^-, & x \in \Omega_i,\ i = 1, 2, \ldots, N, \\ -\left(w - \displaystyle\sum_{j=1,2,\ldots,N} u_j^+ \right), & x \in \mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N). \end{cases}
\end{aligned} \tag{3.1}$$

Consequently, these controls enable the cancellation of sound due to the exterior sources on the domains $\Omega_1, \Omega_2, \ldots, \Omega_N$ regardless of the specific choice of the auxiliary function w. The output of the controls outside $\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N$ is given by $\sum_{j=1,2,\ldots,N} u_j^+ - w$. It basically duplicates the acoustic field generated inside the regions with the correction $-w$. More specifically, for any given Ω_i the controls double the output of the sources interior to a region on the way out and then halve the result as it comes into another region. As such, the overall acoustic field is given by

$$\begin{aligned}
u &= u_1^+ + u_2^+ + \cdots + u_N^+ + u^- + v \\
&= \begin{cases} u_1^+ + u_2^+ + \cdots + u_N^+, & x \in \Omega_i,\ i = 1, 2, \ldots, N, \\ -w + u^- + 2u_1^+ + 2u_2^+ + \cdots + 2u_N^+, & x \in \mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N), \end{cases}
\end{aligned}$$

which means that the subdomains $\Omega_1, \Omega_2, \ldots, \Omega_N$ can communicate freely with each other without interference from outside sources.

THEOREM 3.1. *Suppose that $\Omega_1, \Omega_2, \ldots, \Omega_N$ are given, where $\Omega_i \subseteq \mathbb{R}^n$ are disjoint regions:* $\operatorname{dist}(\Omega_i, \Omega_j) \geq \epsilon > 0$ *if $i \neq j$, with the boundaries $\partial\Omega_1 = \Gamma_1, \partial\Omega_2 =$*

$\Gamma_2, \ldots, \partial\Omega_n = \Gamma_N$. *Assume that the total acoustic field in* $\mathbb{R}^n$ *is governed by* $Lu \equiv \Delta u + k^2 u = f = f_1^+ + f_2^+ + \cdots + f_N^+ + f^-$, *where the sources are located according to* $\mathrm{supp} f_1^+ \subset \Omega_1$, $\mathrm{supp} f_2^+ \subset \Omega_2, \ldots, \mathrm{supp} f_n^+ \subset \Omega_N$, *and* $\mathrm{supp} f^- \subset \mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N)$. *Let the overall acoustic field* u *be represented as* $u = u_1^+ + u_2^+ + \cdots + u_N^+ + u^-$.

Let a control source $g = g(x)$ *be added to the other sources* $f(x)$ *such that the overall field* $\tilde{u}$ *governed by* $L\tilde{u} = f_1^+ + f_2^+ + \cdots + f_N^+ + f^- + g$ *satisfies*

$$\tilde{u} = \sum_{j=1,2,\ldots,N} u_j^+, x \in \Omega_i, \ i = 1, 2, \ldots, N. \tag{3.2}$$

Then the general solution for the desired control is given by

$$g = -Lw|_{\mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N)}, \qquad g = 0|_{(\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N)}, \tag{3.3}$$

where $w = w(x)$ *satisfies the Sommerfeld condition at infinity and the following interface conditions:*

$$w\big|_{\Gamma_i} = u\big|_{\Gamma_i} \tag{3.4a}$$

and

$$\frac{\partial w}{\partial n}\Big|_{\Gamma_i} = \frac{\partial u}{\partial n}\Big|_{\Gamma_i} \tag{3.4b}$$

for all $i = 1, \ldots, N$.

Proof. We need to prove that any control g given by (3.3) is an appropriate control and, conversely, that any appropriate control g can be obtained by using a suitable auxiliary function w. Suppose we have a function $w(x)$ that satisfies (3.4a), (3.4b), and the Sommerfeld condition (2.3a) or (2.3b) at infinity. Then, formula (3.1) implies that the corresponding control (3.3) provides the desired properties eliminating the exterior sound u^- on $\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N$.

Conversely, suppose a control g achieves the desired cancellation, so that equality (3.2) holds. Then, clearly, $g(x) = 0$ for $x \in \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N$. Consequently, the output v of the control g is as follows:

$$v(x) = \int_{\mathbb{R}^n \backslash (\Omega_1 \cup \Omega_2 \cup \ldots \Omega_N)} G g dy = -u^-, \quad x \in \Omega_i, \ i = 1, \ldots N.$$

Consider the equation $-Lw = g - f_1^+ - f_2^+ - \cdots - f_N^+$. Its solution, subject to the Sommerfeld condition at infinity (2.3a) or (2.3b), satisfies

$$\begin{aligned} w &= -v + u_1^+ + u_2^+ + \cdots + u_N^+ \\ &= u, \quad x \in \Omega_i, \ i = 1, 2, \ldots, N. \end{aligned}$$

Since $w(x)$ is at least C^1 smooth on $\mathbb{R}^n$, it satisfies relations (3.4a) and (3.4b). Therefore the control $g(x)$ can be obtained by formula (3.3) applied to this particular $w(x)$, since $\mathrm{supp} f_1^+ \subset \Omega_1$, $\mathrm{supp} f_2^+ \subset \Omega_2, \ldots, \mathrm{supp} f_N^+ \subset \Omega_N$. ∎

3.2.2. Selective cancellation. Now suppose that in each subdomain Ω_i, we would like to eliminate all outside interference and, in addition, selectively eliminate sound from some other subdomains. It will be helpful to formulate a convenient way of keeping track of communications between the subdomains. For that purpose, let us

introduce an $N \times N$ matrix $\mathbf{M}$, such that each row i corresponds to a region Ω_i, and the entry (0 or 1) in each column is used to determine whether this Ω_i hears a region corresponding to that column or not. In other words, if the entry at the intersection of row i and column j is 0, then Ω_i hears Ω_j. If this entry is 1, then it does not. Obviously the diagonal of $\mathbf{M}$ is filled with zeros since the regions hear themselves. So, in the case of Theorem 2.1 we have

$$\mathbf{M} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

and for Theorem 2.2 we get

$$\mathbf{M} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

Notice that no reciprocity in the communication pattern is assumed; i.e., the matrix $\mathbf{M}$ is not necessarily symmetric.

For a given matrix $\mathbf{M}$ that corresponds to a specific communication pattern between the regions $\Omega_1, \Omega_2, \ldots, \Omega_N$, we will now build the auxiliary function $w(x)$ and the controls $g(x)$ as before, i.e., in two stages. At the first stage, we take the auxiliary functions $w_i(x)$ for all Ω_i, $i = 1, \ldots, N$, that satisfy

$$w_i\big|_{\Gamma_i} = u\big|_{\Gamma_i}$$

and

$$\frac{\partial w_i}{\partial n}\Big|_{\Gamma_i} = \frac{\partial u}{\partial n}\Big|_{\Gamma_i},$$

as well as the Sommerfeld condition at infinity. We also require that each w_i be compactly supported near the corresponding Ω_i. Then, we build the supplementary controls:

$$g_i(x) = -Lw_i|_{\mathbb{R}^n \setminus \Omega_i}, \qquad g_i(x) = 0|_{\Omega_i}.$$

According to formula (2.10) applied to a given subdomain Ω_i, the output of these controls on $\mathbb{R}^n \setminus \Omega_i$ is

$$v_i = \int_{\mathbb{R}^n} G g_i dy = u_i^+ - w_i, \quad x \in \mathbb{R}^n \setminus \Omega_i, \tag{3.5}$$

and since w_i is taken compactly supported near Ω_i, we have $v_i = u_i^+$ near Ω_j, where $j = 1, 2, \ldots, N$ and $j \neq i$.

At the second stage, we start with introducing the auxiliary function $w(x)$, which satisfies the Sommerfeld radiation condition (2.3a) or (2.3b) at infinity. In addition, on each Γ_i we require that

$$w\big|_{\Gamma_i} = (u + \mathbf{e}_i^T \mathbf{M} \mathbf{v})\big|_{\Gamma_i}$$

and

$$\frac{\partial w}{\partial n}\Big|_{\Gamma_i} = \frac{\partial (u + \mathbf{e}_i^T \mathbf{M} \mathbf{v})}{\partial n}\Big|_{\Gamma_i}.$$

In these formulae, $\mathbf{e}_i$ is a vector with its ith component equal to 1 and all other components equal to 0, and

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{pmatrix},$$

where each v_i is obtained at the first stage with the help of the supplementary controls $g_i(x)$ according to formula (3.5).

Next, we define the control sources $g(x)$ as

$$g(x) = -Lw|_{\mathbb{R}^n \setminus (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N)}, \qquad g(x) = 0|_{(\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N)}.$$

Their output $v = v(x)$, $x \in \mathbb{R}^n$, is given by

$$\begin{aligned} v(x) &= \int_{\mathbb{R}^n} Ggdy \\ &= -\int_{\mathbb{R}^n \setminus (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N)} GLwdy \\ &= -\left(w(x) - \int_{\Omega_1} GLwdy - \int_{\Omega_2} GLwdy - \cdots - \int_{\Omega_N} GLwdy \right), \\ &= \begin{cases} -(u^- + \mathbf{e}_i^T \mathbf{M} \mathbf{v}), & x \in \Omega_i, \\ -(w - u_1^+ - u_2^+ - \cdots - u_N^+), & x \in \mathbb{R}^n \setminus (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N), \end{cases} \end{aligned} \tag{3.6}$$

which obviously enables the desired cancellation.

We now prove that this is in fact the general solution for the controls with the prescribed properties, i.e., that we obtain all possible controls.

THEOREM 3.2. *Suppose that $\Omega_1, \Omega_2, \ldots, \Omega_N$ are given, where $\Omega_i \subseteq \mathbb{R}^n$ are disjoint regions:* $\operatorname{dist}(\Omega_i, \Omega_j) \geq \epsilon > 0$ *if $i \neq j$, with the boundaries $\partial\Omega_1 = \Gamma_1, \partial\Omega_2 = \Gamma_2, \ldots, \partial\Omega_n = \Gamma_N$. Assume that the total acoustic field in $\mathbb{R}^n$ is governed by $Lu \equiv \Delta u + k^2 u = f = f_1^+ + f_2^+ + \cdots + f_N^+ + f^-$, where the sources are located according to* $\operatorname{supp} f_1^+ \subset \Omega_1$, $\operatorname{supp} f_2^+ \subset \Omega_2, \ldots, \operatorname{supp} f_n^+ \subset \Omega_N$, *and* $\operatorname{supp} f^- \subset \mathbb{R}^n \setminus (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N)$. *Let the overall acoustic field u be represented as $u = u_1^+ + u_2^+ + \cdots + u_N^+ + u^-$.*

Let a control source $g = g(x)$ be added to the other sources $f(x)$ such that the overall field $\tilde{u}$ governed by $L\tilde{u} = f_1^+ + f_2^+ + \cdots + f_N^+ + f^- + g$ satisfies

$$\tilde{u} = \mathbf{e}_i^T (\mathbf{1} - \mathbf{M}) \mathbf{u}, \ x \in \Omega_i, \ i = 1, 2, \ldots, N, \tag{3.7}$$

where $\mathbf{1}$ is an $N \times N$ matrix with all entries equal to 1, *and*

$$\mathbf{u} = \begin{pmatrix} u_1^+ \\ u_2^+ \\ \vdots \\ u_N^+ \end{pmatrix}.$$

Then, the general solution for the desired control is given by

$$g = -Lw|_{\mathbb{R}^n \setminus (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N)}, \qquad g = 0|_{(\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N)}, \tag{3.8}$$

where $w = w(x)$ satisfies the Sommerfeld condition at infinity and the following interface conditions:

$$w\big|_{\Gamma_i} = (u + \mathbf{e}_i^T \mathbf{M}\mathbf{v})\big|_{\Gamma_i} \tag{3.9a}$$

and

$$\frac{\partial w}{\partial n}\Big|_{\Gamma_i} = \frac{\partial (u + \mathbf{e}_i^T \mathbf{M}\mathbf{v})}{\partial n}\Big|_{\Gamma_i} \tag{3.9b}$$

for all $i = 1, \ldots, N$. Note that if $\mathbf{M} = \mathbf{0}$, then (3.7) *reduces to* (3.2)*, and the current theorem becomes the same as Theorem* 3.1.

Similarly to Theorem 2.2, Theorem 3.2 implies that the controls (3.8) are built using a predictor-corrector procedure. The predictor stage consists of computing $\mathbf{v}$ of (3.5), whereas the corrector stage consists of obtaining the overall composite controls $g(x)$ by means of the auxiliary function $w(x)$ defined via (3.9).

Proof. We need to prove that any control g given by (3.8) is an appropriate control and, conversely, that any appropriate control g can be obtained by using a suitable auxiliary function w. Suppose we have a function $w(x)$ that satisfies (3.9a), (3.9b), and the Sommerfeld condition at infinity. Then, according to formula (3.6), the corresponding control (3.8) provides the desired properties as it eliminates u^- on $\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N$ and selectively allows the sound to propagate between the subdomains following a predetermined pattern $\mathbf{M}$.

Conversely, suppose that a control g achieves the desired cancellation; see formula (3.7). Then, $g(x) = 0$ for $x \in \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N$, and the output v of the control g is

$$\begin{aligned} v(x) &= \int_{\mathbb{R}^n \setminus (\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N)} G g dy \\ &= -(u^- + \mathbf{e}_i^T \mathbf{M}\mathbf{v}),\ x \in \Omega_i,\ i = 1, 2, \ldots, N. \end{aligned}$$

Consider the equation $-Lw = g - f_1^+ - f_2^+ - \cdots - f_N^+$. Its solution, subject to the Sommerfeld condition at infinity (2.3a) or (2.3b), satisfies

$$\begin{aligned} w &= -v + u_1^+ + u_2^+ + \cdots + u_N^+ \\ &= \mathbf{e}_i^T (\mathbf{1} - \mathbf{M})\mathbf{u},\ x \in \Omega_i,\ i = 1, 2, \ldots, N. \end{aligned}$$

Since $w(x)$ is at least C^1 smooth on $\mathbb{R}^n$, it satisfies the interface conditions (3.9a) and (3.9b). Therefore, the control $g(x)$ can be obtained by formula (3.8) applied to this $w(x)$, because $\operatorname{supp} f_1^+ \subset \Omega_1$, $\operatorname{supp} f_2^+ \subset \Omega_2, \ldots, \operatorname{supp} f_N^+ \subset \Omega_N$. ∎

4. Generalized Calderon's potentials. We will now show how the split between $u_1^+, u_2^+, \ldots, u_N^+$, and u^- can be conveniently described in terms of the generalized potentials and boundary projection operators of Calderon's type. For more detail, the reader is referred to the work of Lončarić, Ryaben'kii, and Tsynkov [13].

Consider some function $u(x)$ that satisfies $Lu = 0$, where $x \in \Omega_i$ for a given i. Then the Green's formula yields

$$u(x) = \int_{\Gamma_i} \left(u \frac{\partial G}{\partial n} - \frac{\partial u}{\partial n} G \right) ds_y, \quad x \in \Omega_i. \tag{4.1}$$

Note than the direction of the normal n is fixed to always point outward from a given domain Ω_i. A generalized potential of Calderon's type with vector density $\xi_{\Gamma_i} = (\xi_0, \xi_1)$ specified on Γ_i is defined by the following formula:

$$P_{\Omega_i}\xi_{\Gamma_i}(x) = \int_{\Gamma_i} \left(\xi_0 \frac{\partial G}{\partial n} - \xi_1 G\right) ds_y, \quad x \in \Omega_i, \tag{4.2}$$

which is similar to (4.1) except that we do not require ahead of time that ξ_0 and ξ_1 in (4.2) be the boundary values of some function u that solves $Lu = 0$ on Ω_i and its normal derivative. With the help of (4.2), formula (4.1) can be rewritten as

$$u = P_{\Omega_i}\left(u, \frac{\partial u}{\partial n}\right)\Big|_{\Gamma_i}, \ x \in \Omega_i.$$

Next, for any sufficiently smooth function v specified on Ω_i, we define its vector trace on Γ_i as

$$Tr_i\, v = \left(v, \frac{\partial v}{\partial n}\right)\Big|_{\Gamma_i} \tag{4.3}$$

and then introduce the boundary operator as P_{Γ_i} as a combination of the potential P_{Ω_i} of (4.2) and trace Tr_i of (4.3):

$$P_{\Gamma_i}\xi_{\Gamma_i} = Tr_i\, P_{\Omega_i}\xi_{\Gamma_i}. \tag{4.4}$$

Note that the operator P_{Γ_i} is a projection, $P_{\Gamma_i}^2 = P_{\Gamma_i}$.

The previous construction can easily be changed from the use of surface integrals to that of volume integrals. Given a vector density $\xi_{\Gamma_i} = (\xi_0, \xi_1)$, we take a sufficiently smooth auxiliary function $w(x)$ that is compactly supported near Γ_i and such that

$$Tr_i\, w = \xi_{\Gamma_i}. \tag{4.5}$$

Then, the potential (4.2) can be redefined as follows:

$$\begin{aligned} P_{\Omega_i}\xi_{\Gamma_i}(x) &= w(x) - \int_{\Omega_i} GLw dy \\ &= \int_{\mathbb{R}^n \setminus \Omega_i} GLw dy, \ x \in \Omega_i. \end{aligned} \tag{4.6}$$

Note that $P_{\Omega_i}\xi_{\Gamma_i}(x)$ of (4.6) does not depend on the specific choice of $w(x)$ as long as condition (4.5) is satisfied. We can also define the exterior potential, $Q_{\mathbb{R}^n\setminus\Omega_i}\xi_{\Gamma_i}(x)$, $x \in \mathbb{R}^n\setminus\Omega_i$, for the complementary domain $\mathbb{R}^n\setminus\Omega_i$ as

$$\begin{aligned} Q_{\mathbb{R}^n\setminus\Omega_i}\xi_{\Gamma_i}(x) &= w(x) - \int_{\mathbb{R}^n\setminus\Omega_i} GLw dy \\ &= \int_{\Omega_i} GLw dy, \ x \in \mathbb{R}^n\setminus\Omega_i. \end{aligned} \tag{4.7}$$

The exterior projection operator Q_{Γ_i} will be given by

$$Q_{\Gamma_i}\xi_{\Gamma_i} = Tr_i\, Q_{\mathbb{R}^n\setminus\Omega_i}\xi_{\Gamma_i}. \tag{4.8}$$

Combining (4.2), (4.6), and (4.7), we obtain a scalar function defined on both Ω_i and $\mathbb{R}^n\backslash\Omega_i$:

$$\int_{\Gamma_i}\left(\xi_0\frac{\partial G}{\partial n}-\xi_1 G\right)ds_y=\begin{cases}P_{\Omega_i}\xi_{\Gamma_i}(x), & x\in\Omega_i,\\ -Q_{\mathbb{R}^n\backslash\Omega_i}\xi_{\Gamma_i}(x), & x\in\mathbb{R}^n\backslash\Omega_i.\end{cases}\tag{4.9}$$

As has already been seen, we can calculate each branch of (4.9) using volumetric integrals instead of surface integrals.

Now let $u=u_i^+ + u_i^-$, where u_i^+ originates inside its corresponding Ω_i and u_i^- originates from outside of Ω_i. That is, $u_i^- = u^- + \sum_{j\neq i} u_i^+$ is the entire incoming component for Ω_i. Also denote $\xi_{\Gamma_i}=(u,\frac{\partial u}{\partial n})|_{\Gamma_i}$ and

$$\xi_{\Gamma_i}^+=\left(u_i^+,\frac{\partial u_i^+}{\partial n}\right)\Big|_{\Gamma_i},$$

$$\xi_{\Gamma_i}^-=\left(u_i^-,\frac{\partial u_i^-}{\partial n}\right)\Big|_{\Gamma_i}.$$

According to formula (4.9) and definitions of the projections (4.4) and (4.8), we then have

$$\begin{aligned}P_{\Gamma_i}\xi_{\Gamma_i}&=\xi_{\Gamma_i}^-,\\ Q_{\Gamma_i}\xi_{\Gamma_i}&=\xi_{\Gamma_i}^+.\end{aligned}\tag{4.10}$$

Hence the sum of the two projections is the identity $P_{\Gamma_i}+Q_{\Gamma_i}=I$. Formula (4.10) renders the wave split. The space Ξ_{Γ_i} of all two-dimensional vector functions ξ_{Γ_i} is split into a direct sum of two subspaces: $\Xi_{\Gamma_i}=\Xi_{\Gamma_i}^+\oplus\Xi_{\Gamma_i}^-$, where $\Xi_{\Gamma_i}^-=ImP_{\Gamma_i}\equiv KerQ_{\Gamma_i}$ contains traces of all incoming waves and $\Xi_{\Gamma_i}^+=ImQ_{\Gamma_i}\equiv KerP_{\Gamma_i}$ contains traces of all outgoing waves. The split is done only on the boundary, and no knowledge of the wave sources is needed. Any function ξ_{Γ_i} is represented as $\xi_{\Gamma_i}^-+\xi_{\Gamma_i}^+$, where $\xi_{\Gamma_i}^-$ can be extended to Ω_i and $\xi_{\Gamma_i}^+$ can be extended to $\mathbb{R}^n\backslash\Omega_i$, as solutions of the homogeneous equation $Lu=0$. The extensions are given by the incoming and outgoing branches of the potential:

$$P_{\Omega_i}\xi_{\Gamma_i}=P_{\Omega_i}\xi_{\Gamma_i}^-=u_i^-,\quad x\in\Omega_i,$$

and

$$Q_{\mathbb{R}^n\backslash\Omega_i}\xi_{\Gamma_i}=Q_{\mathbb{R}^n\backslash\Omega_i}\xi_{\Gamma_i}^+=u_i^+,\quad x\in\mathbb{R}^n\backslash\Omega_i,$$

respectively. If a given ξ_{Γ_i} satisfies the boundary equation with projection

$$P_{\Gamma_i}\xi_{\Gamma_i}=\xi_{\Gamma_i},\tag{4.11}$$

then this function is the trace of some u_i^-. That is, it is extendible to Ω_i as a solution of $Lu=0$. In other words, those and only those ξ_{Γ_i} that are traces of solutions to the homogeneous equation $Lu=0$ on Ω_i satisfy the Calderon boundary equation (4.11). A reciprocal result holds for $Q_{\Gamma_i}\xi_{\Gamma_i}=\xi_{\Gamma_i}$.

Having defined the potentials and projections for individual domains Ω_i, we will now extend the definitions to the entire composite domain $\Omega = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N$. Denote $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \cdots \cup \Gamma_N$, and let ξ_Γ be a two-dimensional vector function on this composite boundary. The interior branch of the potential with the density ξ_Γ is defined similarly to (4.6):

$$P_\Omega \xi_\Gamma(x) = \int_{\mathbb{R}^n \setminus \Omega} GLw dy, \quad x \in \Omega,$$

where $w = w(x)$ is an auxiliary function that satisfies the interface conditions

$$Tr\, w = \xi_\Gamma \quad \Longleftrightarrow \quad \left\{ Tr_i\, w = \xi_\Gamma\big|_{\Gamma_i},\ i = 1, 2, \ldots, N \right\}$$

and the appropriate Sommerfeld condition (2.3a) or (2.3b) at infinity. Other than that, $w(x)$ may be arbitrary. Likewise, the exterior branch of the potential is given by

$$Q_{\mathbb{R}^n \setminus \Omega} \xi_\Gamma(x) = \int_\Omega GLw dy, \quad x \in \mathbb{R}^n \setminus \Omega.$$

Using definition (4.7), for the exterior region $\mathbb{R}^n \setminus \Omega$ we can write

$$Q_{\mathbb{R}^n \setminus \Omega} \xi_\Gamma(x) = \sum_{i=1}^N Q_{\mathbb{R}^n \setminus \Omega_i} \xi_{\Gamma_i}(x) = \sum_{i=1}^N u_i^+(x), \quad x \in \mathbb{R}^n \setminus \Omega, \tag{4.12}$$

whereas for the interior of Ω_i, $i = 1, 2, \ldots, N$, we have according to (4.6) and (4.7)

$$\begin{aligned} P_\Omega \xi_\Gamma(x) &= P_{\Omega_i} \xi_{\Gamma_i}(x) - \sum_{\substack{j=1 \\ j \neq i}}^N Q_{\mathbb{R}^n \setminus \Omega_j} \xi_{\Gamma_j}(x) \\ &= u_i^- - \sum_{\substack{j=1 \\ j \neq i}}^N u_j^+(x) = u^-(x), \quad x \in \Omega_i. \end{aligned} \tag{4.13}$$

In formula (4.13), u_i^- denotes the entire incoming field with respect to the domain Ω_i. In other words, u_i^- is composed of u^- and u_j^+ from all Ω_j except $j = i$.

The projections for composite domains are defined as traces of the potentials:

$$\begin{aligned} P_\Gamma \xi_\Gamma &= Tr\, P_\Omega \xi_\Gamma, \\ Q_\Gamma \xi_\Gamma &= Tr\, Q_{\mathbb{R}^n \setminus \Omega} \xi_\Gamma. \end{aligned} \tag{4.14}$$

They possess the same properties as the projections built previously for individual subdomains. Namely, $P_\Gamma + Q_\Gamma = I$, and the projections render the wave split at the interface Γ into incoming waves $\xi_\Gamma^- = P_\Gamma \xi_\Gamma$ and outgoing waves $\xi_\Gamma^+ = Q_\Gamma \xi_\Gamma$.

Now that we have defined the potentials and projections for individual subdomains and for the composite domain, we can once again obtain the controls for composite domains. First, we will investigate the simple case of fully eliminating the exterior noise inside $\Omega = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N$, i.e., eliminating the entire incoming

component of the acoustic field with respect to Ω. Our control function $g = g(x)$ is defined as

$$g(x) = -Lw|_{\mathbb{R}^n\setminus\Omega}, \qquad g(x) = 0|_{\Omega}, \tag{4.15}$$

giving the output $v = v(x)$ in the form

$$v(x) = \begin{cases} -P_\Omega \xi_\Gamma(x), & x \in \Omega, \\ -w(x) + Q_{\mathbb{R}^n\setminus\Omega}\xi_\Gamma(x), & x \in \mathbb{R}^n\setminus\Omega. \end{cases} \tag{4.16}$$

Hence we achieve the desired cancellation on Ω, because for $x \in \Omega$ according to (4.13) we have $v(x) = -P_\Omega\xi_\Gamma(x) = -u^-(x)$. As for the exterior region $\mathbb{R}^n\setminus\Omega$, formulae (4.12), (4.13), and (4.16) indicate that the controls basically duplicate the output of a given Ω_i and subsequently halve it as it enters another subdomain Ω_j.

Next, we will explore the operator interpretation of the selective cancellation for individual subdomains. As before, assume that the $N \times N$ communication matrix $\mathbf{M}$ is given that determines which regions are allowed to hear one another. If the entry m_{ij} of this matrix at the intersection of row i and column j is equal to zero, then Ω_i hears Ω_j; otherwise, if $m_{ij} = 1$, then Ω_i does not hear Ω_j. In doing so, no reciprocity is assumed; i.e., the matrix $\mathbf{M}$ is not necessarily symmetric. At the first stage of building the selective controls, we will modify the boundary trace ξ_Γ with the help of the matrix $\mathbf{M}$.

Let $u = u(x)$ be the overall acoustic field from all original sources, and let $\xi_\Gamma = Tr\, u$. Denote $\xi_{\Gamma_i} = \xi_\Gamma\big|_{\Gamma_i}$ and introduce

$$\tilde{\xi}_\Gamma \stackrel{\text{def}}{=} \left\{ \tilde{\xi}_{\Gamma_i},\ i = 1, 2, \ldots, N \,\middle|\, \tilde{\xi}_{\Gamma_i} = \xi_{\Gamma_i} + \sum_{\substack{j=1 \\ m_{ij}=1}}^{N} Tr_i\, Q_{\mathbb{R}^n\setminus\Omega_j}\xi_{\Gamma_j} \right\}. \tag{4.17}$$

At the second stage, we obtain the controls $\tilde{g}$ according to the same formula (4.15) as we used previously, but substituting a different auxiliary function $\tilde{w} = \tilde{w}(x)$. In addition to the appropriate Sommerfeld condition (2.3a) or (2.3b) at infinity, this new auxiliary function is supposed to satisfy an alternative interface condition at Γ:

$$Tr\, \tilde{w} = \tilde{\xi}_\Gamma, \tag{4.18}$$

where $\tilde{\xi}_\Gamma$ is defined by formula (4.17). The output of the control sources $\tilde{g}(x)$ on the domain $\Omega = \Omega_1 \cup \ldots \cup \Omega_N$ is given by the potential

$$\begin{aligned} v(x) =& -P_\Omega\tilde{\xi}_\Gamma(x) = -P_{\Omega_i}\tilde{\xi}_{\Gamma_i}(x) + \sum_{\substack{j=1 \\ j\neq i}}^{N} Q_{\mathbb{R}^n\setminus\Omega_j}\tilde{\xi}_{\Gamma_j} \\ =& -P_{\Omega_i}\xi_{\Gamma_i}(x) - \sum_{\substack{j=1 \\ m_{ij}=1}}^{N} Q_{\mathbb{R}^n\setminus\Omega_j}\xi_{\Gamma_j} + \sum_{\substack{j=1 \\ j\neq i}}^{N} Q_{\mathbb{R}^n\setminus\Omega_j}\tilde{\xi}_{\Gamma_j} \\ =& -u^- - \sum_{\substack{j=1 \\ j\neq i}}^{N} u_j^+ - \sum_{\substack{j=1 \\ m_{ij}=1}}^{N} u_j^+ + \sum_{\substack{j=1 \\ j\neq i}}^{N} u_j^+ \\ =& -u^- - \sum_{\substack{j=1 \\ m_{ij}=1}}^{N} u_j^+, \quad x \in \Omega_i, \end{aligned}$$

where we have taken into account that $P_{\Omega_i} Tr_i Q_{\mathbb{R}^n \setminus \Omega_j} \xi_{\Gamma_j} = Q_{\mathbb{R}^n \setminus \Omega_j} \xi_{\Gamma_j}$ for $x \in \Omega_i$ if $i \neq j$. Consequently, the overall field on Ω after applying the control $\tilde{g}$ is given by

$$\begin{aligned}\tilde{u}(x) &= u(x) + v(x) \\ &= u^-(x) + \sum_{j=1}^{N} u_j^+(x) - u^-(x) - \sum_{\substack{j=1 \\ m_{ij}=1}}^{N} u_j^+(x) \\ &= \sum_{\substack{j=1 \\ m_{ij}=0}}^{N} u_j^+(x), \quad x \in \Omega_i.\end{aligned}$$

In other words, the unwanted exterior noise $u^-(x)$ gets canceled out on all Ω_i, $i = 1, \ldots, N$, as before. Moreover, the sound field on a given Ω_i contains only the contributions from those Ω_j for which $m_{ij} = 0$, i.e., from those regions that Ω_i is allowed to hear. This is precisely the type of selective cancellation that we strived to achieve. Note also that even though we did not formulate the results in this section as theorems, it is clear that they are equivalent to the theorems of section 3.

5. A more realistic formulation. As of yet, we have only used the Calderon potentials and projections of section 4 to recast the results of section 3 in a more convenient yet equivalent operator form. However, the operator framework introduced in section 4 will also allow us to analyze a more elaborate formulation of the problem compared to that from section 3.

Instead of the Helmholtz equation (2.1), consider a general variable coefficient differential (or operator) equation

$$Lu = f, \tag{5.1}$$

where both the unknown solution $u = u(x)$ and the given right-hand side $f = f(x)$ are defined on some domain Ω_0 that may, but does not have to, coincide with the entire space $\mathbb{R}^n$. In the context of acoustics, (5.1) may, for example, govern the propagation of sound through a nonhomogeneous medium, where the propagation speed depends on the location.

A very important consideration is to define the solvability class for (5.1) on Ω_0. In most generic terms, let us require that $u \in U$, where U is a certain linear subspace of the space of all sufficiently smooth functions on Ω_0. We will assume that the solution $u = u(x)$ of (5.1) exists and is unique in U, provided that the right-hand side f belongs to another appropriate class F. Note that in the context of sections 2, 3, and 4, we had $\Omega_0 = \mathbb{R}^n$ and the class U was defined by the Sommerfeld condition (2.3a) or (2.3b) at infinity.

Since for any $f \in F$ there is a unique solution $u \in U$ of (5.1), we can introduce the inverse operator $G: F \longmapsto U$ that provides the solution for a given right-hand side:

$$u = Gf, \quad u \in U, \quad f \in F. \tag{5.2}$$

Note that previously (in the context of constant coefficients) the operator G was introduced by means of the convolution (2.4) with the fundamental solution (2.5) or (2.6). For variable coefficients, and/or when the domain Ω_0 is smaller than the entire

space $\mathbb{R}^n$, the apparatus of fundamental solutions does not apply. Yet the inverse operator G of (5.2) is well defined. In practice, it can be computed; i.e., problem (5.1) subject to the condition $u \in U$ can be discretized on Ω_0 and solved numerically.

Another very important consideration is the structure of the boundary trace that corresponds to the new operator L of (5.1). For the Laplace and Helmholtz operators, the vector traces on Γ are defined as traces of the solution itself and of the normal derivative; see formula (4.3). In the general theory of Calderon's operators (see [19]), the traces are constructed to guarantee a key property of the potentials (4.6), (4.7) and projections (4.4), (4.8), namely, their independence of the auxiliary function $w(x)$ as long as it has the correct trace, i.e., as long as the interface condition (4.5) is satisfied. For the second order variable coefficient operators L that have the form

$$Lv = \nabla(p\nabla v) + \{\text{lower order terms}\}, \quad p = p(x), \tag{5.3}$$

the Neumann data reduce to the standard normal derivative, and, consequently, the previous definition of the trace (see (4.3)) applies with no change. Hereafter, we will assume for simplicity that this is the case. This assumption does not entail a considerable loss of generality because operators (5.3) cover many important applications.

Having introduced the operator equation (5.1), defined the inverse (5.2), and identified the boundary trace Tr (4.3), we can extend all the operator constructions of section 4 in a straightforward manner, as done in [13] for a single domain. The only thing that will change is that every time a volumetric convolution with the fundamental solution appears in an equation, it ought to be replaced by the operator G of (5.2) applied to the corresponding source function. This way, we define the generalized Calderon potentials (cf. formulae (4.6) and (4.7))

$$P_{\Omega_i}\xi_{\Gamma_i}(x) = G\Big\{ Lw\big|_{\mathbb{R}^n\setminus\Omega_i} \Big\}, \quad x \in \Omega_i, \tag{5.4}$$

$$Q_{\mathbb{R}^n\setminus\Omega_i}\xi_{\Gamma_i}(x) = G\Big\{ Lw\big|_{\Omega_i} \Big\}, \quad x \in \mathbb{R}^n\setminus\Omega_i, \tag{5.5}$$

and the boundary projection operators (cf. formulae (4.4) and (4.8))

$$P_{\Gamma_i}\xi_{\Gamma_i} = Tr_i P_{\Omega_i}\xi_{\Gamma_i}, \tag{5.6}$$

$$Q_{\Gamma_i}\xi_{\Gamma_i} = Tr_i Q_{\mathbb{R}^n\setminus\Omega_i}\xi_{\Gamma_i} \tag{5.7}$$

for all $i = 1, 2, \ldots, N$. Combined operators for the composite domain $\Omega = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N$ are also introduced similarly to section 4, according to formulae (4.12), (4.13), and (4.14), where the individual operators are now given by (5.4)–(5.7).

The fundamental properties of the projections (5.6) and (5.7) are the same as before. Namely, the function $u \in U$ is a solution to the homogeneous equation $Lu = 0$ on the domain Ω_i if and only if its boundary trace $\xi_{\Gamma_i} = Tr_i u$ satisfies the boundary equation with projection,

$$P_{\Gamma_i}\xi_{\Gamma_i} = \xi_{\Gamma_i}. \tag{5.8}$$

Similarly, the function $u \in U$ is a solution to the homogeneous equation $Lu = 0$ on the complementary domain $\Omega_0 \setminus \Omega_i$ if and only if its boundary trace $\xi_{\Gamma_i} = Tr_i u$ satisfies the boundary equation with projection,

$$Q_{\Gamma_i}\xi_{\Gamma_i} = \xi_{\Gamma_i}. \tag{5.9}$$

Accordingly, if the solutions to (5.1) are interpreted as waves, then one can say that the boundary equations with projections (5.8) and (5.9) render the wave split into incoming and outgoing with respect to a given Ω_i. If $u \in U$ and $Tr_i u = \xi_{\Gamma_i}$, then

$$\xi_{\Gamma_i} = P_{\Gamma_i}\xi_{\Gamma_i} + Q_{\Gamma_i}\xi_{\Gamma_i} \stackrel{\text{def}}{=} \xi^-_{\Gamma_i} + \xi^+_{\Gamma_i},$$

where the component $\xi^-_{\Gamma_i}$ is the trace of the incoming field due to the sources outside Ω_i,

$$\xi^-_{\Gamma_i} = Tr_i u_i^-, \quad Lu_i^- = 0 \ \text{ for } \ x \in \Omega_i,$$

and the component $\xi^+_{\Gamma_i}$ is the trace of the outgoing field due to the sources inside Ω_i,

$$\xi^+_{\Gamma_i} = Tr_i u_i^+, \quad Lu_i^+ = 0 \ \text{ for } \ x \in \mathbb{R}^n \setminus \Omega_i.$$

In doing so, the entire space $\Xi_{\Gamma_i} = \{\xi_{\Gamma_i}\}$ can be represented as a direct sum of the traces of incoming waves and those of the outgoing waves:

$$\Xi_{\Gamma_i} = \Xi^-_{\Gamma_i} \oplus \Xi^+_{\Gamma_i}.$$

The exact same results automatically extend to the operators built for the composite domain $\Omega = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N$ as well.

Moreover, all the conclusions of sections 3 and 4 regarding the active control sources are also preserved. Namely, to cancel out the unwanted exterior sound u^- on $\Omega = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_N$, we build the controls according to formula (4.15):

$$g(x) = -Lw|_{\mathbb{R}^n \setminus \Omega}, \qquad g(x) = 0|_{\Omega}, \tag{5.10}$$

where the auxiliary function $w = w(x)$ satisfies $w \in U$ and $Trw = Tru$, and $u = u(x)$ is the overall acoustic field. We emphasize that in order to obtain the controls $g(x)$ of (5.10), we only need to know Tru at the boundary $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \cdots \cup \Gamma_N$. Moreover, the coefficients of the operator L, i.e., the properties of the medium, only need to be known outside Ω on the region where the auxiliary function $w(x) \neq 0$. This region can be a narrow layer outside Γ right next to it. This conclusion seems counterintuitive at first glance, because the controls $g(x)$ of (5.10) are supposed to eliminate the unwanted component of the field inside Ω, and yet it seems that the properties of the medium do not need to be known. The explanation, however, is quite simple. Both the unwanted noise $u^-(x)$ and the output of the controls $v(x) = Gg$ propagate across one and the same medium, and to achieve cancellation we do not necessarily need to know what this medium is inside Ω. Equivalently, one can think that the entire incoming component $u^-(x)$ is canceled by the controls (5.10) right at the entry to Ω so that it does not propagate any further; see [13, sections 4.2 and 4.3].

Active controls $\tilde{g}(x)$ that will render the selective cancellation of sound on the system of subdomains Ω_i, $i = 1, 2, \ldots, N$, according to a predetermined communication pattern $\mathbf{M}$ are also obtained with the help of formula (5.10). The only difference is that as before, the application of this formula requires a preliminary stage. At this preliminary stage, we construct a modified boundary trace $\tilde{\xi}_\Gamma$ according to formula (4.17), where the operators $Q_{\mathbb{R}^n \setminus \Omega_j}$ are defined by (5.5). At the final stage, we take an auxiliary function $\tilde{w} = \tilde{w}(x)$ that satisfies $\tilde{w} \in U$ and $Tr\tilde{w} = \tilde{\xi}_\Gamma$ and substitute it into (5.10), thus obtaining the desired selective controls $\tilde{g}(x)$.

For more detail on the theory of generalized Calderon potentials and projections, as well as their efficient computation by means of the method of difference potentials, we refer the reader to the monograph [19].

6. Conclusions. We have introduced and studied the problem of active control of sound for composite regions. This problem is, in fact, a particular inverse source problem for the differential equation (or system) that governs the sound field. Allowing for composite domains is a key innovation proposed here as compared to our previous work on the subject (see [13] and related references). We obtained a closed form general solution for the control sources. This solution allows all individual subdomains to either communicate freely with one another or else be shielded from their peers. In doing so, no reciprocity is assumed; i.e., for a given pair of subdomains one may be allowed to hear the other but not necessarily vice versa.

If the controls in the composite case are built exactly as in the previously analyzed case of simple, i.e., arcwise connected, domains, then the communications between all subdomains is allowed. In other words, by default all subdomains hear one another. If, however, a particular subdomain is not allowed to hear another given subdomain, then the supplementary controls are employed prior to building the final set of controls. The role of the supplementary controls (one can call it the predictor stage) is to communicate the specific acoustic output of the domain not to be heard to the domain that is not allowed to hear it. Subsequently, the final controls (corrector stage) use these data to render the desired sound cancellation.

Moreover, the general solution requires no information on the original acoustic sources and can be constructed based solely on the knowledge of the field quantities at the boundaries of the subdomains. In practice, those quantities can be obtained by measurements. In doing so, the methodology guarantees the exact volumetric cancellation of the unwanted noise, as opposed to many other techniques available in the literature that would only provide for a pointwise or directional cancellation, and would not even offer an approach to selective cancellation on composite domains.

The problem is solved for a general formulation that allows the propagation of sound across a medium with variable characteristics. In doing so, to cancel out the outside sound on a given domain, no actual knowledge of the medium properties on this domain is required. The explanation of this seemingly counterintuitive behavior is simple—both the original sound and the output of the controls propagate across one and the same medium, and for building the control sources we do not necessarily need to know what this medium is.

It is also important to mention that for every subdomain there is a component of the acoustic field to be canceled out and another component to be left unaffected. Yet the quantities at the boundary that need to be measured in order to build the control system can pertain to the overall field rather than only to its unwanted component, and the methodology will automatically distinguish between the two. Of course, the locations and shapes of the subdomains need to be known ahead of time.

Finally, it is clear that in the context of implementation, obtaining the continuous data, as well as providing a continuous excitation (control sources), along the interface Γ is not practical. Instead, the problem needs to be discretized so that only finite arrays of individual sensors (microphones) and actuators (loudspeakers) are used. A powerful apparatus for the analysis of discrete active shielding problems is provided by the method of difference potentials [19]. This method offers a comprehensive finite-difference theory, which is fully analogous to the continuous theory of Calderon's operators [3, 23] and in many instances even goes beyond it. As mentioned in section 1, discrete active controls have been built, and their properties established, for various settings; see [14, 15, 16, 18, 22, 25, 27, 28]. In particular, the case of a composite region in the discrete framework is analyzed in [21]. A brief account of the method of difference potentials, along with the analysis of discrete active shielding problems, can be found in [20, Chapter 14].

REFERENCES

[1] J. C. Burgess, *Active adaptive sound control in a duct: A computer simulation*, J. Acoust. Soc. Amer., 70 (1981), pp. 715–726.

[2] R. H. Cabell and C. R. Fuller, *Active control of periodic disturbances using principal component LMS: Theory and experiment*, in Proceedings of the 3rd AST/HSR Interior Noise Workshop, Part I: Sessions A, B, and C, NASA Langley Research Center, Hampton, VA, 1998.

[3] A. P. Calderon, *Boundary-value problems for elliptic equations*, in Proceedings of the Soviet-American Conference on Partial Differential Equations, Fizmatgiz, Novosibirsk, Moscow, 1963, pp. 303–304.

[4] R. L. Clark, W. R. Saunders, and G. P. Gibbs, *Adaptive Structures: Dynamics and Control*, John Wiley and Sons, New York, 1998.

[5] A. J. Devaney and G. C. Sherman, *Nonuniqueness in inverse source and scattering problems*, IEEE Trans. Antennas and Propagation, 30 (1982), pp. 1034–1037; see also pp. 1037–1042.

[6] A. J. Devaney and E. Wolf, *Radiating and nonradiating classical current distributions and the fields they generate*, Phys. Rev. D, 8 (1973), pp. 1044–1047.

[7] S. J. Elliott, I. M. Stothers, and P. A. Nelson, *A multiple error LMS algorithm and its application to the active control of sound and vibration*, in IEEE Trans. Acoust., Speech, Signal Processing, ASSP-35 (1987), pp. 1423–1434.

[8] C. R. Fuller, S. J. Elliott, and P. A. Nelson, *Active Control of Vibration*, Academic Press, London, 1996.

[9] C. R. Fuller and A. H. von Flotow, *Active control of sound and vibration*, IEEE Control Syst. Mag., 15 (1995), pp. 9–19.

[10] V. Isakov, *Inverse Source Problems*, Mathematical Surveys Monogr. 34, American Mathematical Society, Providence, RI, 1990.

[11] R. K. Kincaid and K. Laba, *Reactive TABU search and sensor selection in active structural control problems*, J. Heuristics, 4 (1998), pp. 199–220.

[12] R. K. Kincaid, K. Laba, and S. L. Padula, *Quelling cabin noise in turboprop aircraft via active control*, J. Comb. Optim., 1 (1997), pp. 229–250.

[13] J. Lončarić, V. S. Ryaben'kii, and S. V. Tsynkov, *Active shielding and control of noise*, SIAM J. Appl. Math., 62 (2001), pp. 563–596.

[14] J. Lončarić and S. V. Tsynkov, *Optimization of acoustic source strength in the problems of active noise control*, SIAM J. Appl. Math., 63 (2003), pp. 1141–1183.

[15] J. Lončarić and S. V. Tsynkov, *Optimization of power in the problems of active control of sound*, Math. Comput. Simulation, 65 (2004), pp. 323–335.

[16] J. Lončarić and S. V. Tsynkov, *Quadratic optimization in the problems of active control of sound*, Appl. Numer. Math., 52 (2005), pp. 381–400.

[17] P. A. Nelson and S. J. Elliott, *Active Control of Sound*, Academic Press, San Diego, 1999.

[18] V. S. Ryaben'kii, *A difference screening problem*, Funct. Anal. Appl., 29 (1995), pp. 70–71.

[19] V. S. Ryaben'kii, *Method of Difference Potentials and Its Applications*, Springer Ser. Comput. Math. 30, Springer-Verlag, Berlin, 2002.

[20] V. S. Ryaben'kii and S. V. Tsynkov, *A Theoretical Introduction to Numerical Analysis*, Chapman & Hall/CRC, Boca Raton, FL, 2007.

[21] V. S. Ryaben'kii, S. V. Tsynkov, and S. V. Utyuzhnikov, *Inverse source problem and active shielding for composite domains*, Appl. Math. Lett., 20 (2007), pp. 511–516.

[22] V. S. Ryaben'kii and S. V. Utyuzhnikov, *Differential and finite-difference problems of active shielding*, Appl. Numer. Math., 57 (2007), pp. 374–382.

[23] R. T. Seeley, *Singular integrals and boundary value problems*, Amer. J. Math., 88 (1966), pp. 781–809.

[24] M. O. Tokhi and S. M. Veres, eds., *Active Sound and Vibration Control: Theory and Applications*, IEE Control Ser. 62, The Institution of Electrical Engineers, London, 2002.

[25] S. V. Tsynkov, *On the definition of surface potentials for finite-difference operators*, J. Sci. Comput., 18 (2003), pp. 155–189.

[26] H. Van der Auweraer, M. Iadevaia, U. Emborg, and M. Gustavsson, *Derivation of experimental vibro-acoustical models for ANC configuration design*, AIAA Paper 97-1618, in Proceedings of the 3rd AIAA/CEAS Aeroacoustics Conference, Atlanta, GA, 1997, pp. 222–234.

[27] R. I. Veizman and V. S. Ryaben'kii, *Difference problems of screening and simulation*, Dokl. Akad. Nauk, 354 (1997), pp. 151–154.

[28] R. I. Veizman and V. S. Ryaben'kii, *Difference simulation problems*, Trans. Moscow Math. Soc., 58 (1997), pp. 239–248.

[29] B. Widrow, D. Shur, and S. Shaffer, *On adaptive inverse control*, in Proceedings of the 15th IEEE Asilomar Conference on Circuits, Systems and Computers, Pacific Grove, CA, 1981, pp. 185–189.

[30] S. E. Wright and B. Vuksanovic, *Active control of environmental noise*, J. Sound Vibration, 190 (1996), pp. 565–585.

[31] S. E. Wright and B. Vuksanovic, *Active control of environmental noise*, II: *Non-compact acoustic sources*, J. Sound Vibration, 202 (1997), pp. 313–359.

SIAM J. APPL. MATH.
Vol. 67, No. 6, pp. 1610–1632

ON THE UPLINK OF A CELLULAR SYSTEM WITH IMPERFECT POWER CONTROL AND MULTIPLE SERVICES*

JOHN A. MORRISON†‡ AND PHIL WHITING†

Abstract. We analyze the reverse link of a single wireless cell in which mobile phones simultaneously transmit to the base station using code division multiple access (CDMA). The mobiles are transmitting data which is delay intolerant, so that scheduling cannot be employed. There is a finite number of data classes, and the users transmit either data or a lower rate synchronizing signal. To overcome the near-far problem, received power control is used, which has a log-normal error. For each class there is an outage probability that the user's signal-to-noise ratio (SNR) will be met (when active and when idle). Refinements of the central limit theorem are used to determine the number of users of each class that can be supported, i.e., the capacity. The approximation can also be used to determine the minimal target powers necessary to meet the outage requirements. Comparison with simulation shows these approximations to be accurate.

Key words. asymptotics, capacity region, central limit approximation, code division multiple access, log-normal errors, minimal target powers, received power control, uplink, wireless

AMS subject classifications. 60K37, 90B15

DOI. 10.1137/050648316

1. Introduction.

This paper considers the capacity of the reverse link of a cellular code division multiple access (CDMA) system in which there are multiple circuit-switched data connections, and the minimal target powers when the link is operating within its capacity. Each data connection is drawn from one of a small number of data classes, with its own quality of service (QoS) requirements. The data in all classes is delay *intolerant*, so that scheduling of the data is infeasible. Furthermore, while users are connected, data is transmitted periodically to the base station in "bursts"; otherwise they are "idle," and a synchronization signal is transmitted. Active transmissions are made up of a series of channel encoded frames. The base station makes no attempt to coordinate user transmissions, and so frames will be undecodable when there is excessive interference from other user transmissions. Interference is generated both when users are active and when they are idle.

One way to manage this interference is to endeavor to directly control the received signal-to-noise-ratio (SNR) of the user frames by power control [15],[16]. However, since data transmissions are bursty, of short duration, and difficult to anticipate, management of *user SNR* by such a power control may be difficult to accomplish. Instead we propose a power control with the simpler objective of maintaining received power only. The interference then fluctuates according to the activity of the users themselves and the target powers. Clearly the target powers should depend on the number of users of each class which are connected at any point as well as their QoS. However it is not a priori clear how to set these targets. One possibility is to set them using stochastic approximation, raising and lowering the targets according to user performance. This is not our approach. Instead we propose to set the targets

*Received by the editors December 23, 2005; accepted for publication (in revised form) May 8, 2007; published electronically September 12, 2007.
http://www.siam.org/journals/siap/67-6/64831.html
†Alcatel-Lucent, Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974 (johnmorrison@alcatel-lucent.com, pawhiting@alcatel-lucent.com).
‡Consultant.

directly using the statistics of activity for each class, which we suppose known. (In some implementations of the system it may be possible to actually control the degree of activity and so make the usual trade-offs between throughput and the number of users supported.)

Users set their transmit powers in conjunction with a pilot signal measured by the mobile receivers. The base station then compares the user's received power with its target and sends corrections to the mobile phone (henceforward referred to as a mobile) via a feedback link. The feedback link increases the accuracy of the power control. A distributed algorithm for perfect power control, and its convergence, were investigated in [12]. Ideal power control (very small error in the received power) is of course difficult to achieve, however, and we suppose that there is a nonnegligible residual error. The marginal distribution of this error at any instant we take to be log-normal, and the underlying standard deviation is supposed known. It will be convenient to quote relative powers in decibels (dB), which is $10\log_{10}$.

The QoS requirements specify the fraction of packets that is allowed to be undecodable at the receiver. We propose to control this fraction by maintaining the received signals so that the instantaneous probability of an outage is small. An outage occurs when the SNR ratio falls below a predetermined threshold, which we also suppose known. Our approach is thus related to the one described in [3] where outage probabilities are determined by modeling the distribution of users as a spatial Poisson process.

Each class i is thus characterized by requirements both when active and idle and which are in general distinct. These are the users' rate, the SNR thresholds $\alpha_i/W, \beta_i/W$, and the outage probabilities L_i, l_i. Here W is the spread bandwidth. Additionally we suppose that the probability that a user is active is w_i. The standard deviation of power control error may also be taken to depend on whether the user is active as well as on the user's data class. In what follows it will be convenient to work with the bit-energy-to-interference density ratio E_b/I_0, which determines the performance of the base station decoder and is usually quoted in the design of wireless communication links. E_b/I_0 is related to the SNR threshold via the processing gain to be $E_\mathrm{b}/I_0 = \alpha_i/R_i$, where R_i is the class i active bit rate and a similar relationship when the user is idle.

To estimate the number of users that can be supported and determine the minimal target powers when the link is operating within its capacity, we suppose that the system is large-scale and use a central limit approximation. Put crudely, we are relying on the approximately normal behavior of the interference from users in the same cell. The statistics of this interference is a combination of random activity and the user power control errors, as already mentioned. In order to have a satisfactory approximation, it will turn out that estimates involving the density of these statistics will be needed and not just their distribution. Indeed we are led to a model in which the scaled interference converges to a fixed quantity plus a normal error with scale $1/\sqrt{K}$, where K is a scale factor for the number of users in the system. The interference from users in adjacent cells is modeled along the lines in [3], [4], with allowance for skewness and kurtosis.

Before continuing, we would like to discuss some other related references. First it should be noted that the model used in our paper is very similar to Model 2 described in [1] and that we also address similar performance questions. We refer readers to [1] but note that both papers are concerned with received power control and corresponding received power targets, both have more than one class of users, and both are concerned with outage, referred to as the service availability probability

(SAP) in [1]; see our (2.8) and (2.9). Both models also consider imperfect power control with a marginal log-normal error. However, the methods of analysis are very different, with the imperfect power control (SAP) in [1] being approximated using an estimate given in [8], whereas our results are obtained via refined asymptotics based on the central limit theorem, as mentioned earlier. Finally, our results lead to the determination of minimum feasible received power targets themselves as well as explicit constraints for the capacity, as in Propositions 2.1 and 2.2.

Next we would like to mention other papers related to our work. Results for a single cell, two service class system were obtained in [13] where in this case resources were allocated according to a mechanism of dynamically adapting the spreading gain (chip-to-data bit ratio). Second there is the well-known paper [6] which is for a pure voice system but does not consider power control errors. (The influence of power control error on reverse link capacity was the subject of detailed investigation in subsequent studies; see, for example, [2] and [14].) Additionally, in our system, load control might be affected by active rate control (via the vocoder or similar controls for streaming video and other near-real-time applications). A related idea of adaptive control of error correction codes, power, and scheduling of users is considered in [10] for the downlink of a CDMA system. Finally, algorithms for joint load balancing and cell assignment were treated by Hanly [7] as well as Yates and Huang [15], which were indeed devised initially for the reverse link of CDMA networks.

The model is formulated in section 2, and the asymptotic results are stated there. The asymptotic analysis is carried out in section 3, under the assumption that the system is large-scale. The capacity of the system, in terms of the number of mobiles requiring each service, is determined to lowest order in section 4, and correction terms are determined in section 5. Asymptotic approximations to the minimal powers are also determined. Numerical results are presented in section 6. These results are used to illustrate the analysis in the relatively simple setting of distance-based path loss laws. Subsequently, there are detailed discussions as to how to incorporate more involved propagation models with log-normal shadowing as well as how to approximate the multicell interference where the control is being used at each cell in a network. Conclusions are presented in section 7.

2. Model and results. Because of the imperfect power control, we assume that the power received at the base station from an active mobile m of class j is $P_j e_m^{(j)}$, where P_j is the target power, and $e_m^{(j)} = e^{\kappa_j \xi_m^{(j)}}$, where $\xi_m^{(j)}$ is normally distributed, with zero mean and unit variance. Analogously, the power received at the base station from an idle mobile of class j is $p_j \epsilon_m^{(j)}$, where $\epsilon_m^{(j)} = e^{\sigma_j \eta_m^{(j)}}$ and $\eta_m^{(j)}$ is normally distributed, with zero mean and unit variance.

Additionally, let $X_m^{(j)}$ be the activity indicator for mobile m of class j. We assume that

$$w_j = \Pr\left\{X_m^{(j)} = 1\right\} = 1 - \Pr\left\{X_m^{(j)} = 0\right\}, \quad m \in \mathcal{M}_j,\ j = 1, \dots, J, \tag{2.1}$$

where $\mathcal{M}_j$ denotes the set of mobiles of class j. We also assume that the random variables $(\xi_m^{(j)}, \eta_m^{(j)})$ and $X_m^{(j)}$, $m \in \mathcal{M}_j$, $j = 1, \dots, J$, are mutually independent, but we allow for possible correlation between $\xi_m^{(j)}$ and $\eta_m^{(j)}$.

The power $I^{(j)}$ received at the base station due to all the mobiles of class j in the cell under consideration is

$$I^{(j)} = \sum_{m \in \mathcal{M}_j} \left[P_j e_m^{(j)} X_m^{(j)} + p_j \epsilon_m^{(j)} (1 - X_m^{(j)}) \right]. \tag{2.2}$$

The total power I_0 received at the base station, including that from mobiles in other cells, is taken to be

$$I_0 = \sum_{l=1}^{J} I^{(l)} + \eta W + MK_0 + \sqrt{vK_0}\, S\,. \tag{2.3}$$

Here η, M, and v are positive constants, and ηW is the local receiver noise power. Since we are considering a wideband system, W is large. Also, $K_0 \gg 1$ is the total number of users in adjacent cells and, motivated by the results of Chan and Hanly [3], we assume that the density of the random variable S has a truncated Edgeworth-like expansion of the form

$$p(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \left[1 + \frac{\mu_3}{6} He_3(y) + \frac{\mu_4}{24} He_4(y) + \frac{\mu_3^2}{72} He_6(y)\right], \tag{2.4}$$

where μ_3 and μ_4 are $O(1)$ constants and, in the notation of [11],

$$He_n(y) = (-1)^n e^{y^2/2} \frac{d^n}{dy^n}(e^{-y^2/2}) \tag{2.5}$$

are Hermite polynomials. The mean MK_0 and variance vK_0 of $MK_0 + \sqrt{vK_0}S$ are both $O(K_0)$, and μ_3 is the skewness of S.

We define

$$J_m^{(j)} = \sum_{m' \neq m} \left[P_j e_{m'}^{(j)} X_{m'}^{(j)} + p_j \epsilon_{m'}^{(j)} (1 - X_{m'}^{(j)})\right], \quad m \in \mathcal{M}_j\,. \tag{2.6}$$

Then the interference $I_m^{(j)}$ to the transmission of any mobile of class j is

$$I_m^{(j)} = \sum_{l \neq j} I^{(l)} + J_m^{(j)} + \eta W + MK_0 + \sqrt{vK_0}\, S\,. \tag{2.7}$$

The SNR requirement for an active mobile is

$$\Pr\left\{P_j e_m^{(j)} \geq \frac{\alpha_j}{W} I_m^{(j)}\right\} \geq 1 - L_j, \quad m \in \mathcal{M}_j,\ j = 1, \ldots, J\,, \tag{2.8}$$

and the SNR requirement for an idle mobile is

$$\Pr\left\{p_j \epsilon_m^{(j)} \geq \frac{\beta_j}{W} I_m^{(j)}\right\} \geq 1 - l_j, \quad m \in \mathcal{M}_j,\ j = 1, \ldots, J\,, \tag{2.9}$$

where L_j and l_j are prescribed loss probabilities.

We define the outage quantiles ν_j and ρ_j by

$$1 - L_j = \frac{1}{\sqrt{2\pi}} \int_{-\nu_j}^{\infty} e^{-z^2/2} dz \tag{2.10}$$

and

$$1 - l_j = \frac{1}{\sqrt{2\pi}} \int_{-\rho_j}^{\infty} e^{-x^2/2} dz\,. \tag{2.11}$$

We assume that the target powers are limited, with

$$P_j \le \bar{P}_j, \quad p_j \le \bar{p}_j, \quad j = 1, \dots, J\,, \tag{2.12}$$

and we define

$$\delta_0 = \max_j \max \left(\frac{\alpha_j}{\bar{P}_j} e^{\kappa_j \nu_j}, \ \frac{\beta_j}{\bar{p}_j} e^{\sigma_j \rho_j} \right). \tag{2.13}$$

We suppose that K_j is the large number of mobiles belonging to class j, and we define

$$\tau = \frac{1}{W} \sum_{j=1}^{J} [\alpha_j e^{\kappa_j \nu_j} e^{\kappa_j^2/2} w_j + \beta_j e^{\sigma_j \rho_j} e^{\sigma_j^2/2} (1 - w_j)] K_j. \tag{2.14}$$

We first derive the following asymptotic result.

PROPOSITION 2.1. *The first order approximation to the admissible set, which gives a constraint on the number of users of each class that can be supported, is given by*

$$\tau \le 1 - \left(\eta + \frac{MK_0}{W} \right) \delta_0\,. \tag{2.15}$$

The corresponding minimal target powers necessary to meet the outage requirements are

$$P_j^* = \frac{\alpha_j e^{\kappa_j \nu_j}}{(1-\tau)} \left(\eta + \frac{MK_0}{W} \right), \qquad p_j^* = \frac{\beta_j e^{\sigma_j \rho_j}}{(1-\tau)} \left(\eta + \frac{MK_0}{W} \right). \tag{2.16}$$

We remark that, to this order, the approximation does not involve the variance or skewness of the interference from users in adjacent cells. We derive a refined asymptotic approximation which does depend on these quantities. We define

$$f_j = \alpha_j e^{\kappa_j \nu_j} e^{\kappa_j^2/2} w_j + \beta_j e^{\sigma_j \rho_j} e^{\sigma_j^2/2} (1 - w_j)\,, \tag{2.17}$$

$$\psi_l = \alpha_l^2 e^{2\kappa_l \nu_l} e^{2\kappa_l^2} w_l + \beta_l^2 e^{2\sigma_l \rho_l} e^{2\sigma_l^2} (1 - w_l) - f_l^2\,, \tag{2.18}$$

$$\psi = vK_0 W \left[\frac{(1-\tau)}{(\eta W + MK_0)} \right]^2, \tag{2.19}$$

and

$$\omega_j = \frac{1}{2} \left(\frac{\nu_j}{\kappa_j} - 1 \right) \left(\frac{1}{W} \sum_{l=1}^{J} \psi_l K_l + \psi \right) - f_j, \tag{2.20}$$

$$\zeta_j = \frac{1}{2} \left(\frac{\rho_j}{\sigma_j} - 1 \right) \left(\frac{1}{W} \sum_{l=1}^{J} \psi_l K_l + \psi \right) - f_j,$$

and

$$m_j = \frac{1}{6} \mu_3 \psi^{3/2} \left[\left(\frac{\nu_j}{\kappa_j} - 1 \right) \left(\frac{\nu_j}{\kappa_j} - 2 \right) - \frac{1}{\kappa_j^2} \right], \tag{2.21}$$

$$q_j = \frac{1}{6} \mu_3 \psi^{3/2} \left[\left(\frac{\rho_j}{\sigma_j} - 1) \right) \left(\frac{\rho_j}{\sigma_j} - 2 \right) - \frac{1}{\sigma_j^2} \right].$$

Also, we let

$$X = \frac{1}{W}\sum_{j=1}^{J}[\alpha_j e^{\kappa_j\nu_j}e^{\kappa_j^2/2}w_j\omega_j + \beta_j e^{\sigma_j\rho_j}e^{\sigma_j^2/2}(1-w_j)\zeta_j]K_j\,, \tag{2.22}$$

$$U = \frac{1}{W}\sum_{j=1}^{J}[\alpha_j e^{\kappa_j\nu_j}e^{\kappa_j^2/2}w_j m_j + \beta_j e^{\sigma_j\rho_j}e^{\sigma_j^2/2}(1-w_j)q_j]K_j\,, \tag{2.23}$$

and

$$\delta = \max_j \max\left[\frac{\alpha_j}{\bar{P}_j}e^{\kappa_j\nu_j}\left(1+\frac{\omega_j}{W}+\frac{m_j}{W^{3/2}}\right),\ \frac{\beta_j}{\bar{p}_j}e^{\sigma_j\rho_j}\left(1+\frac{\zeta_j}{W}+\frac{q_j}{W^{3/2}}\right)\right]. \tag{2.24}$$

We then have the following asymptotic results.

PROPOSITION 2.2. *The refined approximation to the admissible set is given by*

$$\tau + \frac{X}{W} + \frac{U}{W^{3/2}} \le 1 - \left(\eta + \frac{MK_0}{W}\right)\delta\,, \tag{2.25}$$

and the corresponding minimal target powers are

$$P_j^* = \frac{\alpha_j e^{\kappa_j\nu_j}\left(\eta+\frac{MK_0}{W}\right)\left(1+\frac{\omega_j}{W}+\frac{m_j}{W^{3/2}}\right)}{\left(1-\tau-\frac{X}{W}-\frac{U}{W^{3/2}}\right)}\,, \tag{2.26}$$

$$p_j^* = \frac{\beta_j e^{\sigma_j\rho_j}\left(\eta+\frac{MK_0}{W}\right)\left(1+\frac{\zeta_j}{W}+\frac{q_j}{W^{3/2}}\right)}{\left(1-\tau-\frac{X}{W}-\frac{U}{W^{3/2}}\right)}\,.$$

We remark that, to this order, which neglects terms of $O(1/W^2)$, the kurtosis μ_4 in (2.4) *does not appear.*

3. Asymptotic analysis. For $m \in \mathcal{M}_j$, the error terms have first and second moments

$$E(e_m^{(j)}) = e^{\kappa_j^2/2}, \qquad E\left[(e_m^{(j)})^2\right] = e^{2\kappa_j^2} \tag{3.1}$$

and

$$E(\epsilon_m^{(j)}) = e^{\sigma_j^2/2}, \qquad E\left[(\epsilon_m^{(j)})^2\right] = e^{2\sigma_j^2}\,. \tag{3.2}$$

From (2.1), since we have assumed that the activity and error random variables are mutually independent,

$$E\left[P_j e_m^{(j)}X_m^{(j)} + p_j\epsilon_m^{(j)}(1-X_m^{(j)})\right] = P_j e^{\kappa_j^2/2}w_j + p_j e^{\sigma_j^2/2}(1-w_j) \tag{3.3}$$

and

$$E\left\{\left[P_j e_m^{(j)}X_m^{(j)} + p_j\epsilon_m^{(j)}(1-X_m^{(j)})\right]^2\right\} = P_j^2 e^{2\kappa_j^2}w_j + p_j^2 e^{2\sigma_j^2}(1-w_j)\,. \tag{3.4}$$

Hence,

$$\begin{aligned} v_j &\equiv \mathrm{Var}\left[P_j e_m^{(j)}X_m^{(j)} + p_j\epsilon_m^{(j)}(1-X_m^{(j)})\right] \\ &= P_j^2 e^{2\kappa_j^2}w_j + p_j^2 e^{2\sigma_j^2}(1-w_j) - \left[P_j e^{\kappa_j^2/2}w_j + p_j e^{\sigma_j^2/2}(1-w_j)\right]^2. \end{aligned} \tag{3.5}$$

Suppose that K_j is the number of mobiles belonging to class j. From (2.2) and (2.6), since $P_j e_m^{(j)} X_m^{(j)} + p_j \epsilon_m^{(j)}(1 - X_m^{(j)})$, $m \in \mathcal{M}_j$, are independently and identically distributed random variables with finite third moment,

$$I^{(j)} = \left[P_j e^{\kappa_j^2/2} w_j + p_j e^{\sigma_j^2/2}(1 - w_j)\right] K_j + \sqrt{v_j K_j} S^{(j)} \tag{3.6}$$

and

$$J_m^{(j)} = \left[P_j e^{\kappa_j^2/2} w_j + p_j e^{\sigma_j^2/2}(1 - w_j)\right] (K_j - 1) + \sqrt{v_j (K_j - 1)} S_m^{(j)}\,, \tag{3.7}$$

where $S^{(j)}$ and $S_m^{(j)}$ are asymptotically normally distributed with zero mean, unit variance, and error $O(1/\sqrt{K_j})$ as $K_j \to \infty$; see [5, p. 539]. This is an estimate for convergence in distribution, but it is shown in Appendix A that the densities converge as well. The random variables $S^{(l)}$, $l \neq j$, and $S_m^{(j)}$ are mutually independent. We now take the following asymptotic scalings:

$$\eta W = N_0 K,\ K_j = \gamma_j K, \quad j = 0, \ldots, J\,, \tag{3.8}$$

where $N_0 = 0(1)$ in appropriate power units and $\gamma_j = O(1)$ as $K, W \to \infty$. We take $K = \min_j K_j$. If we introduce these scalings into the expression (2.7) for $I_m^{(j)}$, use (3.6) and (3.7), and let

$$N = N_0 + M\gamma_0, \quad V = v\gamma_0, \tag{3.9}$$

we obtain

$$\begin{aligned} \frac{I_m^{(j)}}{K} = {} & \sum_{l=1}^{J} \left[P_l e^{\kappa_l^2/2} w_l + p_l e^{\sigma_l^2/2}(1 - w_l)\right] \gamma_l + N + \sqrt{\frac{V}{K}}\, S \\ & + \frac{1}{\sqrt{K}} \left[\sum_{l \neq j} \sqrt{v_l \gamma_l} S^{(l)} + \sqrt{v_j \left(\gamma_j - \frac{1}{K}\right)} S_m^{(j)}\right] \\ & - \frac{1}{K} \left[P_j e^{\kappa_j^2/2} w_j + p_j e^{\sigma_j^2/2}(1 - w_j)\right]\,. \end{aligned} \tag{3.10}$$

We define

$$G = \sum_{l=1}^{J} \left[P_l e^{\kappa_l^2/2} w_l + p_l e^{\sigma_l^2/2}(1 - w_l)\right] \gamma_l + N\,, \tag{3.11}$$

$$\lambda_j = \frac{P_j}{G}, \quad \theta_j = \frac{p_j}{G}, \quad \Phi = \frac{V}{G^2}\,, \tag{3.12}$$

$$\begin{aligned} \Gamma_j &= \lambda_j e^{\kappa_j^2/2} w_j + \theta_j e^{\sigma_j^2/2}(1 - w_j)\,, \\ \Phi_l &= \lambda_l^2 e^{2\kappa_l^2} w_l + \theta_l^2 e^{2\sigma_l^2}(1 - w_l) - \Gamma_l^2\,, \end{aligned} \tag{3.13}$$

and

$$\Lambda_m^{(j)} = \sum_{l \neq j} \sqrt{\Phi_l \gamma_l} S^{(l)} + \sqrt{\Phi_j \left(\gamma_j - \frac{1}{K}\right)} S_m^{(j)} + \sqrt{\Phi}\, S\,. \tag{3.14}$$

We also introduce the scalings

$$\frac{\alpha_j}{W} = \frac{a_j}{K}, \quad \frac{\beta_j}{W} = \frac{b_j}{K}, \quad j = 1, \ldots, J, \tag{3.15}$$

where $a_j = O(1)$ and $b_j = O(1)$ as $K, W \to \infty$. Then, from (3.5) and (3.10)–(3.15), the inequalities $P_j e_m^{(j)} \geq (\alpha_j/W) I_m^{(j)}$ and $p_j \epsilon_m^{(j)} \geq (\beta_j/W) I_m^{(j)}$ imply that, for $j = 1, \ldots, J$,

$$\frac{\lambda_j}{a_j} e^{\kappa_j \xi_m^{(j)}} \geq 1 + \frac{\Lambda_m^{(j)}}{\sqrt{K}} - \frac{\Gamma_j}{K}, \qquad \frac{\theta_j}{b_j} e^{\sigma_j \eta_m^{(j)}} \geq 1 + \frac{\Lambda_m^{(j)}}{\sqrt{K}} - \frac{\Gamma_j}{K}\,. \tag{3.16}$$

In the next section we consider the implications of these stochastic inequalities.

4. Capacity and minimal powers. We assume that $\kappa_j \geq \kappa > 0$ and $\sigma_j \geq \sigma > 0$, $j = 1, \ldots, J$, and that κ and σ are not small. We define

$$Y_m^{(j)} = \frac{\lambda_j}{a_j} e^{\kappa_j \xi_m^{(j)}} - \frac{\Lambda_m^{(j)}}{\sqrt{K}}, \qquad U_m^{(j)} = \frac{\theta_j}{b_j} e^{\sigma_j \eta_m^{(j)}} - \frac{\Lambda_m^{(j)}}{\sqrt{K}}\,. \tag{4.1}$$

It is shown in Appendix B that the densities of $(\lambda_j/a_j) e^{\kappa_j \xi_m^{(j)}}$ and $(\theta_j/b_j) e^{\sigma_j \eta_m^{(j)}}$ are, respectively,

$$g_j(y) = \frac{1}{\sqrt{2\pi}\kappa_j y} \exp\left\{ -\frac{1}{2} \left[\frac{1}{\kappa_j} \ln\left(\frac{a_j y}{\lambda_j} \right) \right]^2 \right\}, \quad y > 0\,, \tag{4.2}$$

and

$$h_j(y) = \frac{1}{\sqrt{2\pi}\sigma_j y} \exp\left\{ -\frac{1}{2} \left[\frac{1}{\sigma_j} \ln\left(\frac{b_j y}{\theta_j} \right) \right]^2 \right\}, \quad y > 0\,. \tag{4.3}$$

Moreover, it is shown that $Y_m^{(j)}$ and $U_m^{(j)}$ have densities

$$g_j(y) + \frac{1}{2K} \left(\sum_{l=1}^{J} \Phi_l \gamma_l + \Phi \right) \frac{d^2 g_j}{dy^2} + \frac{\mu_3 \Phi^{3/2}}{6K^{3/2}} \frac{d^3 g_j}{dy^3} + O\left(\frac{1}{K^2} \right), \quad y > 0, \tag{4.4}$$

and

$$h_j(y) + \frac{1}{2K} \left(\sum_{l=1}^{J} \Phi_l \gamma_l + \Phi \right) \frac{d^2 h_j}{dy^2} + \frac{\mu_3 \Phi^{3/2}}{6K^{3/2}} \frac{d^3 h_j}{dy^3} + O\left(\frac{1}{K^2} \right), \quad y > 0\,, \tag{4.5}$$

respectively.

From (2.8)–(2.11), (3.16), and (4.1)–(4.3), the QoS requirements are

$$\Pr\left\{ Y_m^{(j)} \geq 1 - \frac{\Gamma_j}{K} \right\} \geq \int_{\frac{\lambda_j}{a_j} e^{-\kappa_j \nu_j}}^{\infty} g_j(y) dy \tag{4.6}$$

and

$$\Pr\left\{ U_m^{(j)} \geq 1 - \frac{\Gamma_j}{K} \right\} \geq \int_{\frac{\theta_j}{b_j} e^{-\sigma_j \rho_j}}^{\infty} h_j(y) dy\,. \tag{4.7}$$

With the densities of $Y_m^{(j)}$ and $U_m^{(j)}$ given by (4.4) and (4.5), these inequalities are asymptotically equivalent to

(4.8)

$$\int_{\frac{\lambda_j}{a_j}e^{-\kappa_j\nu_j}}^{\infty} g_j(y)dy$$
$$\leq \int_{1-\frac{\Gamma_j}{K}}^{\infty} g_j(y)dy - \frac{1}{2K}\left(\sum_{l=1}^{J}\Phi_l\gamma_l+\Phi\right)\frac{dg_j}{dy}(1) - \frac{\mu_3\Phi^{3/2}}{6K^{3/2}}\frac{d^2g_j}{dy^2}(1) + O\left(\frac{1}{K^2}\right)$$

and

(4.9)

$$\int_{\frac{\theta_j}{b_j}e^{-\sigma_j\rho_j}}^{\infty} h_j(y)dy$$
$$\leq \int_{1-\frac{\Gamma_j}{K}}^{\infty} h_j(y)dy - \frac{1}{2K}\left(\sum_{l=1}^{J}\Phi_l\gamma_l+\Phi\right)\frac{dh_j}{dy}(1) - \frac{\mu_3\Phi^{3/2}}{6K^{3/2}}\frac{d^2h_j}{dy^2}(1) + O\left(\frac{1}{K^2}\right).$$

However, from (4.2) and (4.3),

$$\frac{dg_j}{dy}(1) = -\left[1+\frac{1}{\kappa_j^2}\ln\left(\frac{a_j}{\lambda_j}\right)\right]g_j(1), \tag{4.10}$$

$$\frac{d^2g_j}{dy^2}(1) = \left\{\left[1+\frac{1}{\kappa_j^2}\ln\left(\frac{a_j}{\lambda_j}\right)\right]\left[2+\frac{1}{\kappa_j^2}\ln\left(\frac{a_j}{\lambda_j}\right)\right]-\frac{1}{\kappa_j^2}\right\}g_j(1),$$

and

$$\frac{dh_j}{dy}(1) = -\left[1+\frac{1}{\sigma_j^2}\ln\left(\frac{b_j}{\theta_j}\right)\right]h_j(1), \tag{4.11}$$

$$\frac{d^2h_j}{dy^2}(1) = \left\{\left[1+\frac{1}{\sigma_j^2}\ln\left(\frac{b_j}{\theta_j}\right)\right]\left[2+\frac{1}{\sigma_j^2}\ln\left(\frac{b_j}{\theta_j}\right)\right]-\frac{1}{\sigma_j^2}\right\}h_j(1).$$

Also

$$\int_{1-\frac{(\zeta+\xi)}{K}-\frac{\eta}{K^{3/2}}}^{\infty} g_j(y)dy \tag{4.12}$$
$$= \int_{1-\frac{\zeta}{K}}^{\infty} g_j(y)dy + \left(\frac{\xi}{K}+\frac{\eta}{K^{3/2}}\right)g_j(1) + O\left(\frac{1}{K^2}\right)$$

and

$$\int_{1-\frac{(\zeta+\mu)}{K}-\frac{\nu}{K^{3/2}}}^{\infty} h_j(y)dy \tag{4.13}$$
$$= \int_{1-\frac{\zeta}{K}}^{\infty} h_j(y)dy + \left(\frac{\mu}{K}+\frac{\nu}{K^{3/2}}\right)h_j(1) + O\left(\frac{1}{K^2}\right).$$

It follows from (4.8)–(4.13) that, for $j = 1, \ldots, J$,

$$\frac{\lambda_j}{a_j} e^{-\kappa_j \nu_j} \geq 1 - \frac{\Gamma_j}{K} - \frac{1}{2K} \left(\sum_{l=1}^{J} \Phi_l \gamma_l + \Phi \right) \left[1 + \frac{1}{\kappa_j^2} \ln \left(\frac{a_j}{\lambda_j} \right) \right] + \frac{\mu_3 \Phi^{3/2}}{6K^{3/2}} \left\{ \left[1 + \frac{1}{\kappa_j^2} \ln \left(\frac{a_j}{\lambda_j} \right) \right] \left[2 + \frac{1}{\kappa_j^2} \ln \left(\frac{a_j}{\lambda_j} \right) \right] - \frac{1}{\kappa_j^2} \right\} + O\left(\frac{1}{K^2} \right) \tag{4.14}$$

and

$$\frac{\theta_j}{b_j} e^{-\sigma_j \rho_j} \geq 1 - \frac{\Gamma_j}{K} - \frac{1}{2K} \left(\sum_{l=1}^{J} \Phi_l \gamma_l + \Phi \right) \left[1 + \frac{1}{\sigma_j^2} \ln \left(\frac{b_j}{\theta_j} \right) \right] + \frac{\mu_3 \Phi^{3/2}}{6K^{3/2}} \left\{ \left[1 + \frac{1}{\sigma_j^2} \ln \left(\frac{b_j}{\theta_j} \right) \right] \left[2 + \frac{1}{\sigma_j^2} \ln \left(\frac{b_j}{\theta_j} \right) \right] - \frac{1}{\sigma_j^2} \right\} + O\left(\frac{1}{K^2} \right). \tag{4.15}$$

These are the deterministic inequalities implied by the stochastic ones in (3.16).

To lowest order,

$$\lambda_j \geq a_j e^{\kappa_j \nu_j} \left[1 + O\left(\frac{1}{K} \right) \right], \qquad \theta_j \geq b_j e^{\sigma_j \rho_j} \left[1 + O\left(\frac{1}{K} \right) \right]. \tag{4.16}$$

Hence, from (3.12),

$$P_j \geq a_j e^{\kappa_j \nu_j} G \left[1 + O\left(\frac{1}{K} \right) \right], \qquad p_j \geq b_j e^{\sigma_j \rho_j} G \left[1 + O\left(\frac{1}{K} \right) \right]. \tag{4.17}$$

We define

$$\tau = \sum_{j=1}^{J} \left[a_j e^{\kappa_j \nu_j} e^{\kappa_j^2/2} w_j + b_j e^{\sigma_j \rho_j} e^{\sigma_j^2/2} (1 - w_j) \right] \gamma_j . \tag{4.18}$$

Then, from (3.11) and (4.17), we obtain

$$\left[1 - \tau + O\left(\frac{1}{K} \right) \right] G \geq N . \tag{4.19}$$

We assume that the target powers are limited, as in (2.12), and we define

$$\Delta_0 = \max_j \max \left(\frac{a_j}{\overline{P}_j} e^{\kappa_j \nu_j}, \ \frac{b_j}{\overline{p}_j} e^{\sigma_j \rho_j} \right) \tag{4.20}$$

and assume that $N \Delta_0 < 1$. It follows from (4.17) that

$$\frac{1}{G} \geq \Delta_0 \left[1 + O\left(\frac{1}{K} \right) \right], \tag{4.21}$$

and (4.19) implies that

$$\tau \leq 1 - N\Delta_0 + O\left(\frac{1}{K}\right), \tag{4.22}$$

which is the lowest-order approximation to the admissible set. Also, from (4.17) and (4.19), the minimal powers are

$$P_j^* = \frac{Na_j e^{\kappa_j \nu_j}\left[1 + O\left(\frac{1}{K}\right)\right]}{\left[1 - \tau + O\left(\frac{1}{K}\right)\right]}, \qquad p_j^* = \frac{Nb_j e^{\sigma_j \rho_j}\left[1 + O\left(\frac{1}{K}\right)\right]}{\left[1 - \tau + O\left(\frac{1}{K}\right)\right]}. \tag{4.23}$$

We investigate correction terms in the next section.

5. Refined approximation. The minimal powers P_j^* and p_j^* correspond to equality in (4.17) and (4.19) and hence, from (3.12),

$$\lambda_j = a_j e^{\kappa_j \nu_j}\left(1 + \frac{n_j}{K}\right), \qquad \theta_j = b_j e^{\sigma_j \rho_j}\left(1 + \frac{r_j}{K}\right), \tag{5.1}$$

where $n_j = O(1)$ and $r_j = O(1)$, and

$$\frac{1}{G} = \frac{(1-\tau)}{N} + O\left(\frac{1}{K}\right). \tag{5.2}$$

Then, from (3.12), (3.13), (5.1), and (5.2),

$$\Gamma_j = \Theta_j + O(1/K), \quad \Phi_l = \Psi_l + O(1/K), \quad \Phi = \Psi + O(1/K), \tag{5.3}$$

where

$$\Theta_j = a_j e^{\kappa_j \nu_j} e^{\kappa_j^2/2} w_j + b_j e^{\sigma_j \rho_j} e^{\sigma_j^2/2}(1 - w_j) \tag{5.4}$$

and

$$\Psi_l = a_l^2 e^{2\kappa_l \nu_l} e^{2\kappa_l^2} w_l + b_l^2 e^{2\sigma_l \rho_l} e^{2\sigma_l^2}(1 - w_l) - \Theta_l^2, \ \Psi = V\left(\frac{1-\tau}{N}\right)^2. \tag{5.5}$$

Also, (4.14) and (4.15) imply that

$$n_j \geq \Omega_j + \frac{M_j}{\sqrt{K}} + O\left(\frac{1}{K}\right), \qquad r_j \geq Z_j + \frac{Q_j}{\sqrt{K}} + O\left(\frac{1}{K}\right), \tag{5.6}$$

where

$$\Omega_j = \frac{1}{2}\left(\frac{\nu_j}{\kappa_j} - 1\right)\left(\sum_{l=1}^{J} \Psi_l \gamma_l + \Psi\right) - \Theta_j, \tag{5.7}$$

$$Z_j = \frac{1}{2}\left(\frac{\rho_j}{\sigma_j} - 1\right)\left(\sum_{l=1}^{J} \Psi_l \gamma_l + \Psi\right) - \Theta_j,$$

and

$$M_j = \frac{1}{6}\mu_3 \Psi^{3/2}\left[\left(\frac{\nu_j}{\kappa_j} - 1\right)\left(\frac{\nu_j}{\kappa_j} - 2\right) - \frac{1}{\kappa_j^2}\right], \tag{5.8}$$

$$Q_j = \frac{1}{6}\mu_3 \Psi^{3/2}\left[\left(\frac{\rho_j}{\sigma_j} - 1\right)\left(\frac{\rho_j}{\sigma_j} - 2\right) - \frac{1}{\sigma_j^2}\right].$$

We define

$$\chi = \sum_{j=1}^{J} [a_j e^{\kappa_j \nu_j} e^{\kappa_j^2/2} w_j \Omega_j + b_j e^{\sigma_j \rho_j} e^{\sigma_j^2/2} (1 - w_j) Z_j] \gamma_j, \tag{5.9}$$

$$u = \sum_{j=1}^{J} \left[a_j e^{\kappa_j \nu_j} e^{\kappa_j^2/2} w_j M_j + b_j e^{\sigma_j p_j} e^{\sigma_j^2/2} (1 - w_j) Q_j \right] \gamma_j \,.$$

From (3.12), (5.1), and (5.6), we obtain lower bounds on P_j and p_j. If we multiply these expressions by $w_j \gamma_j \exp(\kappa_j^2/2)$ and $(1 - w_j)\gamma_j \exp(\sigma_j^2/2)$, respectively, sum on j, and use the definition of G in (3.11) and those of τ, χ, and u in (4.18) and (5.9), we obtain

$$\left[1 - \tau - \frac{\chi}{K} - \frac{u}{K^{3/2}} + O\left(\frac{1}{K^2}\right) \right] G \geq N \,. \tag{5.10}$$

The minimal powers, corresponding to equality in (5.6) and (5.10), are

$$P_j^* = \frac{N a_j e^{\kappa_j \nu_j} \left[1 + \frac{\Omega_j}{K} + \frac{M_j}{K^{3/2}} + O\left(\frac{1}{K^2}\right) \right]}{\left[1 - \tau - \frac{\chi}{K} - \frac{u}{K^{3/2}} + O\left(\frac{1}{K^2}\right) \right]} \tag{5.11}$$

and

$$p_j^* = \frac{N b_j e^{\sigma_j \rho_j} \left[1 + \frac{Z_j}{K} + \frac{Q_j}{K^{3/2}} + O\left(\frac{1}{K^2}\right) \right]}{\left[1 - \tau - \frac{\chi}{K} - \frac{u}{K^{3/2}} + O\left(\frac{1}{K^2}\right) \right]} \,. \tag{5.12}$$

However, $P_j^* \leq \overline{P}_j$ and $p_j^* \leq \overline{p}_j$, $j = 1, \ldots, J$. We define

$$\Delta = \max_j \max \left[\frac{a_j}{\overline{P}_j} e^{\kappa_j \nu_j} \left(1 + \frac{\Omega_j}{K} + \frac{M_j}{K^{3/2}} \right), \right. \tag{5.13}$$

$$\left. \frac{b_j}{\overline{p}_j} e^{\sigma_j \rho_j} \left(1 + \frac{Z_j}{K} + \frac{Q_j}{K^{3/2}} \right) \right] ,$$

and assume that $N\Delta < 1$. Then, from (3.12), (5.1), and (5.6),

$$\frac{1}{G} \geq \Delta \left[1 + O\left(\frac{1}{K^2}\right) \right] , \tag{5.14}$$

and, from (5.10), the refined approximation to the admissible set is

$$\tau + \frac{\chi}{K} + \frac{u}{K^{3/2}} \leq 1 - N\Delta + O\left(\frac{1}{K^2}\right) . \tag{5.15}$$

We now express the results in terms of the original variables and define

$$f_j = \frac{W}{K} \Theta_j, \qquad \psi_j = \left(\frac{W}{K}\right)^2 \Psi_j, \quad \psi = \frac{W}{K} \Psi \,, \tag{5.16}$$

$$\omega_j = \frac{W}{K} \Omega_j, \qquad \zeta_j = \frac{W}{K} Z_j \,, \tag{5.17}$$

$$m_j = \left(\frac{W}{K}\right)^{3/2} M_j, \qquad q_j = \left(\frac{W}{K}\right)^{3/2} Q_j \,,$$

and

$$\delta_0 = \frac{W}{K}\Delta_0, \quad \delta = \frac{W}{K}\Delta, \quad X = \frac{W}{K}\chi, \quad U = \left(\frac{W}{K}\right)^{3/2} u\,. \tag{5.18}$$

Then, from (3.8), (3.9), (3.15), (5.4), (5.5), (5.7), and (5.8), we obtain (2.17)–(2.21). Also, from (3.8), (3.15), (4.18), and (5.9), we obtain (2.14), (2.22), and (2.23). Finally, from (3.15), (4.20), and (5.13), we obtain (2.13) and (2.24).

To lowest order, from (4.22) and (4.23), ignoring the $O(1/K)$ terms, the admissible set is given by (2.15), and the minimal powers by (2.16). In the refined approximation, from (5.11), (5.12), and (5.15), ignoring the $O(1/K^2)$ terms, we obtain (2.25) and (2.26).

6. Numerical results. The first order approximation is that given by Proposition 2.1. The second order approximation corresponds to setting $\mu_3 = 0$, i.e., to neglecting the effects of skewness, so that, from (2.21) and (2.23), $m_j = 0$, $q_j = 0$, and $U = 0$ in (2.24)–(2.26). The third order approximation retains $\mu_3 \neq 0$.

In the following results the external interference is simulated (as well as analyzed) as being statistically independent of the power control processes which are taking place in the desired or target cell. This is obviously a simplification of what would happen in an actual network. We further suppose that the first three moments (mean, variance, and skewness) of this external interference are known exactly at the desired cell. Thus the impact on capacity of estimation error is neglected. Finally, we neglect background noise, which is supposed to be small in comparison with the interference.

The path loss propagation law is taken to be

$$\gamma_{Loss} \propto R_D^{-\alpha},$$

where γ_{loss} is the absolute path loss, R_D is the distance between the mobile and base, and $\alpha > 0$ is a small positive constant taken to be roughly $2 \leq \alpha \leq 4$ in most wireless network models. The constant of proportionality is neglected for the results presented, as the background noise is not taken into account.

The simulations all use the power settings derived from the preceding analysis for the second order approximation. The sample outages over 100000 trials were then examined for numbers of mobiles varying close to the second order approximation. The largest number of mobiles achieving an outage strictly smaller than the target is given as the result of the simulation. A more thorough trial and error search for the optimal power ratio was not conducted.

In general all three approximations were in complete agreement with the simulation (after rounding down, rather than up, to avoid violating any constraint determined by the asymptotics). However, we did obtain discrepancies between the first order approximation and the simulated results in some cases, and to a lesser extent between the second- and third order approximations and the simulations. The following is an example.

In this example the interference is taken to be normal, with the mean and variance of the interference per mobile being $M = 0.6$ and $v = 0.015$, respectively. There were 6 interfering cells, corresponding to a hexagonal cellular array. The power control error was taken as $10\log_{10}\kappa = 10\log_{10}\rho = 1$ dB. The number K_0 of interfering mobiles and other parameters is given in Table 1. The received power target was set according to the second order approximation.

TABLE 1
Parameters for the normal interference experiment.

K_0	Actv/Idle pwr	w	L/l	W (MHz)	$\alpha/\beta \times 10^3$
12	1.0 (0 dB)/0.12589 (−9 dB)	0.5	0.002/0.01	10	384/48

TABLE 2
Simulation, first-, and second order capacity estimates.

# Simul.	1st order	2nd order
19	21 (21.44578)	19 (19.214888)

TABLE 3
Parameters for capacity with normal interference experiment.

K_0	$\overline{P}$	$\overline{p}$	w	L/l	W (MHz)	$\alpha/\beta \times 10^3$
120	1.0 (0 dB)	0.12589 (−9 dB)	0.5	0.1/0.1	5	48/6

Our results are given in Table 2. These show that the first order approximation overestimates the capacity of the desired cell by 2, whereas the second order and simulation results are in agreement. The received power limits were $\overline{P}, \overline{p} = 1, 0.12589$, respectively. Under the second order approximation the actual targets were $1, 0.1070825$ so that the active power target was at constraint.

Since in all of our numerical results the third order approximation, which includes the skewness of the external interference, differed only slightly from the second order, we have omitted any presentation of these results. However, it was imperative to ascertain that the skewness made only a slight difference; see [3].

6.1. Normal interference. Here we examine the impact of the mean interference on capacity. The external interference is modeled as normal, independent of the target cell. The mean and variance of the interference are $K_0 M$ and $K_0 v$, where K_0 is the total number of mobiles in the interfering cells. $v = 0.0025$, and the remaining parameters for the simulation runs are given in Table 3.

Our results are presented in Table 4. Where the first order approximation does not agree with the simulation it underestimates the capacity only by 1. The second order approximation is in complete agreement with the simulation, with one exception due to rounding down.

6.2. Capacity and control error. In this experiment external interference is generated by K_0 mobiles which are placed at random in cells surrounding the desired cell. The positions of these mobiles are fixed throughout the simulation. Since the path loss for each mobile is determined with an exponent of 3.5, this is also fixed.

The other parameters for the experiment are tabulated in Table 5. Note that since both outages and control errors (see Tables 5, 7) are the same, the ratio of the active to idle transmit powers equals the ratio of α to β. Since this also holds for the interfering cells, in this special case, the statistics of the interference are completely determined. Furthermore, as there is no external noise, performance does not depend on the transmitted powers as such. An increase in transmitted powers in the desired cell by any given factor leads to an increase in transmitted power in the interfering cells by the same factor. Thus performance depends only on the active to idle power ratio. (In more general cases, the idle and active transmit powers are not in this ratio and do depend on the external interference.)

TABLE 4
Capacity versus M with normal interference.

M	1st order	2nd order	Simul.
0.020000	34 (34.4284)	34 (34.9288)	34
0.029000	33 (33.1721)	33 (33.6907)	33
0.038000	31 (31.9158)	32 (32.4527)	32
0.047000	30 (30.6595)	31 (31.2147)	31
0.056000	29 (29.4032)	29 (29.9768)	30
0.065000	28 (28.1469)	28 (28.7390)	28
0.074000	26 (26.8906)	27 (27.5011)	27
0.083000	25 (25.6343)	26 (26.2633)	26
0.092000	24 (24.3780)	25 (25.0255)	25
0.101000	23 (23.1217)	23 (23.7877)	23
0.110000	21 (21.8654)	22 (22.5499)	22
0.119000	20 (20.6091)	21 (21.3121)	21
0.128000	19 (19.3528)	20 (20.0744)	20
0.137000	18 (18.0965)	18 (18.8367)	18
0.146000	16 (16.8402)	17 (17.5990)	17
0.155000	15 (15.5839)	16 (16.3613)	16
0.164000	14 (14.3276)	15 (15.1237)	15
0.173000	13 (13.0713)	13 (13.8860)	13
0.182000	11 (11.8150)	12 (12.6484)	12
0.191000	10 (10.5587)	11 (11.4108)	11
0.200000	9 (9.3024)	10 (10.1732)	10

TABLE 5
Parameters for capacity and control error experiment.

K_0	Actv/Idle pwr	w	L/l	W (MHz)	$\alpha/\beta \times 10^3$
120	1.0 (0 dB) /0.12589 (-9 dB)	0.5	0.1	5	48/6

TABLE 6
Mean and variance per mobile.

M	v	$\hat{M}$	$\hat{v}$
0.026437	0.002286	0.02642	0.0022788

Table 6 shows the results for the mean and variance of the interference. These were calculated using the activity and power variables given in Table 5. The activity of the interfering mobiles was simulated and used in determining interference. The mean and variance of this interference was sampled during the simulation, and illustrative results are also tabulated in Table 6, which are consistent with the calculated values. Our results are depicted in Table 7. These results indicate that capacity is significantly affected by power control error. In fact, the capacity diminishes by a factor of roughly 10 as the power control error increases from 0.5 to 5.0 dB. The second order approximation is in complete agreement with the simulation; however, the first order calculation has a small error in the cases of 0.5 dB (overestimate) and 5.0 dB (underestimate).

6.3. Minimal powers. As described earlier, if the system is underloaded, the powers may be reduced, thus restricting the interference caused to adjacent cells. The parameters used are the same as for the previous simulations except that the powers were set according to Table 8 and the power control error was taken as $10 \log_{10} \kappa =$

TABLE 7
Impact of κ on capacity.

κ	1st order	2nd order	Simul.
0.5/0.5	153	151	151
1.0/ 1.0	128	128	128
1.5/ 1.5	106	106	106
2.0/2.0	86	86	86
2.5/2.5	69	69	69
3.0/3.0	54	54	54
3.5/3.5	42	42	42
4.0/4.0	31	32	32
4.5/4.5	23	23	23
5.0/5.0	16	17	17

TABLE 8
Optimal power settings.

# Mobiles	Active power	Idle power
20	7.882376e-02	9.852970e-03
40	1.080771e-01	1.350964e-02
60	1.733684e-01	2.167105e-02
80	4.486806e-01	5.608508e-02

TABLE 9
Mobiles supported and simulated outages.

# Mobiles	Simul.	Actv outage	Idle outage
20	20	9.983743e-02	9.930380e-02
40	40	9.998671e-02	9.985218e-02
60	60	9.992572e-02	9.991928e-02
80	80	9.988996e-02	9.975853e-02

$10 \log_{10} \rho = 2dB$. The capacity of the system was determined to be 86 mobiles using both the first- and second order approximations.

The results from the simulations are given in Table 9. The left-hand column is the desired number of mobiles to support. The simulation column gives the maximum number of mobiles which achieved an outage strictly smaller than the desired outages. The final columns give the outages achieved for the desired number of mobiles given in the first column. The results show that the actual outages are very close to their desired values and that the simulation and desired number of mobiles are in agreement.

6.4. Two-class example. Table 10 gives the class dependent parameters which were used in the simulation. In addition, $W = 5$ MHz, $M = 0.05$, $v = 0.005, K_0 = 10$. The first- and second order capacity regions were determined in two steps. In the first step the maximum number of users in each class was determined using the asymptotics for a *single* class, yielding $K_1^{\max}, K_2^{\max}$. Then in the second step the maximum number of class 2 users was obtained, with class 1 users fixed at $K_1 = 1, \ldots, K_1^{\max}$. This search was performed by binary chop for both first- and second order approximations. Table 11 shows the results for both the first- and second order capacity regions. The simulation results were also obtained by fixing the number of class 1 users and then the number of class 2 users for an interval centered on the second order approximation. The powers were set for each pair of points (K_1, K_2) according to the second order approximation at the boundary. The largest value of

TABLE 10
Parameters for the two-class experiment.

Class	Actv/Idle pwr	w	L/l	κ/ρ (dB)	$\alpha/\beta \times 10^3$
1	1.0/0.125	0.3	0.1/0.01	2/2	250/5
2	1.0/0.125	0.4	0.1/0.1	2/2	125/5

TABLE 11
Results from the two-class experiment.

K_1	Simul.	1st order	2nd order
1	42	44 (44.442818)	43 (43.877646)
2	41	42 (42.920880)	42 (42.307850)
3	39	41 (41.398941)	40 (40.737816)
4	38	39 (39.877002)	39 (39.167544)
5	36	38 (38.355064)	37 (37.597036)
6	35	36 (36.833125)	36 (36.026291)
7	33	35 (35.311186)	34 (34.455309)
8	31	33 (33.789248)	32 (32.884093)
9	30	32 (32.267309)	31 (31.312641)
10	28	30 (30.745370)	29 (29.740956)
11	27	29 (29.223432)	28 (28.169036)
12	25	27 (27.701493)	26 (26.596883)
13	24	26 (26.179555)	25 (25.024498)
14	22	24 (24.657616)	23 (23.451881)
15	20	23 (23.135677)	21 (21.879032)
16	19	21 (21.613739)	20 (20.305951)
17	17	20 (20.091800)	18 (18.732641)
18	16	18 (18.569861)	17 (17.159100)
19	14	17 (17.047923)	15 (15.585330)
20	13	15 (15.525984)	14 (14.011330)
21	11	14 (14.004045)	12 (12.437102)
22	9	12 (12.482107)	10 (10.862646)
23	8	10 (10.960168)	9 (9.287963)
24	6	9 (9.438229)	7 (7.713053)
25	5	7 (7.916291)	6 (6.137915)
26	3	6 (6.394352)	4 (4.562552)
27	1	4 (4.872413)	2 (2.986963)

K_2 so that all outage probabilities were strictly met was given as the result of the simulation. The maximum number of K_1 users was determined by the second order approximation.

As the results show, the first order approximation and to a lesser extent the second order approximation overestimate the capacity region. This latter approximation agreed with simulation at both end points (the results were 28 (28.24095) and 45 (45.447203)).

The reader is cautioned against using the multiclass results when not all classes have a large number of mobiles, and in particular for empty classes. In the latter case, the maxima in the expressions for δ_0 and δ in (5.27) and (5.28), and for Δ_0 and Δ in (4.22) and (5.13), should be taken over nonempty classes.

6.5. Numerical result extensions. So far we have considered networks in which the mean path loss (average over fading) is determined by distance alone. However, more realistic models of networks of mobile phones often suppose that this quantity is determined as the product of the deterministic distance loss and an addi-

tional random log-normal shadowing factor

$$S_F \sim \text{lognormal}(0, \sigma^2).$$

We now discuss how our results can be used to examine the coverage and capacity of networks in which such log-normal shadowing is present. To do so we consider a single class of mobiles in all cells to avoid obscuring the discussion with unnecessary details. We also go in more detail into the question of obtaining numerical results where there are interfering cells.

As far as coverage is concerned, let P_T be the maximum transmit power of the mobile device used in the network. Once shadowing is determined by a random factor such as S_F, it is no longer the case that all mobiles within a distance R_c from the base station can meet the maximum received power requirement $\overline{Q} = \max(\overline{P}, \overline{p})$. (Here

$$R_c = \left(\frac{P_T C}{\overline{Q}} \right)^{\frac{1}{\alpha}},$$

and $\overline{P}, \overline{p}$ are the common active and idle receive power targets.) Instead, only a random fraction are within coverage, and the cell sizes are chosen to ensure that this fraction is suitably high, say f_c. Thus coverage must now be expressed as the maximum distance R_c such that

$$\Pr\left\{ \frac{P_T C}{R_M^{\alpha}} \geq \overline{Q} \right\} \geq f_c.$$

Of course, similar criteria may be used. In any case, whether or not there is log-normal shadowing, there is always in addition a coverage versus capacity trade-off in that the higher the maximum received power target, the higher the system capacity but the smaller the coverage area which can be supported. This trade-off can be explored using our results. See [1] for typical values of transmit power, acceptable path loss, and so on in both voice systems such as IS-95 and data services.

Log-normal shadowing also impinges on the capacity directly, in that it affects the distribution of interference between mobiles. Note that mobiles are usually assigned to the base station offering the lowest degree of path loss, and log-normal shadowing to distinct base stations is usually taken to be independent. Under these assumptions, the shadowing distribution is determined as order statistics of log-normal random variables, with means varying with mobile positions. If coverage is taken into account, then these log-normal random variables are truncated accordingly.

A final but very important general consideration is that the minimal received power targets (the ones likely to be used in practice) depend on the interference from the surrounding base stations, which in turn depend on their own receive power targets. In symmetrical cases it is reasonable to apply common received power targets at each cell and so establish the corresponding cell capacity. By varying the targets, capacity can be optimized. Moreover, this can be done in conjunction with the independent shadowing distribution so that the moments needed for the Edgeworth expansion of a random mobile's interference can be obtained by using, e.g., simulation or some other numerical technique. Obviously other factors, such as the distribution of position of the mobiles within the cells, numbers of interfering cells, etc., also have to be taken into account. Only results for a symmetrical case with mobiles along the cell edges were obtained in [1].

Of course if we fix the maximum received powers as target at each cell, distinct cell loadings may also be treated. Further work is needed to establish estimates for what the minimum feasible targets should be for heterogeneous cell loadings. As this goes into the question of power control interactions between cells, it is beyond the scope of this paper.

7. Conclusions. We have investigated a model of the reverse link of a single wireless cell with imperfect power control. For each class there is a specified small probability that the user's SNR will fall below a prescribed threshold, when active and when idle. We have investigated the capacity of the system, i.e., the number of users of each class that can be supported, and the minimal transmission powers when the link is operating within capacity. We have performed an asymptotic analysis, based on a large number of users in each class. The mean of the interference from users in adjacent cells affects the lowest order asymptotic approximation to the minimal powers and the capacity region, while the variance, and to a lesser extent the skewness, affects the refined approximation. Comparison with simulation illustrates that the refined approximation is good even for only moderately large numbers of users in each class.

The results have significant practical interest for power control in cellular systems. The power available at a mobile device for transmission in the reverse link is limited, so the minimal powers required to achieve the desired outage probabilities are important. It is essential to consider different classes of users, since voice transmission is often acceptable despite a significant error rate, while data transmission at such significant error rates would not be acceptable.

Also our results do carry implications for current commercial CDMA networks (see, e.g., [6]) even though these do not quite work according to our model. In particular, the received power targets (more precisely, E_b/I_0 targets) are determined by using feedback control instead of being determined in advance as here. (For example, in mobile voice this control reacts to the loss of speech frames by increasing the target sharply, and otherwise gradually decrementing it.) Nevertheless, our results show how sensitive the capacity is to fast power control error, and we anticipate that this will remain the case for the commercial systems too. Nor do we expect any significant difference in the capacities of commercial systems from the values which would be predicted by our approximation. Finally, our results concerning the received power targets should provide insight as to how the E_b/I_0 targets should be set in actual systems.

Further research is needed to examine the accuracy of our approach for a network of CDMA cells. In particular, the accuracy of the independence assumption is yet to be fully assessed. To conduct an analysis taking into account such dependence in detail appears difficult. Such a model would involve asymptotics for the joint state of the mobiles across the cells in the network. In particular, the computation of capacity would involve the asymptotics of the eigenvalues for a matrix determining the received power targets and would involve the propagation between desired base stations and interfering mobiles.

Appendix A. Uniform convergence to Gaussian density. We here establish (3.6), where $S^{(j)}$ is asymptotically normally distributed with zero mean, unit variance, and error $O(1/\sqrt{K_j})$ as $K_j \to \infty$. We define

$$\phi(\kappa;T) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{iTe^{\kappa\xi}} e^{-\xi^2/2} d\xi \,. \tag{A.1}$$

Then, from (2.1), since $e_m^{(j)} = e^{\kappa_j \xi_m^{(j)}}$ and $\epsilon_m^{(j)} = e^{\sigma_j \eta_m^{(j)}}$, where $\xi_m^{(j)}$ and $\eta_m^{(j)}$ are normally distributed, with zero mean and unit variance, and $(\xi_m^{(j)}, \eta_m^{(m)})$ and $X_m^{(j)}$ are mutually independent,

$$u_j(t) \equiv E\left(\exp\left\{it[P_j e_m^{(j)} X_m^{(j)} + p_j \epsilon_m^{(j)}(1 - X_m^{(j)})]\right\}\right) = w_j \phi(\kappa_j; P_j t) + (1 - w_j)\phi(\sigma_j; p_j t)\,. \tag{A.2}$$

We first show that $\phi(\kappa; T)$, where $\kappa > 0$, is absolutely integrable for $-\infty < T < \infty$. From (A.1), we have $|\phi(\kappa; T)| \le 1$. If we integrate twice by parts for $T \neq 0$, we obtain

$$\begin{aligned}\phi(\kappa; T) &= -\frac{i}{\sqrt{2\pi}\kappa T}\int_{-\infty}^{\infty} e^{iTe^{\kappa\xi}}(\kappa + \xi)e^{-\kappa\xi}e^{-\xi^2/2}d\xi \\ &= \frac{1}{\sqrt{2\pi}(\kappa T)^2}\int_{-\infty}^{\infty} e^{iTe^{\kappa\xi}}[1 - (\kappa + \xi)(2\kappa + \xi)]e^{-2\kappa\xi}e^{-\xi^2/2}d\xi\,. \end{aligned} \tag{A.3}$$

Hence,

$$|\phi(\kappa; T)| \le \frac{1}{\sqrt{2\pi}(\kappa T)^2}\int_{-\infty}^{\infty} |[1 - (\kappa + \xi)(2\kappa + \xi)]|\, e^{-2\kappa\xi}e^{-\xi^2/2}d\xi \equiv \frac{C(\kappa)}{T^2}\,, \tag{A.4}$$

which shows that $\phi(\kappa; T)$ is absolutely integrable. We assume that $\kappa_j > 0$ and $\sigma_j > 0$, and that $P_j > 0$, $p_j \ge 0$, and $0 < w_j \le 1$. If $p_j > 0$ or $w_j = 1$, then $u_j(t)$ is absolutely integrable, and (3.6) follows from (1.5.9) of [9].

If $p_j = 0$ and $0 < w_j < 1$, then we must modify the estimate (1.5.9) in [9], since $|u_j(t)|^\nu$ is not absolutely integrable for any $\nu \ge 1$ in this case. We now adopt the notation in [9] and, dropping the subscripts and superscripts and rescaling t, we define

$$\gamma(t) = w\phi(\kappa; t) + 1 - w\,. \tag{A.5}$$

To establish our result, we need to estimate

$$R = \left(\int_{\sqrt{n}c_3}^{\sqrt{n}r} + \int_{-\sqrt{n}r}^{-\sqrt{n}c_3}\right)\left[\gamma\left(\frac{t}{\sqrt{n}}\right)\right]^n e^{-i\mu_1\sqrt{n}t}e^{-iyt}dt \tag{A.6}$$

as $n \to \infty$, for real $y = O(1)$ and $r > c_3 > 0$. Here the mean $\mu_1 = we^{\kappa^2/2} > 0$.

Now, for $|x| < \delta$,

$$\begin{aligned}|(1 + x)^n - 1| &= \left|x\sum_{s=1}^{n}\binom{n}{s}x^{s-1}\right| \\ &\le \frac{|x|}{\delta}|(1 + \delta)^n - 1| \le \frac{|x|}{\delta}(1 + \delta)^n\,. \end{aligned} \tag{A.7}$$

However, $|\phi(\kappa; T)| \le \epsilon < 1$ for $|T| \ge c_3$. It follows that

$$|[\gamma(T)]^n - (1 - w)^n| \le \frac{1}{\epsilon}|\phi(\kappa; T)|(1 - w + \epsilon w)^n, \quad |T| \ge c_3\,. \tag{A.8}$$

Hence,

$$\begin{aligned}&\left|\int_{|t|\ge\sqrt{n}c_3}\left\{\left[\gamma\left(\frac{t}{\sqrt{n}}\right)\right]^n - (1 - w)^n\right\}e^{-i\mu_1\sqrt{n}t}e^{-iyt}dt\right| \\ &\le \frac{\sqrt{n}}{\epsilon}(1 - w + \epsilon w)^n\int_{|T|\ge c_3}|\phi(\kappa; T)|dT \to 0\end{aligned} \tag{A.9}$$

as $n \to \infty$, since $0 < 1 - w + \epsilon w < 1$.

Next, for $y = 0(1)$ and $\mu_1\sqrt{n} + y > 0$,

$$\text{(A.10)} \qquad \left| \left(\int_{\sqrt{n}c_3}^{\sqrt{n}r} + \int_{-\sqrt{n}r}^{-\sqrt{n}c_3} \right) e^{-\mu_1\sqrt{n}t} e^{-iyt} dt \right|$$

$$\leq \left| \frac{i}{(\mu_1\sqrt{n} + y)} \left\{ [e^{-i\mu_1\sqrt{n}t} e^{-iyt}]_{\sqrt{n}c_3}^{\sqrt{n}r} + [e^{-i\mu_1\sqrt{n}t} e^{-iyt}]_{-\sqrt{n}r}^{-\sqrt{n}c_3} \right\} \right|$$

$$\leq \frac{4}{(\mu_1\sqrt{n} + y)} .$$

This holds for all $r > c_3$. Hence, from (A.6), (A.9), and (A.10), we obtain

$$\text{(A.11)} \qquad |R| \leq \frac{\sqrt{n}}{\epsilon}(1 - w + \epsilon w)^n \int_{|T| \geq c_3} |\phi(\kappa; T)| dT + \frac{4(1-w)^n}{(\mu_1\sqrt{n} + y)} .$$

With this modification of the estimate in (1.5.9) in [9], the result in (3.6) follows.

Appendix B. Density calculations. We here determine the densities of the random variables $Y_m^{(j)}$ and $U_m^{(j)}$ defined in (4.1), where $\Lambda_m^{(j)}$ is given by (3.14). Since $\xi_m^{(j)}$ and $\eta_m^{(j)}$ are normally distributed with zero mean and unit variance, the characteristic functions of $(\lambda_j/a_j)e^{\kappa_j \xi_m^{(j)}}$ and $(\theta_j/b_j)e^{\sigma_j \eta_m^{(j)}}$ are

$$\text{(B.1)} \qquad \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{it(\lambda_j/a_j)e^{\kappa_j \zeta}} e^{-\zeta^2/2} d\zeta = \int_0^{\infty} e^{ity} g_j(y) dy$$

and

$$\text{(B.2)} \qquad \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{it(\theta_j/b_j)e^{\sigma_j \zeta}} e^{-\zeta^2/2} d\zeta = \int_0^{\infty} e^{ity} h_j(y) dy,$$

where the densities are given by (4.2) and (4.3).

The random variables $S^{(l)}$ and $S_m^{(j)}$ are asymptotically normally distributed with zero mean and unit variance and densities. Since $K_l = O(K)$, $l = 1, \ldots, J$, it follows from (1.5.3) of [9] that, for $t = O(1)$,

$$\text{(B.3)} \qquad E\left(e^{it\sqrt{\Phi_l \gamma_l} S^{(l)}/\sqrt{K}}\right) = e^{-\Phi_l \gamma_l t^2/(2K)} \left[1 + O\left(\frac{1}{K^2}\right)\right],$$

$$E\left(e^{it\sqrt{\Phi_j(\gamma_j - \frac{1}{K})} S_m^{(j)}/\sqrt{K}}\right) = e^{-\Phi_j \gamma_j t^2/(2K)} \left[1 + O\left(\frac{1}{K^2}\right)\right].$$

From (2.5), after integration by parts n times,

$$\text{(B.4)} \qquad \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{it\sqrt{\Phi} y/\sqrt{K}} e^{-y^2/2} He_n(y) dy = \left(\frac{it\sqrt{\Phi}}{\sqrt{K}}\right)^n e^{-\Phi t^2/(2K)},$$

and hence, from (2.4),

$$\text{(B.5)} \qquad E\left(e^{it\sqrt{\Phi} S/\sqrt{K}}\right) = e^{-\Phi t^2/(2K)} \left[1 - \frac{i\mu_3 \Phi^{3/2}}{6K^{3/2}} t^3 + O\left(\frac{1}{K^2}\right)\right].$$

Thus, from (3.14), since S, $S^{(l)}$, $l \neq j$, and $S_m^{(j)}$ are mutually independent,

$$\begin{aligned}(\text{B.6}) \qquad & E(e^{it\Lambda_m^{(j)}/\sqrt{K}}) \\ & = \exp\left(-\frac{1}{2K}\left[\sum_{l=1}^{J}\Phi_l\gamma_l + \Phi\right]t^2\right)\left[1 - \frac{i\mu_3\Phi^{3/2}}{6K^{3/2}}t^3 + O\left(\frac{1}{K^2}\right)\right] \\ & = 1 - \frac{1}{2K}\left[\sum_{l=1}^{J}\Phi_l\gamma_l + \Phi\right]t^2 - \frac{i\mu_3\Phi^{3/2}}{6K^{3/2}}t^3 + O\left(\frac{1}{K^2}\right)\end{aligned}$$

for $t = O(1)$.

Since the random variables $(\xi_m^{(j)}, \eta_m^{(j)})$ and $\Lambda_m^{(j)}$ are independent, it follows from (4.1), (B.1), (B.2), and (B.6) that the characteristic functions of $Y_m^{(j)}$ and $U_m^{(j)}$ are

$$(\text{B.7}) \qquad \left[1 - \frac{1}{2K}\left(\sum_{l=1}^{J}\Phi_l\gamma_l + \Phi\right)t^2 + \frac{i\mu_3\Phi^{3/2}}{6K^{3/2}}t^3 + O\left(\frac{1}{K^2}\right)\right]\int_0^\infty e^{ity}g_j(y)dy$$

and

$$(\text{B.8}) \qquad \left[1 - \frac{1}{2K}\left(\sum_{l=1}^{J}\Phi_l\gamma_l + \Phi\right)t^2 + \frac{i\mu_3\Phi^{3/2}}{6K^{3/2}}t^3 + O\left(\frac{1}{K^2}\right)\right]\int_0^\infty e^{ity}h_j(y)dy\,,$$

and hence the densities are given by (4.4) and (4.5), respectively.

Acknowledgments. The authors are grateful to Debasis Mitra for posing the capacity problem. They would also like to thank Sem Borst and Dennis Morgan for many helpful comments on earlier versions of this paper, and the referees for their helpful suggestions for improving the presentation.

REFERENCES

[1] D. Ayyagari and A. Ephremides, *Cellular multicode CDMA for integrated (voice and data) services*, IEEE J. Selected Areas of Communications, 17 (1999), pp. 928–938.

[2] R. Cameron and B. Woerner, *Performance analysis of CDMA with imperfect power control*, IEEE Trans. Comm., 44 (1996), pp. 777–781.

[3] C. C. Chan and S. V. Hanly, *Calculating the outage probability in a CDMA network with spatial Poisson traffic*, IEEE Trans. Veh. Tech., 50 (2001), pp. 183–204.

[4] C. C. Chan and S. V. Hanly, *Outage probabilities in CDMA networks with Poisson traffic: A skewness correction and a Chernoff bound*, in Proceedings of the 50th IEEE Vehicular Technology Conference (VTC), Amsterdam, 1999, pp. 2205–2210.

[5] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, John Wiley, New York, 1971.

[6] K. Gilhousen, I. Jacobs, R. Padovani, A. J. Viterbi, Jr., L. Weaver, and C. Wheatley, III, *On the capacity of a cellular CDMA system*, IEEE Trans. Veh. Tech., 40 (1991), pp. 303–312.

[7] S. V. Hanly, *An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity*, IEEE J. Selected Areas in Communications, 13 (1995), pp. 1332–1340.

[8] J. Holtzman, *A simple accurate method to calculate spread spectrum multiple access error probabilities*, IEEE Trans. Comm., 40 (1992), pp. 461–464.

[9] J. L. Jensen, *Saddlepoint Approximations*, Clarendon Press, Oxford, UK, 1995.

[10] Y. Lu and R. Brodersen, *Integrating power control, error correction and scheduling for a CDMA downlink system*, IEEE J. Selected Areas of Communications, 17 (1999), pp. 978–989.

[11] W. Magnus, F. Oberhettinger, and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1966.

[12] D. Mitra and J. A. Morrison, *A novel distributed power control algorithm for classes of service in cellular CDMA networks*, in Advances in Wireless Communications, J. Holtzman and M. Zorzi, eds., Kluwer Academic Publishers, Boston, 1998, pp. 187–202.

[13] S. Oh and K. Wasserman, *Dynamic spreading gain control in multiservice CDMA networks*, IEEE J. Selected Areas of Communications, 17 (1999), pp. 918–927.

[14] F. Priscoli and F. Sestini, *Effects of imperfect power control and user mobility on a CDMA cellular network*, IEEE J. Selected Areas of Communications, 14 (1996), pp. 1809–1817.

[15] R. Yates and C. Y. Huang, *Integrated power control and basestation assignment*, IEEE Trans. Veh. Tech., 44 (1995), pp. 638–644.

[16] Z. Zander, *Performance of optimum transmitter control in cellular radio systems*, IEEE Trans. Veh. Tech., 41 (1992), pp. 57–62.

SIAM J. Appl. Math.
Vol. 67, No. 6, pp. 1633–1653

CAVITATION ON DEFORMABLE GLACIER BEDS*

CHRISTIAN SCHOOF†

Abstract. The formation of water-filled cavities at the interface between a glacier and its bed can significantly affect the drainage of meltwater along the base of a glacier, which in turn is one of the most important controls on glacier sliding. In this paper, we analyze a mathematical model for cavity formation on deformable glacier beds. By contrast with the case of rigid glacier beds, the cavities described here are the result of an interfacial instability in coupled ice-sediment flow. This instability causes bumps on the ice-sediment interface to grow until normal stress in the lee of bed bumps drops to the local porewater pressure, at which point the ice begins to lose contact with the surface of the sediment. We extend the basic instability model to cover the case of cavity formation, and analyze the corresponding traveling wave problem. This takes the form of a viscous contact problem in which the obstacle on the boundary—the traveling bed bump caused by the initial instability—must be determined as part of the solution. A classical complex variable method allows the traveling wave problem to be cast as an eigenvalue problem which is straightforward to solve numerically. Our results show that solutions for different wavelengths can be obtained from an apparently unique solution to a scaled problem, and that the amplitude of traveling waves increases with wavelength, while their speed decreases with wavelength.

Key words. cavitation, contact problem, unilateral constraint, ice sheet

AMS subject classifications. 35J85, 76D07

DOI. 10.1137/050646470

1. Introduction.

The ice contained in many glaciers, especially those in mid-latitude mountain ranges, is close to the melting point, and meltwater generated at the glacier surface can reach the glacier bed through a network of conduits or *moulins* in the ice. The subsequent routing of meltwater via the glacier bed not only affects water discharge from the glacier, but also the dynamics of the glacier itself. The downslope motion of a glacier—which can be treated as a slowly flowing viscous body—consists of shearing in the ice and sliding at its base. High water pressure at the base of a glacier weakens the contact between ice and bed, and consequently reduces the amount of friction generated at the bed by sliding [7, 20, 17].

The presence of water-filled cavities at the interface between a glacier and its bed can play an important role in the drainage of meltwater along the glacier bed. The purpose of this paper is to analyze a model for the formation of such cavities. Previously developed mathematical models for subglacial cavitation [6, 17] deal with the flow of ice over a rigid glacier bed, and are essentially viscous analogues of elastic Signorini-type contact problems [13], although the subtle differences between the elastic and viscous cases appear to preclude a variational formulation for the latter. In contrast, the model considered here describes the spontaneous formation of cavities on a bed composed of deformable sediment. Consequently, the bed no longer represents a fixed "obstacle," but evolves as a result of stresses at the ice-sediment interface.

The instability mechanism which causes the formation of cavities in our model was

*Received by the editors November 30, 2005; accepted for publication (in revised form) May 30, 2007; published electronically September 14, 2007. This work was supported by the UK Engineering and Physical Sciences Research Council through a doctoral studentship at the Mathematical Institute, Oxford University, and by the U.S. National Science Foundation under grant DMS-03227943.
http://www.siam.org/journals/siap/67-6/64647.html

†Department of Earth and Ocean Sciences, University of British Columbia, 6339 Stores Rd., Vancouver, BC, V6T 1Z4 Canada (cschoof@eos.ubc.ca).

first proposed by Hindmarsh [11] and Fowler [8], and is described in detail in Schoof [19]. It relies on the pressure-dependence of the viscosity of subglacial sediment. (More precisely, sediment viscosity is assumed to depend on *effective pressure*, the difference between total pressure and porewater pressure, which controls how hard sediment grains are pressed together.) The mechanism may be summed up as follows: A shallow bump in the interface between ice and sediment causes a perturbation in the flow of ice over the bed, which leads to higher compressive normal stress being exerted on the upstream side of the bump than its lee. In turn, this causes increased effective pressure in the sediment layer upstream of the bump compared with downstream. Moreover, if the viscosity of the sediment is much lower than that of ice, the horizontal velocity of the interface is approximately constant (i.e., independent of position). Then, if the sediment rheology is such that flux in a thin sediment layer increases with effective pressure when the surface velocity of the layer is fixed, this implies that more sediment flows into the bump than out, causing it to grow.

This mechanism can be shown to work for a variety of viscous sediment rheologies, and numerical solutions of a simplified model show that growth of the instability is generally unbounded before the onset of cavitation [16, 19], which occurs when compressive normal stress in the lee of a bed bump drops to the local porewater pressure. In this paper, we will be concerned with the extension of that model to the case of cavitation. Our interest in this problem is largely motivated by the fact that cavity formation introduces a nonlinearity into the model, which may be sufficient to lead to bounded growth of the instability. Due to the complexity of the problem, we do not, however, consider the full time-dependent problem of cavity evolution, but restrict ourselves mostly to the case of traveling wave solutions, in the hope that these represent the fully evolved ice-bed interface.

The paper is structured as follows. In section 2, we describe the extension of the basic instability model to the case of cavitation. Subsequently, we formulate the traveling wave problem in section 3 and present a method of solution based on a classical complex variable approach. Results are discussed in section 4.

2. The model. We consider a simplified two-dimensional model for the spatially periodic flow of ice over a thin layer of water-saturated subglacial sediment in the absence of cavitation, as detailed in Schoof [19]. We set out the full time-dependent model here, first in the absence of cavitation and subsequently with cavitation. The analysis in this paper will, however, deal almost exclusively with the corresponding traveling wave problem, which we describe in the next section.

The basic assumptions of the model derived in [19] are the following. Ice is treated as an incompressible Newtonian fluid, while subglacial sediment is modeled as an incompressible shear-thinning viscous material whose viscosity additionally depends on *effective pressure*, defined as the difference between the ordinary pressure variable (the spherical part of the stress tensor in the language of continuum mechanics) and a prescribed porewater pressure. In particular, the rheology of subglacial sediment may be taken to be of the form

$$D_{ij} = KN^{-n}\tau^{m-1}\tau_{ij}, \tag{2.1}$$

where D_{ij} is strain rate, N is effective pressure, and τ_{ij} is deviatoric stress with second invariant $\tau = \sqrt{\tau_{ij}\tau_{ij}/2}$, while K, m, and n are positive constants. This rheological model has the qualitative features that strain rate increases with stress, as required for a viscous material, while strain rate decreases with effective pressure, corresponding to sediment grains less able to move past each other when pressed

harder together. The model described below makes a further approximation, treating only the parametric limit $n \approx m \gg 1$ in the rheology (2.1). This corresponds to a "nearly plastic" behavior, in which shear stress is only weakly dependent on strain rate [12]. As described in [19, section 7], this greatly simplifies expressions for volume flux in the sediment layer and shear stress at its surface, and leads to a more tractable ice flow problem.

Moreover, the model assumes that unstable waves generated at the ice-sediment interface have wavelengths that are long compared with the thickness of the sediment. Consequently, the sediment layer is treated as thin, with small surface slopes that allow the ice flow domain to be approximated by a half-space. Treating the sediment layer as thin further allows it to be described by a depth-integrated model that appears in the Stokes flow problem for the ice in the form of boundary conditions at the lower boundary. These boundary conditions can be derived essentially by integrating (2.1). For reasons of space, we refer to [19] for a more detailed derivation of the model.

Below, x and y are Cartesian coordinates parallel and perpendicular, respectively, to the mean bed elevation, while t is time and subscripts $_x$, $_y$, and $_t$ denote the corresponding partial derivatives. $\mathbf{u}(x, y, t)$ is a dimensionless velocity perturbation in the ice relative to a mean shearing flow, and $p(x, y, t)$ a dimensionless pressure perturbation about a mean hydrostatic pressure field. If a is the spatial period of the bed, $\mathbf{u} = (u, v)$ and p satisfy Stokes' equations on a semi-infinite strip in the upper half-plane:

$$\nabla^2 \mathbf{u} - \nabla p = \mathbf{0}, \qquad \nabla \cdot \mathbf{u} = 0 \qquad \text{on } (x, y) \in (0, a) \times (0, \infty), \tag{2.2}$$

with periodic boundary conditions applied at $x = 0$ and $x = a$. The interface between ice and sediment remains, at leading order, at $y = 0$ if waves on the bed are shallow. We denote the amplitude of these waves by $h(x, t)$, and sediment flux in the direction of the x-axis by $q(x, t)$. Effective pressure (the difference between confining normal stress and a prescribed porewater pressure) at the sediment surface will be denoted by $N(x, t)$, and shear stress at the ice-sediment interface by $\tau_b(x, t)$. Lastly, the velocity of the ice-sediment interface will be denoted by U. As before, all of these quantities have been scaled as in [19] and are dimensionless. Boundary conditions for the Stokes flow problems (2.2) are then

$$\left.\begin{aligned} u_y + v_x &\to \gamma^{-1}, \\ p &\to 0 \end{aligned}\right\} \qquad \text{as } y \to \infty, \tag{2.3}$$

$$\left.\begin{aligned} \gamma\,(u_y + v_x) &= \tau_b, \\ 1 + p - 2v_y &= N, \\ v &= U h_x + h_t \end{aligned}\right\} \qquad \text{on } y = 0. \tag{2.4}$$

Above, $\gamma > 0$ is the ratio of mean dimensional effective pressure to far-field shear stress. Hence γ^{-1} is a dimensionless far-field shear stress in $(2.3)_1$, while $(2.3)_2$ ensures that the pressure perturbation p vanishes at large distances from the bed. The boundary conditions $(2.4)_{1,2}$ at the bed relate the appropriate stress components in the Stokes problem to interfacial shear stress and effective pressure, while $(2.4)_3$ relates normal velocity at the bed to the evolution of bed wave amplitude h. h itself satisfies the evolution equation

$$h_t + q_x = 0. \tag{2.5}$$

As described at the beginning of this section, interfacial shear stress τ_b and flux q in the boundary conditions (2.4) must be determined through a model for the thin-

film flow of subglacial sediment. With the rheological specifications based on (2.1) described above, we find the appealingly simple (and *linear*) relationships

$$\tau_b(x,t) = N(x,t), \qquad q(x,t) = N(x,t). \tag{2.6}$$

Meanwhile, interface velocity is determined by a large-scale ice flow problem [18, 19], and in scaled terms we can simply set

$$U = 1, \tag{2.7}$$

while the x-component of the velocity $\mathbf{u} = (u, v)$ is subject to $\int_0^a u(x,0)\,\mathrm{d}x = 0$ (this condition being necessary to ensure a unique solution for $\mathbf{u}$).

As described above, the "constitutive relations" (2.6) are appropriate for a sediment layer flowing in simple shear with rheology given by (2.1) in the parametric limit $m \sim n \gg 1$ [19, section 7]. The relations are thus not completely general, though rheological tests support them [12]. As described in [16, 19] more general sediment rheologies can be introduced into the model simply by changing the prescriptions for τ_b and N in (2.6). We persist with (2.6) in part because it is supported by empirical evidence [12], and also because it is the simplest physically motivated choice we can make, yielding a linear relationships between τ_b, q, and N. However, as we point out in section 4, other rheological models for sediment may also be of practical interest, but these introduce the additional complication of nonlinear constitutive relations in (2.6), which are beyond the scope of this paper.

It is straightforward to show that the trivial solution $h(x,t) \equiv 0$ to (2.2)–(2.6) is unstable: The model admits Fourier mode solutions of the form $h(x,t) = \mathrm{Re}(\exp(ikx + \sigma t))$, where $\sigma = 2|k|^3/(1+2ik|k|)$ has a positive real part, and growth of the instability is apparently unbounded. However, this is physical only while effective pressure N is positive everywhere, which ensures that normal stress at the top of the sediment layer exceeds the porewater pressure within. Once $N = 0$ somewhere, compressive normal stress at the top of the sediment layer at that location equals porewater pressure, and porewater starts to leak out of the sediment. The ice loses contact with the sediment, and a water-filled cavity forms, as also happens in glacier sliding over undeformable beds [6, 17].

When a cavity has formed, different boundary conditions apply to (2.2) on those parts of the bed where cavities are present from those in effect where ice is in contact with sediment. Let the cavitated part of the bed at time t be denoted by $C(t)$, and the contact areas by $C'(t)$; the closure of $C \cup C'$ is then the interval $[0, a]$. The boundary conditions (2.4) together with the constraint $N \geq 0$ (which ensures that normal stress in contact areas cannot drop below the porewater pressure) still hold on $y = 0$, $x \in C'$. On cavitated parts of the bed, we require that effective pressure and shear stress vanish, as water pressure equals normal stress in the ice, and the water is assumed to be inviscid. This is tantamount to setting $\tau_b = N = 0$ in (2.4). In addition, the cavity roof must be above the surface of the sediment over cavitated parts of the bed and must satisfy a kinematic boundary condition analogous to $(2.4)_3$. We denote the dimensionless elevation of the cavity roof by $h_C(x,t)$ (see Figure 2.1) and assume that the cavity roof has a low aspect ratio, comparable with that of the sediment layer. Then we have [16, Chapter 6]

$$\left.\begin{aligned} 1 + p - 2v_y &= 0, \\ u_y + v_x &= 0, \\ v &= Uh_{Cx} + h_{Ct} \end{aligned}\right\} \qquad \text{on } y = 0,\ x \in C, \tag{2.8}$$

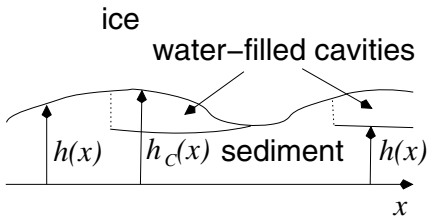


FIG. 2.1. *Illustration of cavity and contact areas, and the definition of h and h_C. The model assumes that bed slopes are small, so that the lower boundary of the ice flow domain can be reduced to $y = 0$ in dimensionless terms. Note that we generally assume discontinuities in h at downstream endpoints of contact areas. These sediment shocks are shown as dotted lines.*

while the absence of traction at the surface of the sediment in the cavities also implies

$$h_t = 0, \qquad q = 0, \qquad \text{on } y = 0,\ x \in C, \tag{2.9}$$

combined with the constraint $h_C > h$.

In addition to the boundary conditions (2.4) and (2.8) for the two-dimensional Stokes equations (2.2) and the evolution equation (2.5), we require jump conditions on h and h_C at the boundary points of C and C'. Based on physical considerations [16], we require that there be no discontinuities in the elevation of the lower boundary of the ice. Defining h_C everywhere as the scaled elevation of the base of the ice, so that $h_C(x) = h(x)$ for $x \in C'$, this implies that h_C is continuous across the endpoints of C and C'. It turns out that the same degree of continuity cannot be imposed on the sediment wave amplitude h, and a discontinuity must be expected at at least one endpoint of each individual contact area. Such discontinuities are not entirely unexpected: For the somewhat similar problem of dune formation in deserts and on river beds [9], the equivalent would be a slip face. The flow of sediment close to a contact point where h is discontinuous cannot be resolved by our thin-film approximation for the sediment. However, instead of attempting to solve an extremely complicated local sediment flow problem, we argue heuristically. We assume that the propagation speed of the contact point is determined by a Rankine–Hugeniot condition which ensures conservation of sediment:

$$V_s = [q]^+_-/[h]^+_-, \tag{2.10}$$

where V_s is the propagation speed of the sediment shock and $[\cdot]^+_-$ denotes the jump in the bracketed quantity across the shock. With $q = N$, it follows from the continuity of h_C and the inequality constraints on h and N that the jump in q must have the same sign as the jump in h, and hence the sediment shock must propagate downstream with $V_s > 0$.

If there were a sediment shock at the upstream end of a contact area, then effective pressure would be positive downstream of the contact point and zero in the cavity. The resulting pressure difference should drive a local sediment flow (not resolved by our thin-film approximation) into the cavity, that is, in the upstream direction and therefore opposite to that required by (2.10). As a consequence, we permit discontinuities in h only at the downstream ends of contact areas and require that h be continuous across the upstream end of each contact area.

As we shall see below, these jump conditions on h and h_C lead to an apparently

well-posed traveling wave problem. Whether or not they can in fact be applied to the general time-dependent problem is a matter for future research.

3. Traveling waves. In what follows, we consider solutions in which h and h_C depend on x and t only through the traveling wave coordinate

$$\eta = x - Vt, \tag{3.1}$$

where $V > 0$ is the unknown pattern speed of the traveling wave. Note that negative pattern speeds are not possible because sediment shocks must propagate downstream, as explained above. Writing $\hat{\nabla} = (\partial/\partial\eta, \partial/\partial y)$, the model can then be cast in the form

$$\hat{\nabla}^2\mathbf{u} - \hat{\nabla}p = \mathbf{0}, \qquad \hat{\nabla}\cdot\mathbf{u} = 0 \qquad \text{on } (\eta, y) \in (0, a) \times (0, \infty), \tag{3.2}$$

$$\left.\begin{aligned} u_y + v_\eta &\to \gamma^{-1}, \\ p &\to 0 \end{aligned}\right\} \qquad \text{as } y \to \infty, \tag{3.3}$$

$$\left.\begin{aligned} \gamma\,(u_y + v_\eta) &= \tau_b, \\ 1 + p - 2v_y &= N, \\ v &= (U - V)h', \\ q' - Vh' &= 0, \\ N &\geq 0 \end{aligned}\right\} \qquad \text{on } \eta \in \hat{C}',\ y = 0, \tag{3.4}$$

$$\left.\begin{aligned} u_y + v_\eta &= 0, \\ 1 + p - 2v_y &= 0, \\ v &= (U - V)h_C', \\ h' &= 0, \\ h_C &> h \end{aligned}\right\} \qquad \text{on } \eta \in \hat{C},\ y = 0, \tag{3.5}$$

$$q = N \qquad \text{on } \eta \in \hat{C}', \tag{3.6}$$

$$q = 0 \qquad \text{on } \eta \in \hat{C}, \tag{3.7}$$

$$\tau_b = N, \qquad U = 1. \tag{3.8}$$

Here primes on h, h_C, and q denote differentiation with respect to η, and $\hat{C}' = C'(0)$ and $\hat{C} = C(0)$ denote contact and cavity areas, respectively, at the bed in the (η, y) coordinate system. In addition, we consider only the case of a single cavity per bed period, and without further loss of generality we can set

$$\hat{C}' = (0, b), \qquad \hat{C} = (b, a), \tag{3.9}$$

where the contact point $\eta = b$ is to be determined as part of the solution. As before, we impose periodic boundary conditions on $\eta = 0$ and $\eta = a$. The continuity requirements on h and h_C at the contact points are

$$h(0^+) = h(a^-) = h_C(a^-), \qquad h(b^-) = h_C(b^+), \tag{3.10}$$

where superscripts $+$ and $-$ indicate limits taken from above and below, respectively (and where, in an abuse of notation, we have replaced the arguments (x, t) by η). Lastly, the jump condition (2.10) for a sediment shock at $\eta = b$ propagating at the pattern speed $V_s = V$ becomes

$$V = -q(b^-)/(h(b^+) - h(b^-)), \tag{3.11}$$

where we recognize that $q(b^+) = 0$ from (3.7).

Some small simplifications are immediately possible. As the model is invariant under changes of h and h_C by the same constant, we can without loss of generality set $h = 0$ in $\hat{C}$ by $(3.5)_4$. $(3.4)_4$ and (3.11) combined then require that

$$q = Vh \qquad \text{on } \hat{C}, \tag{3.12}$$

and the jump conditions on h and h_C become

$$h(0^+) = h_C(a^-) = 0, \qquad h(b^-) = h_C(b^+). \tag{3.13}$$

Equations (3.12) and (3.13) then replace $(3.4)_4$, (3.10), and (3.11) in the subsequent analysis.

The model described above is in many ways similar to the viscous contact problems considered by Fowler [6] and Schoof [17], in the sense that we have a Stokes flow problem with mixed boundary conditions prescribed on parts of the boundary which are not known a priori but must be found as part of the solution so as to satisfy the inequality constraints $(3.4)_5$ and $(3.5)_5$. This introduces an important nonlinearity into the problem which is missing in the original linear evolution problem and, as we shall see, allows for the existence of (nontrivial) traveling wave solutions, which are not possible without cavitation. The crucial difference between the model considered here and that in [6, 17] is that the bed elevation h is not fixed here but forms part of the solution. This renders the problem considerably more complicated.

The remainder of this section will be devoted to constructing a method of solution. Our approach consists of the following steps. First, we represent the solution of the Stokes equations (3.2) in terms of complex potentials. When mapped conformally so as to make use of the periodic boundary conditions at the sides of the domain, the boundary conditions (3.3)–(3.5) lead to a pair of Hilbert problems for these complex potentials, which admit explicit solutions in terms of the unknown functions N and τ_b and the contact point position b. Finally, applying the remaining conditions (3.8), (3.12), and (3.13) allows the problem of finding N and τ_b to be recast as an eigenvalue problem for h (which simultaneously determines the pattern speed V), and b can be determined through an additional integral constraint which ensures that the lower ice surface has no discontinuities.

3.1. Complex variable formulation. We introduce a stream function ψ such that

$$u = \psi_y + y/\gamma, \qquad v = -\psi_\eta, \tag{3.14}$$

where the additional shearing term in the definition of u accounts for the far-field shear stress. Further, we define the complex variables $z = \eta + iy$ and $\overline{z} = \eta - iy$. ψ satisfies the biharmonic equation, which can be written in terms of z and $\overline{z}$ as

$$\hat{\nabla}^4 \psi = 4\psi_{zz\overline{z}\overline{z}} = 0. \tag{3.15}$$

Using standard methods in complex analysis [5], ψ can be shown to take the general form

$$\psi = (\overline{z} - z)\theta(z) + \phi(z) + (z - \overline{z})\overline{\theta(z)} + \overline{\phi(z)} \tag{3.16}$$

for z in the semi-infinite strip $0 < \eta < a$, $0 < y$, where θ and ϕ are analytic functions and an overbar denotes complex conjugation. Furthermore, the Stokes equations (3.2) become

$$\hat{\nabla}^2 \psi_y = p_\eta, \qquad -\hat{\nabla}^2 \psi_\eta = p_y, \tag{3.17}$$

which are the Cauchy–Riemann relations for $p + i\hat{\nabla}^2\psi$ (i.e., pressure p and vorticity $\hat{\nabla}^2\psi$ are harmonic conjugates). Using standard differentiation rules [5], $\hat{\nabla}^2\psi = 4(\theta'(z) + \overline{\theta'(z)}) = 8\mathrm{Im}(i\theta'(z))$, where a prime denotes differentiation. Consequently,

$$p = C_0 + 8\mathrm{Re}(i\theta'(z)) = C_0 + 4i(\theta'(z) - \overline{\theta'(z)}), \tag{3.18}$$

where C_0 is a real constant. The value of this constant can be chosen arbitrarily: If $C_0 \neq 0$, we can add $iC_0 z/8$ to $\theta(z)$ and simultaneously add $iC_0 z^2/8$ to $\phi(z)$ while leaving the stream function ψ in (3.16) unchanged. By redefining θ and ϕ in this way, we ensure that $C_0 = 0$ in (3.18), and we will henceforth assume this to be the case.

In order to satisfy the periodic boundary conditions imposed on the problem, it suffices to ensure that u, v, and p can be extended to sufficiently smooth functions in the half-space $y > 0$, which are periodic in η with period a. Anticipating therefore that p and vorticity $\nabla^2\psi$ can be extended to harmonic functions which are appropriately periodic, we conclude that $\theta'(z)$ can be continued to an analytic function in the upper half-plane which is periodic in $\mathrm{Re}(z)$ with period a. Furthermore, we have for $y > 0$

$$u = 4y\mathrm{Re}\left[\theta'(z)\right] + 2\mathrm{Im}\left[2\theta(z) - \phi'(z)\right], \tag{3.19}$$

$$v = -4y\mathrm{Im}\left[\theta'(z)\right] - 2\mathrm{Re}\left[\phi'(z)\right]. \tag{3.20}$$

Hence, if $\theta'(z)$ can be continued to a periodic analytic function, it suffices to ensure in addition that $\mathrm{Im}[2\theta(z) - \phi'(z)]$ and $\mathrm{Re}[\phi'(z)]$ can be extended to harmonic functions in the upper half-plane with period a in $\mathrm{Re}(z)$. From the periodicity of $\mathrm{Re}[\phi'(z)]$, it follows that $\phi''(z)$ can be continued to an appropriately periodic analytic function in the entire upper half-plane $\mathrm{Im}(z) > 0$. Moreover, the periodicity of $\mathrm{Re}[\phi'(z)]$ and $\mathrm{Im}[2\theta(z) - \phi'(z)]$ are equivalent to

$$\mathrm{Re}\left[\phi'(a+iy) - \phi'(iy)\right] = \mathrm{Re}\left[\int_0^a \phi''(\eta + iy)\,\mathrm{d}\eta\right] = 0, \tag{3.21}$$

$$\mathrm{Im}\left[\int_0^a 2\theta'(\eta + iy) - \phi''(\eta + iy)\,\mathrm{d}\eta\right] = 0 \tag{3.22}$$

for all $y > 0$.

We complete our complex variable formulation by casting the boundary conditions (3.3)–(3.5) in terms of θ and ϕ. At the lower boundary $y = 0$

$$2i(\phi''(\eta) - \overline{\phi''(\eta)}) = \begin{cases} N(\eta) - 1, & \eta \in \hat{C}', \\ -1, & \eta \in \hat{C}, \end{cases} \tag{3.23}$$

$$2(2\theta'(\eta) - \phi''(\eta) + 2\overline{\theta'(\eta)} - \overline{\phi''(\eta)}) = \begin{cases} \gamma^{-1}\left(\tau_b(\eta) - 1\right), & \eta \in \hat{C}', \\ -\gamma^{-1}, & \eta \in \hat{C}, \end{cases} \tag{3.24}$$

$$-(\phi'(\eta) + \overline{\phi'(\eta)}) = \begin{cases} (U - V)h'(\eta), & \eta \in \hat{C}', \\ (U - V)h'_C(\eta), & \eta \in \hat{C}, \end{cases} \tag{3.25}$$

combined with (3.8), (3.12), and (3.13). Naturally, θ', ϕ', and ϕ'' are defined on the real axis as boundary values taken as z approaches the axis from above. As $\mathrm{Im}(z) = y \to \infty$, we have from (3.3)

$$4i\left[\theta'(z) - \overline{\theta'(z)}\right] \to 0, \tag{3.26}$$

$$-2\left[\phi''(z) - 2iy\theta''(z) - 2\theta'(z) + \overline{\phi''(z)} + 2iy\overline{\theta''(z)} - 2\overline{\theta'(z)}\right] \to 0. \tag{3.27}$$

3.2. Reformulation as a Hilbert problem. In order to exploit the periodicity of the problem, and to obtain a straightforwardly solved pair of Hilbert problems for proxies of θ' and ϕ'', we map conformally to the ζ-plane as

$$\zeta = \exp(i2\pi z/a), \qquad \xi = \exp(i2\pi\eta/a), \tag{3.28}$$

where $0 < \eta = \mathrm{Re}(z) < a$. We denote by Γ and Γ' the images of $\hat{C}$ and $\hat{C}'$ under this mapping. Γ and Γ' are then disjoint arcs of the unit circle in the ζ-plane, and the closure of $\Gamma \cup \Gamma'$ is the unit circle itself. We also define $\tilde{N}(\xi) = N(\eta)$, $\tilde{\tau}_b(\xi) = \tau_b(\eta)$, and let

$$\Omega(\zeta) = \begin{cases} \overline{\phi''(\overline{z})}, & |\zeta| > 1, \\ \phi''(z), & 0 < |\zeta| < 1, \end{cases} \tag{3.29}$$

$$\omega(\zeta) = \phi'(z), \qquad 0 < |\zeta| < 1, \tag{3.30}$$

$$\Theta(\zeta) = \begin{cases} \overline{\theta'(\overline{z})}, & |\zeta| > 1, \\ \theta'(z), & 0 < |\zeta| < 1. \end{cases} \tag{3.31}$$

From the Schwarz reflection principle and the periodicity requirements above, it follows that Ω and Θ are analytic in the finite ζ-plane cut along the unit circle and punctured at the origin, while ω is analytic inside the open unit disk cut along the nonnegative part of the real axis in the ζ-plane, where ω may be discontinuous because we know only that $\mathrm{Re}(\phi')$ is periodic. As we shall show later, ω is in fact analytic across that branch cut. Moreover, using $\mathrm{d}/\,\mathrm{d}z = (i2\pi/a)\zeta\,\mathrm{d}/\,\mathrm{d}\zeta$,

$$\Omega(\zeta) = (i2\pi/a)\zeta\omega'(\zeta) \tag{3.32}$$

for $|\zeta| < 1$, except on the branch cut.

The boundary conditions (3.23)–(3.25) become

$$2i\left[\Omega^+(\xi) - \Omega^-(\xi)\right] = \begin{cases} \tilde{N}(\xi) - 1, & \xi \in \Gamma', \\ -1, & \xi \in \Gamma, \end{cases} \tag{3.33}$$

$$2\left[2\Theta^+(\xi) - \Omega^+(\xi) + 2\Theta^-(\xi) - \Omega^-(\xi)\right] = \begin{cases} \gamma^{-1}(\tilde{\tau}_b(\xi) - 1), & \xi \in \Gamma', \\ -\gamma^{-1}, & \xi \in \Gamma, \end{cases} \tag{3.34}$$

$$-2\mathrm{Re}(\omega^+(\xi)) = \begin{cases} (U - V)h'(\eta), & \xi \in \Gamma', \\ (U - V)h'_C(\eta), & \xi \in \Gamma, \end{cases} \tag{3.35}$$

where superscripts + and − denote limits taken as the unit circle is approached from within and without, respectively. The first two of these equations take the form of standard Hilbert problems, whose solutions depend on the behavior of Ω and Θ at infinity. Because of the symmetry inherent in the definitions of Ω and Θ, their behavior at infinity is determined by their behavior at the origin. From (3.27), we have for $\zeta \to 0$

$$4i\left[\Theta(\zeta) - \overline{\Theta(\zeta)}\right] \to 0, \tag{3.36}$$

$$\Omega(\zeta) + \overline{\Omega(\zeta)} - 2\Theta(\zeta) - 2\overline{\Theta(\zeta)} - 2\log|\zeta|\left[\zeta\Theta'(\zeta) + \overline{\zeta\Theta'(\zeta)}\right] \to 0. \tag{3.37}$$

It follows from (3.36) that Θ is analytic at the origin with $\Theta(0) = C_1$, where C_1 is a real constant. Hence $\lim_{\zeta\to 0} \zeta\log|\zeta|\Theta'(\zeta) = 0$, and from (3.37) it follows that Ω is also analytic at the origin with $\Omega(0) = 4C_1 + iC_2$, where C_2 is another real constant.

It can then be shown from the periodicity requirements (3.22) that both C_1 and C_2 vanish. Specifically, (3.22) can be written as

$$\operatorname{Re}\left[\frac{a}{i2\pi}\oint_L \frac{\Omega(\zeta)\,\mathrm{d}\zeta}{\zeta}\right] = \operatorname{Im}\left[\frac{a}{i2\pi}\oint_L \frac{[2\Theta(\zeta)-\Omega(\zeta)]}{\zeta}\,\mathrm{d}\zeta\right] = 0, \tag{3.38}$$

where L is a circular contour about the origin with radius less than 1, traversed anticlockwise. Applying the residue theorem, this implies that $C_1 = C_2 = 0$, and hence

$$\Omega(0) = \Theta(0) = 0. \tag{3.39}$$

From (3.39) and the definitions of Ω and Θ in (3.31) it finally follows that

$$\Omega(\zeta) = \overline{\Omega(1/\bar{\zeta})}, \qquad \Omega(\infty) = 0, \qquad \Theta(\zeta) = \overline{\Theta(1/\bar{\zeta})}, \qquad \Theta(\infty) = 0, \tag{3.40}$$

which taken together also ensure the appropriate behavior at the origin.

We can now solve (3.33), (3.34), and (3.40) explicitly in terms of the as yet unknown effective pressure $N(\eta)$ and shear stress $\tau_b(\eta)$. For the sake of simplicity, define Ξ as a proxy for Θ through

$$\Xi(\zeta) = \begin{cases} 2\Theta(\zeta) - \Omega(\zeta), & |\zeta| < 1, \\ -2\Theta(\zeta) + \Omega(\zeta), & |\zeta| > 1, \end{cases} \tag{3.41}$$

so that (3.34) becomes

$$\Xi^+(\xi) - \Xi^-(\xi) = \begin{cases} \gamma^{-1}(\tilde{\tau}_b(\xi) - 1), & \xi \in \Gamma', \\ -\gamma^{-1}, & \xi \in \Gamma, \end{cases} \tag{3.42}$$

subject to $\Xi(\infty) = 0$, $\Xi(\zeta) = -\overline{\Xi(1/\bar{\zeta})}$. Assuming that $\tilde{N}$ and $\tilde{\tau}_b$ are Hölder continuous and bounded on Γ' (we exclude the possibility of integrable singularities at the endpoints of Γ' because N and τ_b are related to h through (3.8) and (3.12), and h is clearly bounded), (3.33) and (3.42) admit solutions vanishing at infinity of the form (see [14])

$$\Omega(\zeta) = -\frac{1}{4\pi}\int_{\Gamma'} \frac{\tilde{N}(\xi)}{\xi - \zeta}\,\mathrm{d}\xi + \frac{1}{4\pi}\int_{\Gamma\cup\Gamma'} \frac{1}{\xi-\zeta}\,\mathrm{d}\xi, \tag{3.43}$$

$$\Xi(\zeta) = \gamma^{-1}\left[\frac{1}{2\pi i}\int_{\Gamma'} \frac{\tilde{\tau}_b(\xi)}{\xi-\zeta} - \frac{1}{2\pi i}\int_{\Gamma\cup\Gamma'} \frac{1}{\xi-\zeta}\,\mathrm{d}\xi\right], \tag{3.44}$$

where integrals over Γ and Γ' are taken (here and in what follows) as the unit circle is traversed in the anticlockwise direction. Note that the second integral on the right-hand side of each equation may be recognized as $\int_{\Gamma\cup\Gamma'}(\xi-\zeta)^{-1}\,\mathrm{d}\xi = 2\pi i$ if ζ is inside the unit circle, $\int_{\Gamma\cup\Gamma'}(\xi-\zeta)^{-1}\,\mathrm{d}\xi = 0$ for ζ outside the unit circle.

It remains to ensure that Ω and Ξ satisfy $(3.40)_{1,3}$. Using the fact that $\tilde{N}$ is real, while $\bar{\xi} = 1/\xi$ and $\overline{\mathrm{d}\xi} = -1/\xi^2\,\mathrm{d}\xi$, it follows after some manipulation that

$$\overline{\Omega(1/\bar{\zeta})} = -\frac{1}{4\pi}\int_{\Gamma'} \tilde{N}(\xi)\left[\frac{1}{\xi-\zeta} - \frac{1}{\xi}\right]\mathrm{d}\xi + \frac{1}{4\pi}\int_{\Gamma'\cup\Gamma}\left[\frac{1}{\xi-\zeta} - \frac{1}{\xi}\right]\mathrm{d}\xi, \tag{3.45}$$

with a similar expression for $\overline{\Xi(1/\bar{\zeta})}$. In order to satisfy $\Omega(\zeta) = \overline{\Omega(1/\bar{\zeta})}$ and $\Xi(\zeta) = -\overline{\Xi(1/\bar{\zeta})}$, we therefore require

$$\int_{\Gamma'} \frac{\tilde{N}(\xi)}{\xi}\,\mathrm{d}\xi - \int_{\Gamma\cup\Gamma'} \frac{\mathrm{d}\xi}{\xi} = 0, \qquad \int_{\Gamma'} \frac{\tilde{\tau}_b(\xi)}{\xi}\,\mathrm{d}\xi - \int_{\Gamma\cup\Gamma'} \frac{\mathrm{d}\xi}{\xi} = 0. \tag{3.46}$$

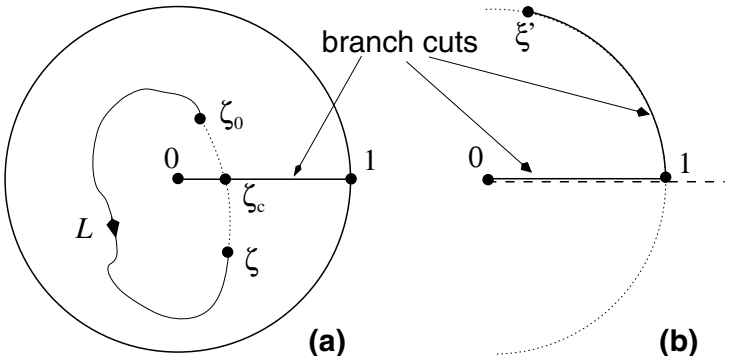


FIG. 3.1. *The contour L in* (3.48) *is shown in panel* (a)*, where solid lines indicate branch cuts in ω. The branch cuts for* $\log(\xi'/\zeta - 1)$ *(solid line) and* $\log(\zeta)$ *(dashed line) are shown in panel* (b).

In real terms, these equations are simply

$$\frac{1}{a}\int_0^b N(\eta)\,\mathrm{d}\eta = 1, \qquad \frac{1}{a}\int_0^b \tau_b(\eta)\,\mathrm{d}\eta = 1, \tag{3.47}$$

which physically state that the mean shear stress at the ice-sediment interface is the far-field shear stress, and that mean effective pressure is fixed by hydrostatic ice overburden and by porewater pressure in the bed (the difference between the two being scaled to unity). With the particular prescription of τ_b in (3.8), the two equations in (3.47) are identical, and we are left with the single constraint $(3.47)_1$, which states that mean effective pressure at the bed is fixed at unity.

The expressions for Ω and Ξ in (3.43) and (3.44) as well as the solvability constraints in (3.47) contain the unknown functions N and τ_b, which depend on sediment thickness h through the constitutive relations (3.6) and (3.8) as well as through (3.12), while h conversely depends on N and τ_b through (3.35). Moreover, the limits of integration appearing in (3.43) and (3.44) are not known a priori, as we have yet to determine the position of the contact point $\eta = b$. In the next section, we show how h (and hence N and τ_b) can be calculated from an eigenvalue problem arising from (3.35), which also determines the pattern speed V as its eigenvalue. Additionally, the continuity requirements in (3.13) allow us to derive an integral constraint which fixes the position of the contact point b.

3.3. Reduction to an eigenvalue problem. We exploit (3.35), (3.32), and (3.43) as well as (3.6), (3.12), (3.13) and the inequality constraints in (3.4) and (3.5) in order to obtain integral equations for h and h_C. Our first task is to calculate ω, which determines the bed slopes h' and h'_C through (3.35). For ζ_0 and ζ in the open unit disk of the ζ-plane cut along the nonnegative half of the real axis, where ω has a branch cut, we have from (3.32)

$$\omega(\zeta) - \omega(\zeta_0) = \frac{a}{i2\pi}\int_L \frac{\Omega(\zeta')}{\zeta'}\,\mathrm{d}\zeta', \tag{3.48}$$

where primes on ζ' indicate a dummy variable, not differentiation. L is any arc connecting ζ_0 to ζ such that L lies entirely in the open unit disk and does not cross the branch cut in ω (see Figure 3.1). It follows from (3.48) that ω is in fact continuous and therefore analytic across that branch cut: Let ζ_c lie on the branch cut, and let

ζ_0 and ζ approach ζ_c from the first and fourth quadrants, respectively. In the limit, L becomes a closed contour encircling the origin (Figure 3.1), and from the residue theorem we have

$$\omega(\zeta) - \omega(\zeta_0) \to a\Omega(0) = 0, \tag{3.49}$$

so $\omega(\zeta_c)$ can be assigned a unique limiting value regardless of which side the real axis is approached from. As a corollary, we can in fact allow the arc L above to cross the positive half of the real axis (which later allows us to establish that $\omega(\zeta)$ has a unique limit when $\zeta \to 1$ from inside the unit circle).

Using (3.43), we can evaluate the integral in (3.48) explicitly:

$$\begin{aligned}\omega(\zeta) - \omega(\zeta_0) = &-\frac{a}{i8\pi^2}\int_{\Gamma'} \tilde{N}(\xi')\left[\log\left(\frac{\xi'}{\zeta_0} - 1\right) - \log\left(\frac{\xi'}{\zeta} - 1\right)\right]\frac{\mathrm{d}\xi'}{\xi'} \\ &+ \frac{a}{4\pi}[\log(\zeta) - \log(\zeta_0)],\end{aligned} \tag{3.50}$$

where for definiteness $\log(\xi'/\zeta_0 - 1)$ and $\log(\xi'/\zeta - 1)$ denote a branch of the logarithm which has a branch cut as indicated in Figure 3.1. Similarly, $\log(\zeta_0)$ and $\log(\zeta)$ denote a branch which has a branch cut along the positive half of the real axis.

Next, we let $\zeta \to \xi$ and $\zeta_0 \to 1$ from inside the unit circle. It is easy to show from (3.48), Cauchy's theorem, and the continuity properties of Ω [14, pp. 53–55 and Chapter 4] that the limit $\omega^+(\xi)$ exists and is continuous as a function of ξ for all ξ on the unit circle. Importantly, this result holds true at both endpoints of Γ', where $\xi = 1$ and $\xi = \exp(i2\pi b/a)$ (with continuity at $\xi = 1$ resulting from the continuity of ω across the real axis). Using (3.35) and (3.50) and taking care with the branches of the logarithms involved, we find after some elementary manipulations that

$$\begin{aligned}-2\mathrm{Re}(\omega^+(\xi) - \omega^+(1)) &= -\frac{1}{2\pi}\int_0^b N(\eta')\log\left|\frac{\sin(\pi(\eta' - \eta)/a)}{\sin(\pi\eta'/a)}\right|\mathrm{d}\eta' \\ &= \begin{cases} (U - V)(h'(\eta) - h'(0^+)), & \eta \in \hat{C}', \\ (U - V)(h'_C(\eta) - h'(0^+)), & \eta \in \hat{C}, \end{cases}\end{aligned} \tag{3.51}$$

where $\xi = \exp(i2\pi\eta/a)$ as before and $\log(\cdot)$ is the ordinary logarithm defined for positive real numbers.

To proceed further, we require $h'(0^+)$. Since ω^+ is continuous at $\xi = 1$, we conclude from (3.35) that $h'_C(a^-) = h'(0^+)$, provided $V \neq U$. In other words, the continuity of the y-component of velocity precludes any breaks in the slope of the lower boundary of the ice. The local behavior of h and h_C near the contact points $\eta = 0$, $\eta = a$ then requires that $h'_C(a^-) = h'(0^+) = 0$ if the inequality constraints $(3.4)_5$ and $(3.5)_5$ on N and h_C are to be satisfied close to the contact points. To see this, note that the cavity roof is above the sediment surface and we have $h_C(\eta) > h(\eta) = 0$ for $b < \eta < a$, while the cavity roof recontacts the bed at $\eta = a$, so that $h_C(a^-) = 0$ and hence $h'(0^+) = h'_C(a^-) \leq 0$. Meanwhile, $h(\eta) \geq 0$ in contact areas $0 < \eta < b$ (as flux $Vh = q = N \geq 0$ and $V > 0$), and sediment thickness is continuous at the downstream cavity endpoint, so that $h(0) = 0$, which implies $h'_C(a^-) = h'(0^+) \geq 0$. These two inequalities on $h'(0^+)$ can be satisfied simultaneously only if bed and cavity roof slope vanish at the downstream cavity endpoint, $h'(0^+) = h'_C(a^-) = 0$.

An integral equation for h is now straightforward to obtain by integrating (3.51)

once more and using $h(0^+) = 0$:

$$\begin{aligned}(U-V)h(\eta) &= \int_0^\eta (U-V)h'(\eta')\,\mathrm{d}\eta' \\ &= -\frac{1}{2\pi}\int_0^\eta \int_0^b N(\eta')\log\left|\frac{\sin(\pi(\eta'-\eta'')/a)}{\sin(\pi\eta'/a)}\right|\,\mathrm{d}\eta'\,\mathrm{d}\eta'' \\ &= -\frac{1}{2\pi}\int_0^b \left[\int_0^\eta \log\left|\frac{\sin(\pi(\eta'-\eta'')/a)}{\sin(\pi\eta'/a)}\right|\,\mathrm{d}\eta''\right] N(\eta')\,\mathrm{d}\eta' \qquad (3.52)\end{aligned}$$

for $\eta \in (0, b)$. As $h_C(a^-) = 0$ from (3.13), the cavity roof similarly satisfies

$$\begin{aligned}(U-V)h_C(\eta) &= -\int_\eta^a (U-V)h_C(\eta')\,\mathrm{d}\eta' \\ &= \frac{1}{2\pi}\int_0^b \left[\int_\eta^a \log\left|\frac{\sin(\pi(\eta'-\eta'')/a)}{\sin(\pi\eta'/a)}\right|\,\mathrm{d}\eta''\right] N(\eta')\,\mathrm{d}\eta' \qquad (3.53)\end{aligned}$$

for $\eta \in (b, a)$. Solutions of these integral equations automatically satisfy $(3.13)_1$. It remains to ensure that $h_C(b^+) = h(b^-)$. Using (3.52) and (3.53), this can be written in the form

$$\frac{1}{2\pi}\int_0^b \left[\int_0^a \log\left|\frac{\sin(\pi(\eta'-\eta'')/a)}{\sin(\pi\eta'/a)}\right|\,\mathrm{d}\eta''\right] N(\eta')\,\mathrm{d}\eta' = 0. \qquad (3.54)$$

This equation ensures that there is no discontinuity in the ice surface at $\eta = b$, but it generally requires a discontinuity in h (in the sense that (3.54) does not ensure $h(b^-) = 0$). This justifies our statement in section 2 that we cannot generally expect sediment thickness to be continuous across all contact points; this is at least true for traveling wave solutions.

Equation (3.54) allows the integrals above to be simplified somewhat. The kernel on the left-hand side of (3.54) can be rewritten as

$$\begin{aligned}&\int_0^a \log\left|\frac{\sin(\pi(\eta'-\eta'')/a)}{\sin(\pi\eta'/a)}\right|\,\mathrm{d}\eta'' \\ &\quad = \int_0^a \frac{1}{2}\left\{\log\left[4\sin^2\left(\frac{\pi(\eta'-\eta'')}{a}\right)\right] - \log\left[4\sin^2\left(\frac{\pi\eta'}{a}\right)\right]\right\}\,\mathrm{d}\eta''. \qquad (3.55)\end{aligned}$$

But, as shown in the appendix,

$$\int_0^a \log\left[4\sin^2(\pi(\eta'-\eta'')/a)\right]\,\mathrm{d}\eta'' = \int_0^a \log\left[4\sin^2(\pi\eta''/a)\right]\,\mathrm{d}\eta'' = 0, \qquad (3.56)$$

and (3.54) becomes more simply

$$\int_0^b \log\left[4\sin^2(\pi\eta/a)\right] N(\eta)\,\mathrm{d}\eta = 0. \qquad (3.57)$$

Similarly rewriting the integral kernel in (3.52) and (3.53) and using (3.56) and (3.57) yields the integral equations

$$(U-V)h(\eta) = -\frac{1}{4\pi}\int_0^b \left\{\int_0^\eta \log\left[4\sin^2\left(\frac{\pi(\eta'-\eta'')}{a}\right)\right]\,\mathrm{d}\eta''\right\} N(\eta')\,\mathrm{d}\eta' \qquad (3.58)$$

for $\eta \in (0, b)$, and

$$(U-V)h_C(\eta) = -\frac{1}{4\pi}\int_0^b \left\{ \int_0^\eta \log\left[4\sin^2\left(\frac{\pi(\eta'-\eta'')}{a}\right)\right] \mathrm{d}\eta'' \right\} N(\eta')\,\mathrm{d}\eta' \tag{3.59}$$

for $\eta \in (b, a)$.

On writing $N = q = Vh$ and setting $U = 1$ from (3.6) and (3.12), the structure of the problem finally becomes apparent: For a given contact point position b, $h(\eta)$ satisfies the eigenvalue problem

$$\lambda h(\eta) + \int_0^b k(\eta, \eta')h(\eta')\,\mathrm{d}\eta' = 0 \tag{3.60}$$

for $\eta \in (0, b)$, where the kernel k is given by

$$k(\eta, \eta') = \frac{1}{4\pi}\int_0^\eta \log\left(4\sin^2\left[\pi(\eta'-\eta'')/a\right]\right)\,\mathrm{d}\eta'', \tag{3.61}$$

and the real eigenvalue λ is a proxy for pattern velocity V, $\lambda = (1-V)/V$. The contact point b is constrained by (3.54), which reads

$$\int_0^b \log\left[4\sin^2(\pi\eta/a)\right] h(\eta)\,\mathrm{d}\eta = 0. \tag{3.62}$$

Finally, the eigenfunction $h(\eta)$ satisfies the normalization condition (3.47), which becomes

$$\frac{1}{a}\int_0^b h(\eta)\,\mathrm{d}\eta = \lambda + 1. \tag{3.63}$$

Once h, λ, and b have been found, the cavity roof shape $h_C(\eta)$ for $\eta \in (b, a)$ can be calculated from (3.53):

$$\lambda h_C(\eta) = -\int_0^b k(\eta, \eta')h(\eta')\,\mathrm{d}\eta'. \tag{3.64}$$

As mentioned previously, we allow only positive pattern speeds $V > 0$, so $\lambda > -1$, and the mean of h is positive by (3.63). A solution h must further satisfy the stronger pointwise constraint $h(\eta) \geq 0$ for $\eta \in (0, b)$, and similarly $h_C(\eta) > 0$ for $\eta \in (b, a)$. It is by no means obvious that this will be the case: We have so far employed the inequality constraints on h and h_C only locally in order to determine the slopes of h and h_C at the contact points $\eta = 0$, $\eta = a$. Compliance with these constraints must therefore be checked *a posteriori* once a solution has been found. Further, our solution of the Hilbert problems (3.33) and (3.42) requires $N(\eta)$ and $\tau_b(\eta)$ to be Hölder continuous on C'; that is, $h(\eta)$ must be Hölder continuous on $(0, b)$ and (to make sense of the jump conditions (3.13)) continuous up to $\eta = 0$ and $\eta = b$ from the left and right, respectively. Moreover, $h(\eta)$ must satisfy the original integrodifferential equation (3.51), rather than simply the integrated version (3.52). In the appendix, we show that any continuous solution $h \in C([0, b])$ of the eigenvalue problem (3.60)–(3.62) does in fact satisfy this equation and is in $C^1([0, b])$, which takes care of the Hölder continuity of h. Hence it is sufficient to look for continuous $h \in C([0, b])$.

3.4. Numerical method. In order to eliminate the arbitrary wavelength a from the eigenvalue problem, we define

$$X = \eta/a, \qquad B = b/a, \qquad \mu = \lambda/a^2 = (1-V)/(a^2V),$$

$$S(X) = h(\eta), \qquad S_C(X) = h_C(\eta). \tag{3.65}$$

The equations we wish to solve are then

$$\mu S(X) + \frac{1}{4\pi}\int_0^B \left[\int_0^X \log\left[4\sin^2(\pi(X'-X''))\right]\,\mathrm{d}X''\right] S(X')\,\mathrm{d}X' = 0, \tag{3.66}$$

$$\int_0^B \log\left[4\sin^2(\pi X)\right] S(X)\,\mathrm{d}X = 0, \tag{3.67}$$

where the eigenfunction S is to be normalized as

$$\int_0^B S(X)\,\mathrm{d}X = 1 + a^2\mu. \tag{3.68}$$

We are not interested in calculating the entire spectrum of the integral operator in (3.66), but merely seek real eigenvalues $\mu > -1/a^2$. To this end, we can exploit the structure of the integral operator by rewriting (3.66) in the form

$$\begin{aligned}\mu S(X) + \frac{1}{4\pi}\int_0^B \left[\int_0^{X-X'} \log\left[4\sin^2(\pi X'')\right]\,\mathrm{d}X''\right] S(X')\,\mathrm{d}X' \\ = -\frac{1}{4\pi}\int_0^B \left[\int_0^{X'} \log\left[4\sin^2(\pi X'')\right]\,\mathrm{d}X''\right] S(X')\,\mathrm{d}X'.\end{aligned} \tag{3.69}$$

The right-hand side of this equation is simply a constant, while the kernel of the convolution-type integral operator on the left-hand side is antisymmetric and therefore has purely imaginary eigenvalues. It follows that the constant on the right-hand side cannot vanish unless we have the trivial solution $S \equiv 0$, which is, however, precluded by the normalization condition (3.68). If we dispense temporarily with this normalization condition by rescaling S—which we are at liberty to do because (3.66) and (3.67) are homogeneous in S—we can therefore fix the constant on the right-hand side of (3.69) at unity. Denoting this rescaled version of S by $\tilde{S}$, we obtain the pair of equations

$$\mu\tilde{S}(X) + \frac{1}{4\pi}\int_0^B \left[\int_0^{X-X'} \log\left[4\sin^2(\pi X'')\right]\,\mathrm{d}X''\right] \tilde{S}(X')\,\mathrm{d}X' = 1, \tag{3.70}$$

$$-\frac{1}{4\pi}\int_0^B \left[\int_0^{X'} \log\left[4\sin^2(\pi X'')\right]\,\mathrm{d}X''\right] \tilde{S}(X')\,\mathrm{d}X' = 1. \tag{3.71}$$

The advantage of (3.70) is precisely that the kernel of the integral operator on the left-hand side is antisymmetric and hence has purely imaginary eigenvalues. By the Fredholm alternative, (3.70) has a unique solution $\tilde{S}(X;\mu,B) \in C([0,B])$ for every

real nonzero μ and every $B \in (0,1]$. We denote this solution by $\tilde{S}(X;\mu,B)$. In terms of $\tilde{S}(X;\mu,B)$, equations (3.71) and (3.67) can then be written in the form

$$f_1(\mu,B) := \frac{1}{4\pi}\int_0^B \left[\int_0^X \log\left[4\sin^2(\pi X')\right]\,\mathrm{d}X'\right]\tilde{S}(X;\mu,B)\,\mathrm{d}X + 1 = 0, \tag{3.72}$$

$$f_2(\mu,B) := \int_0^B \log\left[4\sin^2(\pi X)\right]\tilde{S}(X;\mu,B)\,\mathrm{d}X = 0. \tag{3.73}$$

The task of finding μ and B can therefore be reduced to solving two nonlinear equations, which must be done numerically. Here we use a backtracking line-search modification of Newton's method [4], where the Jacobian is approximated by finite differences.

In order to evaluate the functions $f_1(\mu,B)$ and $f_2(\mu,B)$, $\tilde{S}(X;\mu,B)$ must be calculated from (3.70). We use a degenerate kernel approximation [2, Chapter 2]: As shown in the appendix, the kernel

$$K(X-X') = \frac{1}{4\pi}\int_0^{X-X'} \log\left[4\sin^2(\pi X'')\right]\,\mathrm{d}X'' \tag{3.74}$$

can be approximated uniformly by the truncated Fourier series

$$K_{n_0}(X-X') = \sum_{n=-n_0,\,n\neq 0}^{n_0} \frac{i\exp(i2n\pi X)\exp(-i2n\pi X')}{8\pi^2|n|n} \tag{3.75}$$

as $n_0 \to \infty$. Replacing $K(X-X')$ by $K_{n_0}(X-X')$, the solution of (3.70) follows the standard method for degenerate kernels.

4. Results and discussion. Regardless of the initial guess for μ and B, only a single solution was found numerically, with $\mu = 2.971\times 10^{-3}$ and $B = 0.2285$. Visual inspection of the surfaces generated by f_1 and f_2 also suggests that this solution is unique. In Figure 4.1, we plot the corresponding shape of $S(X)$ and $S_C(X)$, normalized so that $\int_0^B S(X)\,\mathrm{d}X = 1$ (formally, this is (3.68) with $a = 0$, that is, the short-wave limit). The sediment in the traveling wave is confined to a relatively short wedge upstream of an extended cavity, and the vanishing sediment surface and cavity roof slopes at the contact points $X = 0$, 1 are clearly visible. Moreover, the solution appears to satisfy the constraints $S(X) \geq 0$ for $X \in (0,B)$, $S_C(X) > 0$ for $X \in (B,1)$. Note that a solution $S(X)$ for a given wavelength $a \neq 0$ can be obtained from that plotted in Figure 4.1 simply by multiplying it by $1 + a^2\mu$. This ensures that the normalization condition (3.68) is satisfied. Since the amplitude of S varies with wavelength as $1+a^2\mu$, we see that long waves are also taller than short ones. Furthermore, the pattern speed V can be calculated as $V = 1/(1+a^2\mu)$, which states that pattern speed is inversely proportional to amplitude and hence that shorter waves travel faster than longer ones. Moreover, since $\mu > 0$, the pattern speed V is always less than the ice velocity U. If we take a to be given as the fastest growing wavelength in the original instability model in section 2, then $a = 2\pi[2/\sqrt{3}]^{1/2} = 6.752$ and $V = 0.881$. These traveling waves are advected downstream at 88% of the velocity of ice at the bed, and their amplitude is 1.135 times that shown in Figure 4.1.

The existence of a traveling wave solution suggests that cavity formation may be sufficient to lead to bounded growth in the instability mechanism proposed by Hindmarsh [11] and Fowler [8]. However, it also poses some interesting open questions

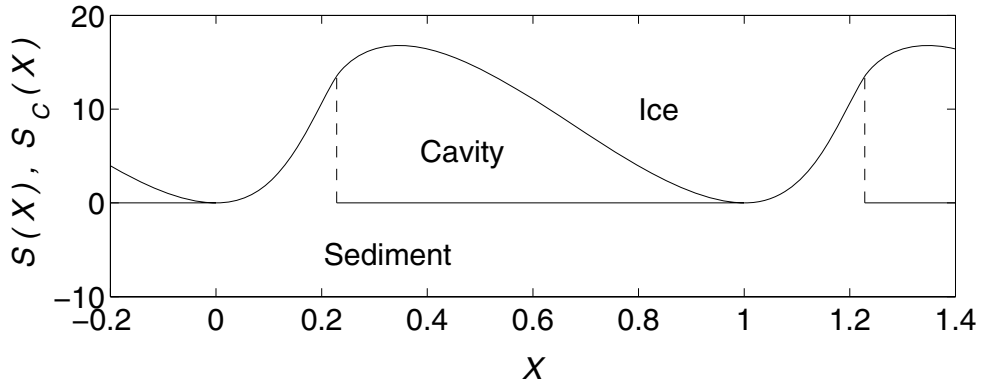


FIG. 4.1. *Solution $S(X)$ for sediment surface elevation and $S_C(X)$ for cavity roof elevation, normalized so that $\int_0^B S(X)\,\mathrm{d}X = 1$ and extended periodically. The sediment shock at $X = B$ is shown as a dashed vertical line.*

regarding the dynamical problem of section 2: First, are the traveling waves described here stable to small perturbation, and if so, does their stability depend on their wavelength? Second, how do waves of different wavelengths interact? Do shorter waves merge with larger ones, and does wavelength coarsening occur? The first of these questions may be amenable to a complex variable approach of the type used above, using an appropriate linearization. The second is much harder and requires an understanding of the fully nonlinear time-dependent problem.

Physically, our results also pose some additional questions. Following [11, 8], our model assumes that porewater pressure in the sediment, and hence mean effective pressure, are fixed by a pre-existing subglacial drainage system, most likely taking the form of channels at the ice-bed interface [15]. However, once cavities have formed, it is likely that they will serve as both storage volume and conduits for water flowing beneath the glacier, and that drainage through these cavities will be instrumental in controlling water pressures at the bed. The simplest way to understand how this type of drainage might work is by considering cavities as a kind of macroscopic pore space, and to consider drainage through cavities at the bed as being a two-dimensional analogue of drainage in ordinary porous media, giving rise to a Biot-type problem [7, 10] for water pressure on an outer length scale associated with the length of the glacier as a whole (which is assumed to be large compared with the instability length scale considered in the model studied in this paper). The problem with applying this approach here is that the size of cavities increases with mean effective pressure (scaled to unity in the dimensionless model), as can be shown from the scalings used in [19]. Specifically, the dimensional scale for h is

$$[h] = \frac{[N]}{(n-2)(1-\phi)(\rho_s - \rho_w)g},$$

where $[N]$ is mean effective pressure at the bed ("mean" being an average taken over the cavity length scale), n is an exponent in the assumed power-law rheology for sediment (which gives rise to the constitutive relations for shear stress τ_b and flux q in (2.6) for $n \gg 1$; see [19]), ϕ is the porosity of the sediment, ρ_s and ρ_w are the densities of sediment grains and of water, and g is acceleration due to gravity. Hence $[h]$ increases linearly with effective pressure, and it follows that cavity size *decreases*

with water pressure. This in turn means that the advocated Biot-type drainage model takes the form of a backward diffusion problem and is therefore ill-posed.

It is unclear to what extent this result is due to the particular rheological model employed for subglacial sediment, and an obvious avenue for further research is to consider alternative prescriptions for τ_b and q from those introduced in (2.6). The eigenvalue problem (3.58) combined with constraints of the form (3.47) and (3.54) generalizes relatively easily to other forms of τ_b and q, though the resulting nonlinear eigenvalue problem is considerably more complicated [16], and we leave a solution as an open problem.

Appendix. Smoothness of solutions. It can be shown in the usual way from the Arzelà–Ascoli theorem that the integral operator in (3.70) is compact on $C([0,B])$, and the antisymmetry of the kernel further ensures that all its eigenvalues are purely imaginary. By the Fredholm alternative [3, Chapter 7.5], the integral equation (3.70) therefore has a unique solution in $C([0,B])$ for every real μ. In what follows, we will show that this solution is in fact in $C^1([0,B])$ and satisfies the integrodifferential equation (3.51). In the process, we will also prove (3.56) and show that the degenerate kernel approximation (3.75) converges uniformly to K in the limit $n_0 \to \infty$.

At issue is thus whether a solution of (3.70) satisfies (3.51), which in view of (3.54), (3.6), (3.12), and the rescaling in (3.65) may be rewritten in the form

$$-\frac{1}{4\pi}\int_0^b \log\left(4\sin^2[\pi(X-X')]\right)\tilde{S}(X)\,\mathrm{d}X' = \mu\tilde{S}'(X) \tag{A.1}$$

for $X \in (0,B)$. If $\tilde{S} \in C([0,B])$ satisfies (A.1), it follows immediately from the properties of convolution integrals that $\tilde{S}' \in C([0,B])$, and $\tilde{S} \in C^1([0,B])$, as required. Equation (A.1) can be obtained by differentiating (3.70) (noting that the right-hand side is simply a constant) and by exchanging the order of differentiation and integration on the integral term. In order to prove that integration and differentiation do commute—which is not obvious because the integrand in (A.1) has a singularity—we approximate the integrand by a sequence of bounded integrands.

A.1. Degenerate kernel approximation. The power series

$$\sum_{n=1}^{\infty}\frac{\zeta^{n-1}}{n} = -\frac{\log(1-\zeta)}{\zeta}$$

has radius of convergence one, and therefore converges everywhere inside the unit circle in the complex ζ-plane. The branch of the logarithm must be continuous on the open unit disk with $\log(1)=0$. Note that the singularity at the origin is removable: We can assign $-\frac{1}{\zeta}\log(1-\zeta)$ its limiting value of 1. The series also converges pointwise on the unit circle except at $\zeta = 1$ [1, p. 409]. For ξ on the unit circle and $r \in (0,1)$, we have

$$\left|\sum_{n=1}^{n_0}\frac{r^{n-1}\xi^n}{n}\right| \le \sum_{n=1}^{n_0}\left|\frac{r^{n-1}\xi^n}{n}\right| = \sum_{n=1}^{n_0}\frac{r^{n-1}}{n} < \frac{\log(1-r)}{r} \tag{A.2}$$

for any finite n_0. Since $\frac{1}{r}\log(1-r)$ is integrable over $r \in (0,1)$, so is

$$\sum_{n=1}^{\infty}\frac{r^{n-1}\xi^n}{n} = -\frac{\log(1-r\xi)}{r\xi}\xi,$$

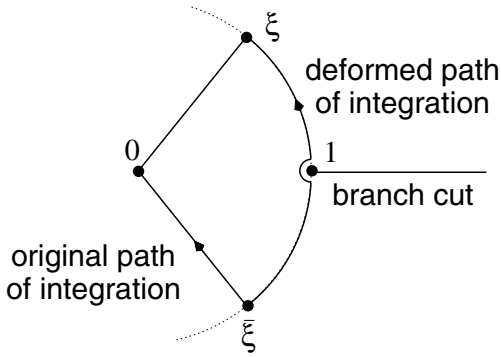


FIG. A.1. *Integration paths in the degenerate kernel approximation of* $K(x)$.

and the order of summation and integration may be interchanged by the dominated convergence theorem, i.e.,

$$\int_0^1 -\frac{\log(1-r\xi)}{r\xi}\xi\,\mathrm{d}r = \sum_{n=1}^{\infty}\int_0^1 \frac{(r\xi)^{n-1}}{n}\xi\,\mathrm{d}r = \sum_{n=1}^{\infty}\frac{\xi^n}{n^2}. \tag{A.3}$$

Similarly,

$$\int_0^1 -\frac{\log(1-r\bar{\xi})}{r\bar{\xi}}\bar{\xi}\,\mathrm{d}r = \sum_{n=1}^{\infty}\frac{\bar{\xi}^n}{n^2}. \tag{A.4}$$

Subtracting the last two expressions and recognizing that $\bar{\xi} = 1/\xi$ yields

$$\sum_{n=-\infty,\, n\neq 0}^{\infty}\frac{\xi^n}{n|n|} = \int_0^1 -\frac{\log(1-r\xi)}{r\xi}\xi\,\mathrm{d}r - \int_0^1 -\frac{\log(1-r\bar{\xi})}{r\bar{\xi}}\bar{\xi}\,\mathrm{d}r. \tag{A.5}$$

However, the right-hand side is just the integral $\int -\frac{1}{\zeta}\log(1-\zeta)\,\mathrm{d}\zeta$ taken along the radial path from $\bar{\xi}$ to ξ via the origin (see Figure A.1). Since the integrand $-\frac{1}{\zeta}\log(1-\zeta)$ can be made holomorphic in the complex plane cut along the interval $[1,\infty)$ on the real line, the curve along which the integral is taken can be deformed to lie on the unit circle, with a small indentation at the branch point $\zeta = 1$ (Figure A.1). The indentation does not contribute to the integral along the deformed curve in the limit where the radius of the indentation tends to zero. Setting $\xi = \exp(i2\pi X)$, the integral may thus be expressed as

$$\begin{aligned}\sum_{n=-\infty,\, n\neq 0}^{\infty}\frac{\xi^n}{n|n|} &= -i2\pi\int_{-X}^{X}\log\left[1-\exp(i2\pi X')\right]\,\mathrm{d}X' \\ &= -i2\pi\int_{0}^{X}\log\left[1-\exp(i2\pi X')\right] + \log\left[1-\exp(-i2\pi X')\right]\,\mathrm{d}X' \\ &= -i2\pi\int_{0}^{X}\log\left[4\sin^2(\pi X')\right]\,\mathrm{d}X'. \end{aligned} \tag{A.6}$$

Consequently, the kernel K defined by (3.74) may be written as

$$K(X) = \frac{1}{4\pi}\int_0^X \log\left[4\sin^2(\pi X')\right]\,\mathrm{d}X' = \sum_{n=-\infty,\, n\neq 0}^{\infty} \frac{i\xi^n}{8\pi^2 n|n|}. \tag{A.7}$$

The argument above has shown that the series on the right converges pointwise to the integral in the middle. By the Weierstrass M-test, it also converges uniformly; that is, $K_{n_0}(X)$ defined in (3.75) converges to $K(X)$ in the $C([0,B])$ norm as $n_0 \to \infty$.

Lastly, it follows immediately that $K(1) = 0$, and (3.56) holds.

A.2. Differentiation under the integral sign. We can use the series representation (A.7) of the kernel K to show that (A.1) holds. Specifically, let

$$\mu\tilde{S}_{n_0}(X) = -\int_0^B K_{n_0}(X - X')\tilde{S}(X')\,\mathrm{d}X' + 1, \tag{A.8}$$

where K_{n_0} is defined in (3.75). It is easy to see that $\tilde{S}_{n_0}$ converges to $\tilde{S}$ in the $C([0,B])$ norm. Since K'_{n_0} is continuous and hence bounded, we can further differentiate directly,

$$\mu\tilde{S}'_{n_0}(X) = -\int_0^B K'_{n_0}(X - X')\tilde{S}(X')\,\mathrm{d}X', \tag{A.9}$$

and $\tilde{S}_{n_0} \in C^1([0,B])$. But $K'_{n_0}(X) = -\sum_{n=-n_0,n\neq 0}^{n_0} \exp(i2\pi X)/(4\pi n)$ converges to $K'(X)$ almost everywhere in $[0,1]$ by the results of section A.1. Moreover, because $\sum_{n=1}^{\infty} 1/|n|^2 < \infty$, $K'_{n_0}(X)$ also converges as a Fourier series to $K'(X)$ in the $L^2([0,1])$-norm, and hence in the $L^1([0,1])$-norm. Hence, for $B \in [0,1]$,

$$\begin{aligned}\sup_{x\in[0,B]} &\left|\int_0^B K'_{n_0}(X - X')\tilde{S}(X')\,\mathrm{d}X' - \int_0^B K'(X - X')\tilde{S}(X')\,\mathrm{d}X'\right| \\ &\leq \sup_{X\in[0,B]} |\tilde{S}(X)| \int_0^1 \left|K'_{n_0}(X') - K'(X')\right|\,\mathrm{d}X' \to 0\end{aligned} \tag{A.10}$$

as $n_0 \to \infty$. By the completeness of $C^1([0,B])$, we therefore have

$$\mu\tilde{S}'(X) = \mu\frac{\mathrm{d}}{\mathrm{d}X}\lim_{n_0\to\infty}\tilde{S}_{n_0}(X) = -\int_0^B K'(X - X')\tilde{S}(X')\,\mathrm{d}X', \tag{A.11}$$

which is (A.1).

REFERENCES

[1] T. M. Apostol, *Calculus*, 2nd ed., Blaisdell Publishing, Waltham, MA, 1969.

[2] K. E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge Monogr. Appl. Comput. Math. 88, Cambridge University Press, Cambridge, UK, 1997.

[3] E. W. Cheney, *Analysis for Applied Mathematics*, Grad. Texts in Math. 208, Springer-Verlag, New York, 2001.

[4] J. E. Dennis, Jr., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Classics in Appl. Math. 16, SIAM, Philadelphia, 1996.

[5] A. H. England, *Complex Variable Methods in Elasticity*, J. Wiley & Sons, London, 1971.

[6] A. C. Fowler, *A sliding law for glaciers of constant viscosity in the presence of subglacial cavitation*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 407 (1986), pp. 147–170.

[7] A. C. Fowler, *Sliding with cavity formation*, J. Glaciol., 33 (1987), pp. 255–267.
[8] A. C. Fowler, *An instability mechanism for drumlin formation*, in Deformation of Glacial Materials, A. Maltman, M. J. Hambrey, and B. Hubbard, eds., Spec. Pub. Geol. Soc. 176, The Geological Society, London, 2000, pp. 307–319.
[9] A. C. Fowler, *Evolution equations for dunes and drumlins*, Rev. P. Acad. Cien. Ser. A Mat., 96 (2002), pp. 377–387.
[10] A. C. Fowler and C. G. Noon, *Mathematical models of compaction, consolidation and regional groundwater flow*, Geophys. J. Int., 136 (1999), pp. 251–260.
[11] R. C. A. Hindmarsh, *The stability of a viscous till sheet coupled with ice flow, considered at wavelengths less than the ice thickness*, J. Glaciol., 44 (1998), pp. 285–292.
[12] B. Kamb, *Basal zone of the West Antarctic ice streams and its role in lubrication of their rapid motion*, in The West Antarctic Ice Sheet: Behaviour and Environment, R. B. Alley and R. A. Binschadler, eds., American Geophysical Union, Washington, DC, 2001, pp. 157–199.
[13] N. Kikuchi and J. T. Oden, *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*, SIAM Stud. Appl. Math. 8, SIAM, Philadelphia, 1988.
[14] N. I. Muskhelishvili, *Singular Integral Equations*, Dover Publications, New York, 1992.
[15] F. S. L. Ng, *Mathematical Modelling of Subglacial Drainage and Erosion*, D. Phil. Thesis, Mathematical Institute, Oxford University, 1998; available online at http://www.maths.ox.ac.uk/research/theses/.
[16] C. Schoof, *Mathematical Models of Glacier Sliding and Drumlin Formation*, D. Phil. Thesis, Mathematical Institute, Oxford University, 2002; available online at http://www.maths.ox.ac.uk/research/theses/.
[17] C. Schoof, *The effect of cavitation on glacier sliding*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 461 (2005), pp. 609–627.
[18] C. Schoof, *A variational approach to ice-stream flow*, J. Fluid Mech., 556 (2006), pp. 227–251.
[19] C. Schoof, *Pressure-dependent viscosity and interfacial instability in coupled ice-sediment flow*, J. Fluid Mech., 570 (2007), pp. 227–252.
[20] S. Tulaczyk, *Ice sliding over weak, fine-grained till: Dependence of ice-till interactions on till granulometry*, Spec. Paper Geol. Soc. Am., 337 (1999), pp. 159–177.

SIAM J. APPL. MATH.
Vol. 67, No. 6, pp. 1654–1674

ATTRACTORS IN CONFINED SOURCE PROBLEMS FOR COUPLED NONLINEAR DIFFUSION*

D. V. STRUNIN†

Abstract. In processes driven by nonlinear diffusion, a signal from a concentrated source is confined in a finite region. Such solutions can be sought in the form of power series in a spatial coordinate. We use this approach in problems involving coupled agents. To test the method, we consider a single equation with (a) linear and (b) quadratic diffusivity in order to recover the known results. The original set of PDEs is converted into a dynamical system with respect to the time-dependent series coefficients. As an application we consider an expansion of a free turbulent jet. Some example trajectories from the respective dynamical system are presented. The structure of the system hints at the existence of an attracting center manifold. The attractor is explicitly found for a reduced version of the system.

Key words. nonlinear diffusion, attractor, turbulence

AMS subject classifications. 35K55, 76F20, 76F60

DOI. 10.1137/060657923

1. Introduction.

A variety of physical processes are described by the nonlinear diffusion equations

$$\partial_t K = (-1)^n \nabla \left(K^m \nabla \, \Delta^n K \right), \tag{1}$$

where $n \geq 0$ is an integer and $m > 0$. In the particular case of $n = 0$, (1) becomes the second-order diffusion equation

$$\partial_t K = \nabla \left(K^m \nabla K \right). \tag{2}$$

An important common property of the nonlinear diffusion (2) and its higher-order generalizations (1) is the finiteness of the speed of a signal propagation. If the signal is initially confined in a finite region so that it is identically equal to zero beyond the region, the signal remains confined during the dynamics. This property distinguishes the nonlinear diffusion from the linear diffusion ($m = 0$), where the signal instantaneously propagates to infinity.

The range of processes described by (1)–(2) is wide. The second-order equations (2) are known as the porous medium equations and appear in models of gas filtration in porous media [1, 2] and thin fluid films in a gravitational field ($m = 3$) [3]. The fourth-order equations ($n = 1$) emerge in lubrication models for thin viscous films ($m = 3$) and Hele–Shaw flows ($m = 1$). The sixth-order models ($n = 2$) are relevant to the process of isolation oxidation of silicon ($m = 3$) [4]. The models with $m = 3$ and different values of n describe thin viscous droplets driven by different factors: gravity ($n = 0$), surface tension ($n = 1$), and an elastic plate ($n = 2$).

Various mathematical aspects of the second-order equation (2) are analyzed in [5, 6, 7]; the fourth-order model is investigated, for example, in [8]. Numerical schemes for

*Received by the editors April 22, 2006; accepted for publication (in revised form) April 24, 2007; published electronically September 26, 2007.
http://www.siam.org/journals/siap/67-6/65792.html
†Department of Mathematics and Computing, University of Southern Queensland, Toowoomba, QLD 4350, Australia (strunin@usq.edu.au).

solving the fourth- and sixth-order equations using finite differences or finite elements are developed in [9, 10, 11].

In this paper we focus on attractors in coupled nonlinear diffusion with confined sources when there are more than one diffusing agent.

To give an example of an attractor in nonlinear diffusion we consider the second-order equation (2) in one dimension, $\partial_t K = \partial_x(K^m \partial_x K)$. Its solution evolving from a confined initial profile is attracted to the universal regime [12, 13],

$$K(x,t) = \frac{\gamma(m)}{t^{1/(m+2)}} \left(\xi_0^2 - \xi^2\right)^{1/m}, \tag{3}$$

where

$$\xi = \frac{x}{t^{1/(m+2)}}, \quad \xi_0 = \left[\frac{\Gamma\left(\frac{1}{m}+\frac{3}{2}\right)}{\gamma(m)\sqrt{\pi}\,\Gamma\left(\frac{1}{m}+1\right)} E\right]^{\frac{m}{m+2}}, \quad \gamma(m) = \left[\frac{m}{2(m+2)}\right]^{\frac{1}{m}},$$

Γ is the gamma-function, and E is the integral

$$E = \int_{-\infty}^{\infty} K(x,t)\,dx,$$

which conserves during the evolution. For our purposes it is convenient to write the attractor (3) in the form

$$K(x,t) = \frac{\alpha}{t^{1/(m+2)}} \left(1 - \frac{\beta}{t^{2/(m+2)}}\, x^2\right)^{1/m}, \tag{4}$$

where α and β are the coefficients depending on m and E.

Similar attractors exist for the higher-order equations (1). For example, in one dimension for $m = 1$ and $n = 1$, i.e., for the fourth-order equation $\partial_t K = \partial_x(K\partial_x^3 K)$, the attractor is [14]

$$K(x,t) = \frac{1}{120\,t^{1/5}} \left(\xi_0^2 - \xi^2\right)^2,$$

where

$$\xi = \frac{x}{t^{1/5}}, \quad \xi_0 = \left(\frac{225E}{2}\right)^{1/5}.$$

Recently in [11], (1) was analyzed numerically for $n = 2$ and $m = 1$, that is, the sixth-order equation $\partial_t K = \partial_x(K\partial_x^5 K)$. It was proved that the solution converges to the attractor found in [14].

Many processes involve more than one diffusing agent. For example, in a free turbulent jet the diffusing turbulent energy is coupled with the diffusing momentum. We analyze this process later in the paper. In the multicomponent problems an important question to answer is whether there exists an attractor.

Another motivation for us to analyze this particular phenomenon stems from fluid mechanics, where an expansion of a jet from a narrow pulse is a natural problem formulation. For other processes, other regimes can be of interest. For example, in the isolation oxidation of silicon, traveling waves are of major interest [10].

In this paper we take an approach in which confined solutions are sought as power-series in a spatial coordinate. The diffusion of the turbulent jet is one of many problems to which this approach is applicable.

Previously [15] we studied the expansion of the turbulent jet using a *nonlocal* version of the K-ℓ model of turbulence (see, e.g., the review [16]).

In the present paper we use the K-ε model [17, 18], which is *local*. The locality enables us to convert the governing PDEs into a set of ODEs. Thus, the problem transforms into a standard dynamical system formulation, in which framework we look for an attractor.

As a first step, in section 2 we test our approach on some standard problems to recover known results. Then, in section 3 the approach is applied to the turbulent jet. In section 4 we analyze in detail its reduced version. The conclusions are given in section 5.

2. Power-series approach. In this section we formulate our approach and test it on simple problems. We consider the nonlinear diffusion with linear and then quadratic diffusivity.

2.1. Diffusion with linear diffusivity. Consider (2) with $m = 1$:

$$\partial_t K = \partial_x (K \partial_x K). \tag{5}$$

The long-term asymptotics (4) of its pulse solution is

$$K = \frac{\alpha}{t^{1/3}} \left(1 - \beta \frac{x^2}{t^{2/3}} \right) \tag{6}$$

or

$$K = a(t) \left[1 - b(t) x^2 \right] \tag{7}$$

with

$$a(t) = \frac{\alpha}{t^{1/3}}, \quad b(t) = \frac{\beta}{t^{2/3}}.$$

For finite times, assuming an initial pulse is symmetric, we seek a solution in the form

$$K(x,t) = A(t) \left[1 - B_2(t) x^2 - B_4(t) x^4 - B_6(t) x^6 - \cdots \right], \tag{8}$$

where $A(t) > 0$ is the value of K at $x = 0$. Expression (8) is acceptable as long as it gives a positive answer. Thus, (8) represents the solution on some interval $0 \le x \le h(t)$, where $h(t)$ is the position of the front in which $K[h(t), t] = 0$. Beyond the front, for $x > h(t)$, formula (8) does not apply; we assume $K(x,t) \equiv 0$ instead. With various initial values $B_k(0)$, $k = 2, 4, \ldots$, the form (8) expresses a wide class of symmetric initial conditions. Apparently $B_k(t)$ are proportional to the Taylor-series coefficients of $K(x,t)$.

Substituting (8) into (5), collecting the terms with the same powers of x, and equating the coefficients gives

$$\begin{aligned} \dot{A} &= -2A^2 B_2, \\ \dot{B}_2 &= -4AB_2^2 + 12AB_4, \\ \dot{B}_4 &= -28AB_2B_4 + 30AB_6, \\ \dot{B}_6 &= -54AB_2B_6 - 28AB_4^2 + 56AB_8, \\ &\ldots. \end{aligned} \tag{9}$$

The system (9) contains no linear terms; however, we can "create" those by modifying time. Divide all the equations in (9) by AB_2 and introduce the new time τ by

$$\frac{d}{AB_2\,dt} = \frac{d}{d\tau} \equiv ()'. \tag{10}$$

Then system (9) transforms into

$$\begin{aligned} A' &= -2A, \\ B_2' &= -4B_2 + 12\frac{B_4}{B_2}, \\ B_4' &= -28B_4 + 30\frac{B_6}{B_2}, \\ B_6' &= -54B_6 - 28\frac{B_4^2}{B_2} + 56\frac{B_8}{B_2}, \\ &\dots. \end{aligned} \tag{11}$$

Looking at the coefficients of the linear terms we notice a considerable spectral gap between the coefficient -4 at B_2 in the equation $B_2' = \dots$ and the coefficient -28 at B_4 in the equation $B_4' = \dots$. Therefore we can expect that B_4 and the subsequent B_k, $k = 6, 8, \dots$, will decay much faster than A and B_2. This is confirmed numerically as demonstrated by Figure 1. The plot shows a family of trajectories for the truncated system formed by the dynamic equations for B_2, B_4, and B_6 in (11) with the term containing B_8 removed. The figure gives three different views of the same trajectories to expose the faster decay of B_4 and B_6 in comparison to B_2.

The numerical results in this paper are obtained with the MATLAB solver DAE2 developed by Roberts [19].

It is interesting to evaluate the contribution of different terms, $(-B_k x^k)$, in the function

$$1 - B_2x^2 - B_4x^4 - B_6x^6 - \cdots \tag{12}$$

defining the shape of $K(x,t)$. Let us compare the terms for the largest value of x inside the signal, that is, the coordinate of the front, $x = h(t)$. Retain the first three terms in (12), presuming that the input of the sixth- and higher-order terms is negligible. On the front the signal vanishes, and therefore approximately

$$1 - B_2h^2 - B_4h^4 = 0.$$

From here

$$h^2 = \left(-B_2 + \sqrt{B_2^2 + 4B_4}\right) \Big/ (2B_4). \tag{13}$$

Further, if we suppose that

$$B_4 \ll B_2/h^2, \tag{14}$$

then the fourth-order term appears to be negligible compared to the quadratic term. Inserting (13) into (14) and rearranging, we obtain

$$B_4 \ll 2B_2^2. \tag{15}$$

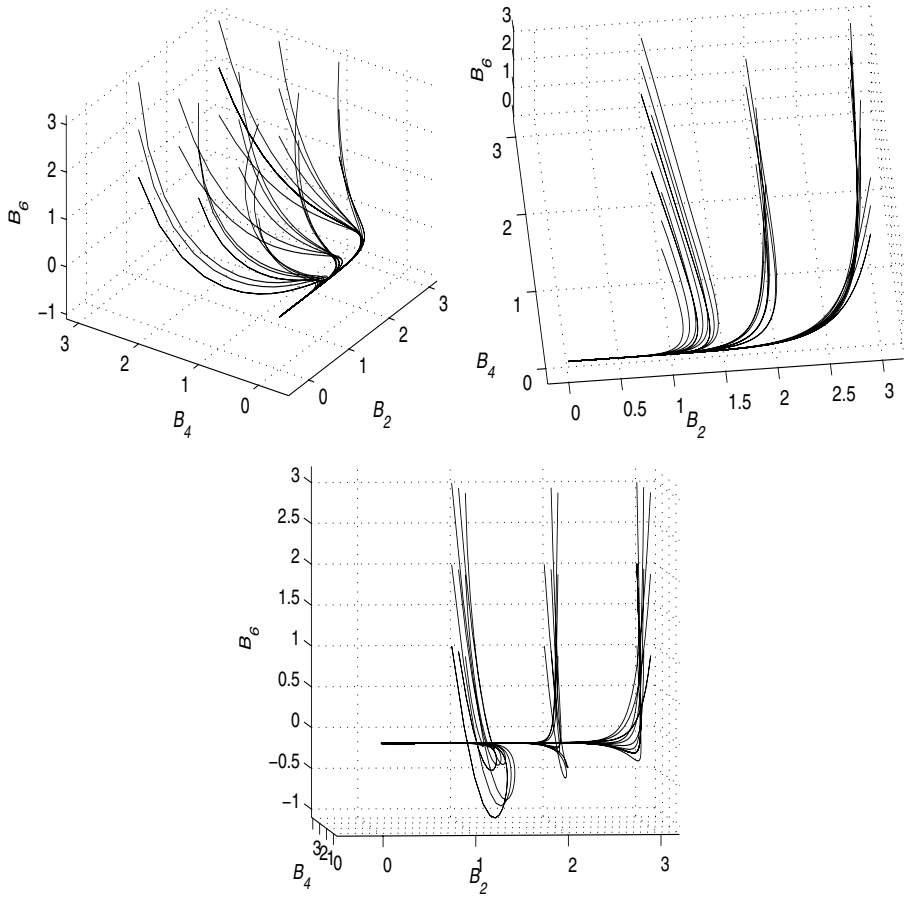


FIG. 1. *Solutions of truncated system* (11).

In the numerical experiments the condition (15) was satisfied several orders over. Tracking the value of $h(t)$ directly from the numerical experiment confirms the negligible contribution of the fourth- and higher-order terms.

Thus, the power-series approach recovered the expected result that (6) is indeed the attractor for the diffusion problem with the linear diffusivity.

2.2. Diffusion with quadratic diffusivity. In this section we consider the diffusion equation with quadratic diffusivity,

$$\partial_t K = \partial_x \left(K^2 \partial_x K \right). \tag{16}$$

Its pulse solution (4) has the form

$$K = \frac{\alpha}{t^{1/4}} \left(1 - \beta \frac{x^2}{t^{1/2}} \right)^{1/2}. \tag{17}$$

Expanding (17) into the Taylor-series, we get

$$K(x,t) = a(t) \left[1 - b_2(t) x^2 - b_4(t) x^4 - b_6(t) x^6 - \cdots \right], \tag{18}$$

where

$$
\begin{aligned}
&b_2(t) = \beta\frac{1}{2\sqrt{t}}, \quad b_4(t) = \beta^2\frac{1}{8t},\\
&b_6(t) = \beta^3\frac{1}{16t^{3/2}}, \quad b_8(t) = \beta^4\frac{5}{128t^2},\\
&\ldots.
\end{aligned}
\tag{19}
$$

It converges for $\beta x^2/t^{1/2} \le 1$, that is, for all x of interest, $0 \le x \le h(t) = t^{1/2}/\beta$. Any b_k in (19) can be expressed through a selected one, for instance, b_2:

$$
b_4 = \frac{1}{2}b_2^2, \quad b_6 = \frac{1}{2}b_2^3, \quad b_8 = \frac{5}{8}b_2^4, \quad \ldots. \tag{20}
$$

Expressions (20) describe the attractor which we intend to reproduce by our approach. As in the previous section, we seek a power-series solution of (16),

$$
K(x,t) = A(t)\,[1 - B_2(t)x^2 - B_4(t)x^4 - B_6(t)x^6 - \cdots]. \tag{21}
$$

Substituting (21) into (16) leads to

$$
\begin{aligned}
&\dot{A} = -2A^3B_2,\\
&\dot{B}_2 = -10A^2B_2^2 + 12A^2B_4,\\
&\dot{B}_4 = -58A^2B_2B_4 + 30A^2B_6 + 10A^2B_2^3,\\
&\dot{B}_6 = -110A^2B_2B_6 + 56A^2B_2^2B_4 - 56A^2B_4^2 + 56A^2B_8,\\
&\dot{B}_8 = -178A^2B_2B_8 + 90A^2B_2^2B_6 + 90A^2B_2B_4^2 - 180A^2B_4B_6 + 90A^2B_{10},\\
&\ldots.
\end{aligned}
\tag{22}
$$

We divide all the equations in (22) by A^2B_2 and introduce the new time τ by

$$
\frac{d}{A^2B_2\,dt} = \frac{d}{d\tau} \equiv ()'. \tag{23}
$$

As a result, system (22) transforms into the following form with linear terms:

$$
\begin{aligned}
&A' = -2A,\\
&B_2' = -10B_2 + \frac{12B_4}{B_2},\\
&B_4' = -58B_4 + \frac{30B_6}{B_2} + 10B_2^2,\\
&B_6' = -110B_6 + 56B_2B_4 - \frac{56B_4^2}{B_2} + \frac{56B_8}{B_2},\\
&B_8' = -178B_8 + 90B_2B_6 + 90B_4^2 - \frac{180B_4B_6}{B_2} + \frac{90B_{10}}{B_2},\\
&\ldots.
\end{aligned}
\tag{24}
$$

Consider only three dynamic equations for B_2, B_2, and B_4 with the term containing B_8 omitted. A set of trajectories for such a system is shown in Figure 2. It is clearly seen from different angles that the trajectories are attracted to a single curve or a one-dimensional manifold. It can be shown that the curve is described by

$$
B_4 = \gamma B_2^3, \quad B_6 = \mu B_2^3, \tag{25}
$$

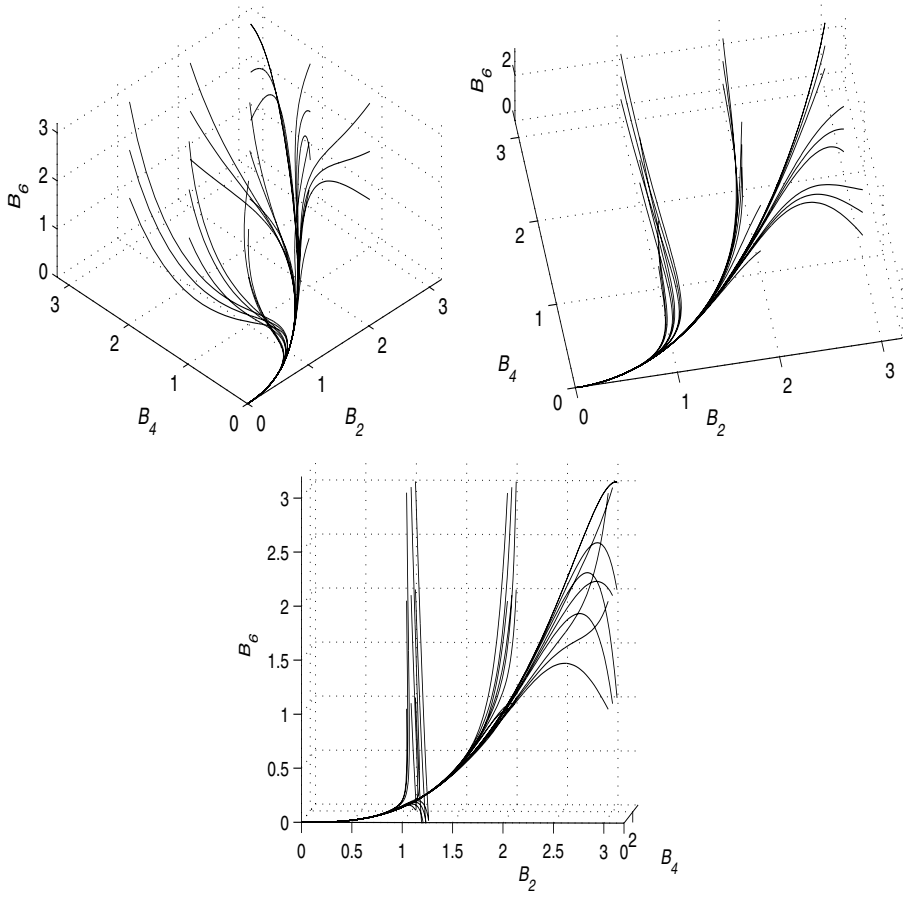


FIG. 2. *Solutions of truncated system* (24).

where γ and μ are parameters. It is interesting to compare them to their values on the attractor defined by (20). We substitute (25) into (24) to obtain

$$\begin{aligned} B_2' &= (-10 + 12\gamma)\, B_2, \\ B_2' &= \left(-29 + \frac{15\mu}{\gamma} + \frac{5}{\gamma}\right) B_2, \\ B_2' &= \left(-\frac{110}{3} + \frac{56\gamma}{3\mu} - \frac{56\gamma^2}{3\mu}\right) B_2. \end{aligned} \tag{26}$$

As each of the equations in (26) describes the motion on the attractor, they all must coincide. So must the coefficients at B_2 in their right-hand sides. This leads to two algebraic equations with respect to γ and μ, giving approximately $\gamma = 0.31$, $\mu = 0.14$. Compare these to the exact values from (20),

$$\gamma_* = 1/2 = 0.5, \quad \mu_* = 1/2 = 0.5.$$

The departure from the exact values can be diminished by involving more equations. For the four-equation system with respect to B_2, B_4, B_6, and B_8, with the term

containing B_{10} excluded, we have

$$B_4 = \gamma B_2^3, \quad B_6 = \mu B_2^3, \quad B_8 = \lambda B_2^4. \tag{27}$$

Inserting (27) into (24) leads to a system of three equations, with respect to γ, μ, and λ, having the approximate solution $\gamma = 0.35$, $\mu = 0.20$, $\lambda = 0.11$. Note that from (20) the exact value for λ is $\lambda_* = 5/8 = 0.62$. To improve the approximation further, more equations have to be involved.

In summary, we reproduced the well-known fact that for the diffusion with quadratic diffusivity the trajectories converge to a one-dimensional manifold representing the similarity regime (17).

3. The K-ε model of a turbulent jet. Consider a turbulent jet created in an unbounded motionless fluid by a quick impulse shaped in space as a narrow flat layer. For instance, some amount of fluid is quickly injected. The velocity shear between the jet and surrounding fluid pumps up the kinetic energy of turbulence. The turbulent region expands and, in the long term, the energy decays due to the geometric effect of expansion and the loss into heat caused by intersections of vortices (we will call this loss simply dissipation).

The expansion is driven by the turbulent diffusion which is essentially nonlinear. As a consequence, the jet has a sharp front similar to the above examples. However, the dynamics is complicated by the coupling between the kinetic energy, dissipation, and momentum. The K-ε model of turbulence [17, 18] is written

$$\begin{aligned}
\partial_t K &= \alpha_1 \partial_x \left(\frac{K^2}{\varepsilon} \partial_x K \right) + \alpha_2 \frac{K^2}{\varepsilon} \left(\partial_x u \right)^2 - \alpha_3 \varepsilon, \\
\partial_t \varepsilon &= \beta_1 \partial_x \left(\frac{K^2}{\varepsilon} \partial_x \varepsilon \right) + \beta_2 K \left(\partial_x u \right)^2 - \beta_3 \frac{\varepsilon^2}{K}, \\
\partial_t u &= \chi \partial_x \left(\frac{K^2}{\varepsilon} \partial_x u \right).
\end{aligned} \tag{28}$$

In (28) the coordinate x is directed across the flat turbulent layer originating in its middle, K stands for the kinetic energy of turbulent pulsations per mass unit, and ε is the dissipation of the turbulent energy; $\alpha_{1,2,3}$, $\beta_{1,2,3}$, and χ are nondimensional coefficients. The system (28) is nondimensional, obtained from dimensional form by using some useful scales, for example, the average initial velocity across the jet, U, as the velocity scale; the initial width of the jet, $2h$, as the length scale; U^2 as the turbulent energy scale; U^3/h as the dissipation rate scale; and h/U as the time scale.

The initial profiles of K, ε, and u across the turbulent layer are assumed to have dome-like forms. We assume that they are symmetric with respect to the middle of the layer. On the edge, or front, of the jet the functions descend to zero and remain zero beyond the front (see the discussion further in this section).

We look for the power-series solutions of (28),

$$\begin{aligned}
K &= A(t) \left[1 - B_2(t) x^2 - B_4(t) x^4 - B_6(t) x^6 - \cdots \right], \\
\varepsilon &= P(t) \left[1 - R_2(t) x^2 - R_4(t) x^4 - R_6(t) x^6 - \cdots \right], \\
u &= M(t) \left[1 - N_2(t) x^2 - N_4(t) x^4 - N_6(t) x^6 - \cdots \right].
\end{aligned} \tag{29}$$

Substituting the series (29) into the dynamic equations (28) leads to

$$
\begin{aligned}
\dot{A} &= -\alpha_1 \frac{2A^3 B_2}{P} - \alpha_3 P,\\
\dot{P} &= -\beta_1 2A^2 R_2 - \beta_3 \frac{P^2}{A},\\
\dot{M} &= -\chi \frac{2A^2 M N_2}{P},\\
\dot{B}_2 &= -\alpha_1 \frac{10A^2 B_2^2}{P} + \alpha_3 \frac{PB_2}{A} + \alpha_1 \frac{6A^2 B_2 R_2}{P} + \alpha_1 \frac{12A^2 B_4}{P}\\
&\quad - \alpha_2 \frac{4AM^2 N_2^2}{P} - \alpha_3 \frac{PR_2}{A},\\
\dot{R}_2 &= -\beta_1 \frac{12A^2 B_2 R_2}{P} + \beta_1 \frac{8A^2 R_2^2}{P} - \beta_3 \frac{PR_2}{A} + \beta_1 \frac{12A^2 R_4}{P}\\
&\quad - \beta_2 \frac{4AM^2 N_2^2}{P} + \beta_3 \frac{PB_2}{A},\\
\dot{N}_2 &= -\chi \frac{12A^2 B_2 N_2}{P} + \chi \frac{2A^2 N_2^2}{P} + \chi \frac{6A^2 N_2 R_2}{P} + \chi \frac{12A^2 N_4}{P},\\
\dot{B}_4 &= -\alpha_1 \frac{58A^2 B_2 B_4}{P} + \alpha_3 \frac{PB_4}{A} + \alpha_1 \frac{10A^2 B_2^3}{P} - \alpha_1 \frac{20A^2 B_2^2 R_2}{P}\\
&\quad + \alpha_1 \frac{10A^2 B_2 R_2^2}{P} + \alpha_1 \frac{10A^2 B_2 R_4}{P} + \alpha_1 \frac{20A^2 B_4 R_2}{P} + \alpha_1 \frac{30A^2 B_6}{P}\\
&\quad + \alpha_2 \frac{8AB_2 M^2 N_2^2}{P} - \alpha_2 \frac{4AM^2 N_2^2 R_2}{P} - \alpha_2 \frac{16AM^2 N_2 N_4}{P} - \alpha_3 \frac{PR_4}{A},\\
\dot{R}_4 &= -\beta_1 \frac{40A^2 B_2 R_4}{P} + \beta_1 \frac{2A^2 R_2 R_4}{P} - \beta_3 \frac{PR_4}{A} + \beta_1 \frac{10A^2 B_2^2 R_2}{P}\\
&\quad - \beta_1 \frac{20A^2 B_2 R_2^2}{P} - \beta_1 \frac{20A^2 B_4 R_2}{P} + \beta_1 \frac{10A^2 R_2^3}{P} + \beta_1 \frac{30A^2 R_2 R_4}{P}\\
&\quad + \beta_1 \frac{30A^2 R_6}{P} + \beta_2 \frac{4AB_2 M^2 N_2^2}{P} - \beta_2 \frac{16AM^2 N_2 N_4}{P} + \beta_3 \frac{PB_2^2}{A}\\
&\quad - \beta_3 \frac{2B_2 PR_2}{A} + \beta_3 \frac{B_4 P}{A} + \beta_3 \frac{PR_2^2}{A},\\
\dot{N}_4 &= -\chi \frac{40A^2 B_2 N_4}{P} + \chi \frac{2A^2 N_2 N_4}{P} + \chi \frac{10A^2 B_2^2 N_2}{P} - \chi \frac{20A^2 B_2 N_2 R_2}{P}\\
&\quad - \chi \frac{20A^2 B_4 N_2}{P} + \chi \frac{10A^2 N_2 R_2^2}{P} + \chi \frac{10A^2 N_2 R_4}{P} + \chi \frac{20A^2 N_4 R_2}{P}\\
&\quad + \chi \frac{30A^2 N_6}{P},\\
&\ldots.
\end{aligned}
\tag{30}
$$

An immediate idea of how to solve (30) could be to truncate the system by removing higher-order variables and solve the resulting closed system under some initial conditions. However, such an approach has a serious flaw since there is no guarantee that the three fronts—the energy front, dissipation front, and velocity front—would coincide during the evolution. By the physics of diffusion, if the fronts do not coincide initially, they must catch up with each other. Suppose, for example, that initially the velocity front is behind the energy and dissipation fronts (suppose that these two coincide). Then the turbulent diffusion will instantaneously transfer the momentum forward up to the energy/dissipation front position. Conversely, if the velocity front

is initially ahead of the energy/dissipation front, it will be motionless for some time, as there is no turbulence in its vicinity. The velocity front would move only when the energy/dissipation front catches up, after which all the fronts move together. The front $x = h(t)$, where the energy, dissipation, and velocity decrease to a zero level, is a special point, yet the system (30) "does not know about it." We should explicitly impose the physical condition that the three profiles (29) must meet at the point $(K = \varepsilon = u = 0,\ x = h)$.

Let us demonstrate with a simple example that the lack of such a condition leads to the growth of the gap between the fronts. Consider a relatively simple model

$$\partial_t K = \partial_x(K\partial_x K),$$
$$\partial_t u = \partial_x(K\partial_x u),$$

without attributing any physical sense to K and u. We look for power-series solutions in the form of (29) and transfer to the new time by using $AB_2\,dt = d\tau$. Retaining only two leading equations for the series coefficients and removing terms with B_4 and N_4, we have

$$\begin{aligned} B_2' &= -4B_2, \\ N_2' &= -6N_2 + \frac{2N_2^2}{B_2}. \end{aligned} \tag{31}$$

Upon solving (31), the front of K can be determined from

$$1 - B_2 h_K^2 = 0,$$

and the front of u can be found from

$$1 - N_2 h_u^2 = 0.$$

Clearly $N_2 = B_2$ satisfies (31), but is this solution stable? Introduce the perturbation s by

$$N_2 = B_2 - s. \tag{32}$$

Substituting (32) into (31) and linearizing, we get

$$s' = -2s.$$

The perturbation decays as $\exp(-2\tau)$, whereas B_2 decays, according to (31), as $\exp(-4\tau)$. Thus, the perturbation goes to zero slower than the function itself. This leads to a large discrepancy between the values of B_2 and N_2, and hence in the front positions, $h_u^2 = 1/N_2$ and $h_K^2 = 1/B_2$. A similar effect occurs with the system (30).

Let us see how the situation changes if we require that the fronts coincide. We augment (31) by two extra equations stating that the functions turn into zero at the same point $x = h(t)$, that is, $K(h,t) = 0$ and $u(h,t) = 0$, where K and u are represented by the truncated series (29). The two extra equations bring one extra unknown, h. Therefore we need to add another unknown to have as many equations as unknowns. Let this new unknown be N_4. We get

$$\begin{aligned} B_2' &= -4B_2, \\ N_2' &= -6N_2 + \frac{12N_4}{B_2} + \frac{2N_2^2}{B_2}, \\ 1 - B_2 h^2 &= 0, \\ 1 - N_2 h^2 - N_4 h^4 &= 0. \end{aligned} \tag{33}$$

Excluding h and N_4 from (33) leads to

$$\begin{aligned} B_2' &= -4B_2, \\ N_2' &= -6N_2 + \frac{2N_2^2}{B_2} + 12(B_2 - N_2). \end{aligned} \tag{34}$$

Substituting (32) into (34) and linearizing gives

$$s' = -14s.$$

Now s decays much faster than B_2 so that, in contrast to the previous case, B_2 and N_2 become closer to each other.

Applying a similar approach to the system (30), we require that K, ε, and u turn into zero at the same location $x = h(t)$. Retaining in the power-series (29) only terms up to the fourth order, we require

$$\begin{aligned} 1 - B_2h^2 - B_4h^4 &= 0, \\ 1 - R_2h^2 - R_4h^4 &= 0, \\ 1 - N_2h^2 - N_4h^4 &= 0. \end{aligned} \tag{35}$$

Equations (35) are complemented by the truncated dynamic equations (30),

$$\begin{aligned} \dot{A} &= -\alpha_1\frac{2A^3B_2}{P} - \alpha_3 P, \\ \dot{P} &= -\beta_1 2A^2R_2 - \beta_3\frac{P^2}{A}, \\ \dot{M} &= -\chi\frac{2A^2MN_2}{P}, \\ \dot{B}_2 &= -\alpha_1\frac{10A^2B_2^2}{P} + \alpha_3\frac{PB_2}{A} + \alpha_1\frac{6A^2B_2R_2}{P} + \alpha_1\frac{12A^2B_4}{P} \\ &\quad - \alpha_2\frac{4AM^2N_2^2}{P} - \alpha_3\frac{PR_2}{A}, \\ \dot{R}_2 &= \beta_1\frac{8A^2R_2^2}{P} - \beta_3\frac{PR_2}{A} - \beta_1\frac{12A^2B_2R_2}{P} + \beta_1\frac{12A^2R_4}{P} \\ &\quad - \beta_2\frac{4AM^2N_2^2}{P} + \beta_3\frac{PB_2}{A}, \\ \dot{N}_2 &= \chi\frac{2A^2N_2^2}{P} - \chi\frac{12A^2B_2N_2}{P} + \chi\frac{6A^2N_2R_2}{P} + \chi\frac{12A^2N_4}{P}, \\ \dot{B}_4 &= -\alpha_1\frac{58A^2B_2B_4}{P} + \alpha_3\frac{PB_4}{A} + \alpha_1\frac{10A^2B_2^3}{P} - \alpha_1\frac{20A^2B_2^2R_2}{P} \\ &\quad + \alpha_1\frac{10A^2B_2R_2^2}{P} + \alpha_1\frac{10A^2B_2R_4}{P} + \alpha_1\frac{20A^2B_4R_2}{P} + \alpha_1\frac{30A^2B_6}{P} \\ &\quad + \alpha_2\frac{8AB_2M^2N_2^2}{P} - \alpha_2\frac{4AM^2N_2^2R_2}{P} - \alpha_2\frac{16AM^2N_2N_4}{P} - \alpha_3\frac{PR_4}{A}. \end{aligned} \tag{36}$$

The system (35)–(36) contains 10 equations with respect to 10 time-dependent functions: A, P, M, B_2, R_2, N_2, B_4, R_4, N_4, and h.

As before, we "create" linear terms by modifying time:

$$\frac{d}{(A^2B_2/P)\,dt} = \frac{d}{d\tau} \equiv ()'. \tag{37}$$

Dividing (36) by A^2B_2/P and converting to τ results in

$$\begin{aligned}
A' &= -\alpha_1 2A - \alpha_3 \frac{P^2}{A^2B_2},\\
P' &= -\beta_1 \frac{2R_2P}{B_2} - \beta_3 \frac{P^3}{A^3B_2},\\
M' &= -\chi \frac{2MN_2}{B_2},\\
B_2' &= -\alpha_1 10B_2 + \alpha_3 \frac{P^2}{A^3} + \alpha_1 6R_2 + \alpha_1 \frac{12B_4}{B_2}\\
&\quad - \alpha_2 \frac{4M^2N_2^2}{AB_2} - \alpha_3 \frac{P^2R_2}{A^3B_2},\\
R_2' &= -\beta_1 12R_2 + \beta_1 \frac{8R_2^2}{B_2} - \beta_3 \frac{P^2R_2}{A^3B_2} + \beta_1 \frac{12R_4}{B_2}\\
&\quad - \beta_2 \frac{4M^2N_2^2}{AB_2} + \beta_3 \frac{P^2}{A^3},\\
N_2' &= -\chi 12N_2 + \chi \frac{2N_2^2}{B_2} + \chi \frac{6N_2R_2}{B_2} + \chi \frac{12N_4}{B_2},\\
B_4' &= -\alpha_1 58B_4 + \alpha_3 \frac{P^2B_4}{A^3B_2} + \alpha_1 10B_2^2 - \alpha_1 20B_2R_2\\
&\quad + \alpha_1 10R_2^2 + \alpha_1 10R_4 + \alpha_1 \frac{20B_4R_2}{B_2} + \alpha_1 \frac{30B_6}{B_2}\\
&\quad + \alpha_2 \frac{8M^2N_2^2}{A} - \alpha_2 \frac{4M^2N_2^2R_2}{AB_2} - \alpha_2 \frac{16M^2N_2N_4}{AB_2} - \alpha_3 \frac{P^2R_4}{A^3B_2}.
\end{aligned} \tag{38}$$

Figures 3, 4, and 5 display some trajectories for the system (35)–(38). We used $\alpha_1 = 0.09$, $\alpha_2 = 0.09$, $\alpha_3 = 1$, $\beta_1 = 0.07$, $\beta_2 = 0.13$, $\beta_3 = 1.92$, $\chi = 0.09$. The initial conditions were chosen to ensure the same position for the three fronts.

Figure 6 shows the front propagation. Note that the seeming acceleration occurs only in terms of the artificial time τ. In terms of the real time t the graph will have opposite curvature showing deceleration.

We notice a considerable spectral gap between the linear decay rates in (36): the coefficient at B_4, $(-58\alpha_1)$, is from 5 to 6 times larger than that of B_2, $(-10\alpha_1)$, of R_2, $(-12\beta_1)$, and of N_2, (-12χ). The numerical data show that the linear terms are the largest in absolute value in each dynamic equation.

The numerical data also show that on the initial sections of the trajectories the velocity terms M, N_2, and N_4 in the equation $B_4' = \ldots$ are much smaller than the terms associated with the energy and dissipation. However, after some period of time the velocity-associated terms become comparable to the other terms.

This points to a mechanism characteristic of center manifolds, where some variables, such as B_4 in our problem, rapidly decay until they are small enough to be comparable with the nonlinear terms which come into play.

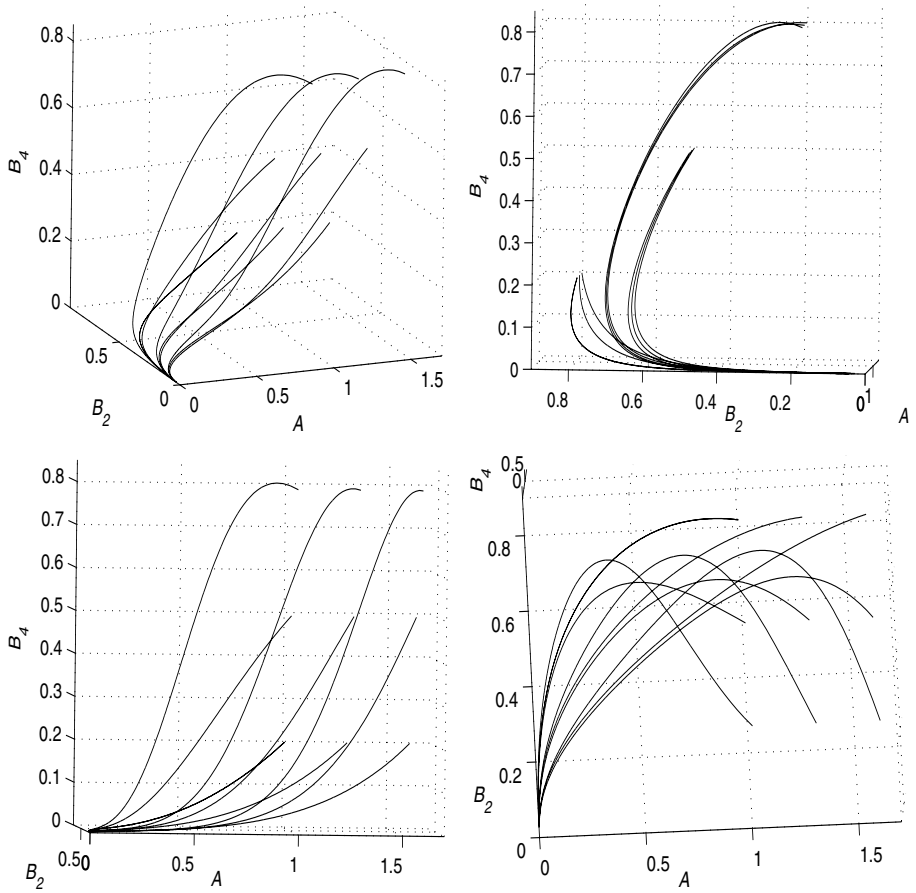


FIG. 3. *Trajectories for the model* (35)–(38) *in the space of the energy variables.*

To illustrate this mechanism we use a simple example from [20]:

$$\begin{aligned} \dot{x} &= -px - xy, \\ \dot{y} &= -y + x^2 - 2y^2. \end{aligned} \tag{39}$$

The linear decay rate p of x is much smaller than that of y, say $p = 0.1 \ll 1$. A set of trajectories for the system (39) is shown in Figure 7. See that the trajectories are attracted to a single curve. It can be shown that in the limit $p = 0$ the attractor is exactly

$$y = x^2, \tag{40}$$

which is called the center manifold. Driven by the linear term $(-y)$ the trajectories quickly drop onto the manifold on which the nonlinear terms $(x^2 - 2y^2)$ are comparable to $(-y)$. On the attractor, in view of (40), the motion is described by $\dot{x} = -xy = -x^3$. The variable y depends on t through x to which it is rigidly linked by (40).

We anticipate that a similar situation takes place in our problem with B_4 being analogous to y in the above example. However, the problem is complicated by a multitude of variables. In this paper we investigate a simplified version of the model.

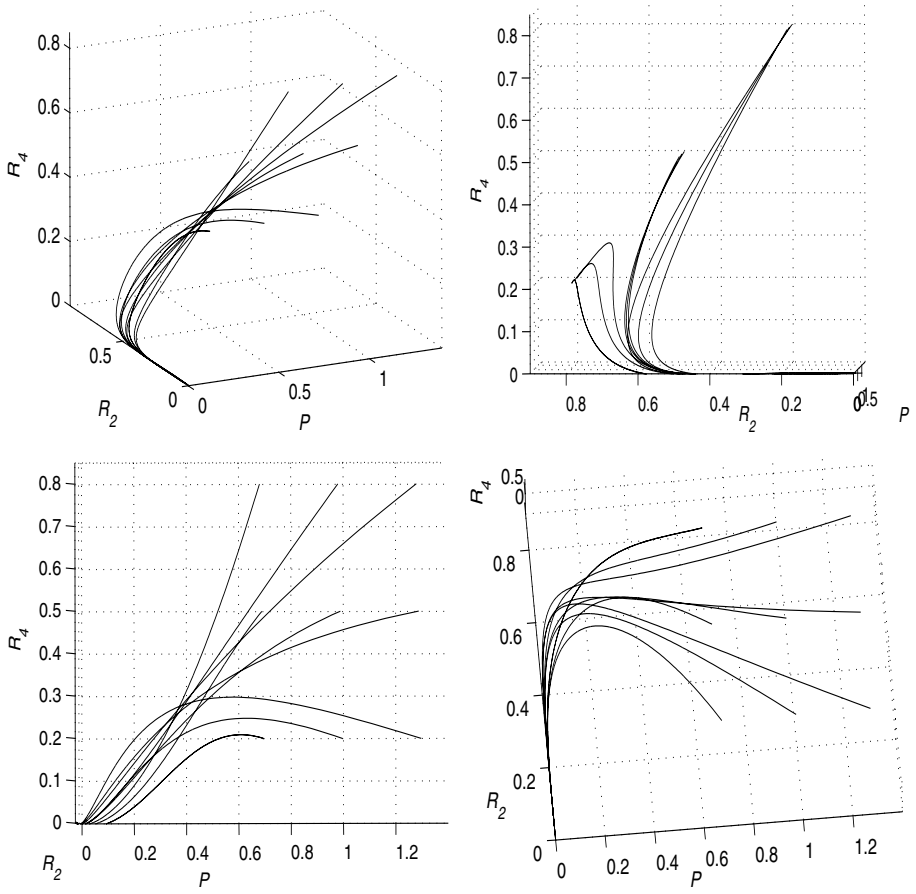


FIG. 4. *Trajectories for the model* (35)–(38) *in the space of the dissipation variables.*

4. Reduced version of the model. In this section we simplify the K-ε model (28) to a great extent, yet make sure that the key physical factors remain. These factors are the nonlinear diffusion and the coupling via the velocity shear. We will keep calling K the energy and u the velocity for consistency with the previous section. However, these terms should not be directly associated with the physical quantities. If we find attractors, our approach will provide a useful basis for studies of more complicated systems.

We assume that (a) $\alpha_3 = \beta_3 = 0$ to remove the dissipation terms, (b) $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 1$ for simplicity, and (c) initial conditions for K and ε coincide. Thus, the problem formulations for K and ε become identical; therefore $K \equiv \varepsilon$ at all times, that is, $A(t) \equiv P(t)$ and $B_k(t) \equiv R_k(t)$, $k = 2, 4, \ldots$. As a result, system (28) reduces to the two equations

$$\begin{aligned} \partial_t K &= \partial_x \left(K \partial_x K \right) + K \left(\partial_x u \right)^2 , \\ \partial_t u &= \partial_x \left(K \partial_x u \right) . \end{aligned} \tag{41}$$

The manipulations in section 3 automatically apply to (41). The definition of new time (37) transforms into

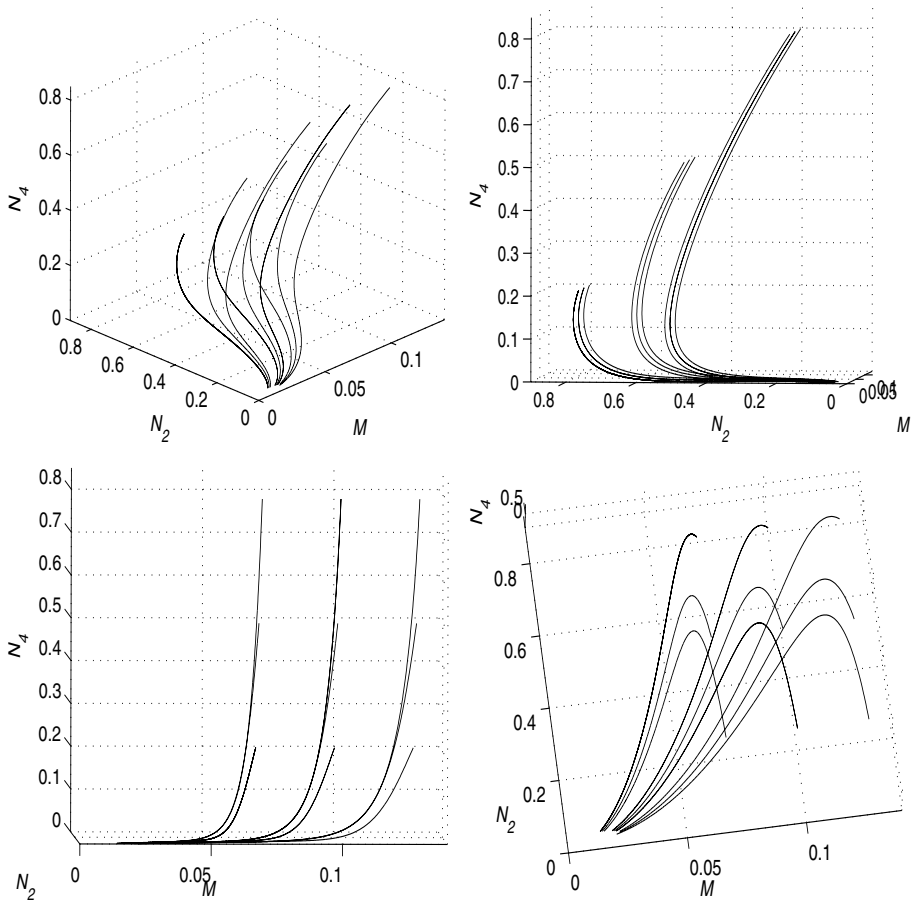


FIG. 5. *Trajectories for the model* (35)–(38) *in the space of the velocity variables.*

$$\frac{d}{(AB_2)\,dt} = \frac{d}{d\tau} \equiv ()'. \tag{42}$$

We introduce the new function

$$T \equiv \frac{M^2}{A}. \tag{43}$$

It turns out that it is possible to derive a dynamic equation for T, where M and A appear in combination (43). This equation will replace the two dynamic equations for A and M. Differentiating (43) gives

$$T' = \left(\frac{M^2}{A}\right)' = \frac{2MM'A - M^2A'}{A^2}. \tag{44}$$

The following expressions for A' and M' are obtained from (38) under conditions (a), (b), and (c): $A' = -2A$ and $M' = -2MN_2/B_2$. Substituting these into (44) gives the equation shown below.

Under assumptions (a), (b), and (c), and in view of (44), system (35)–(38) becomes

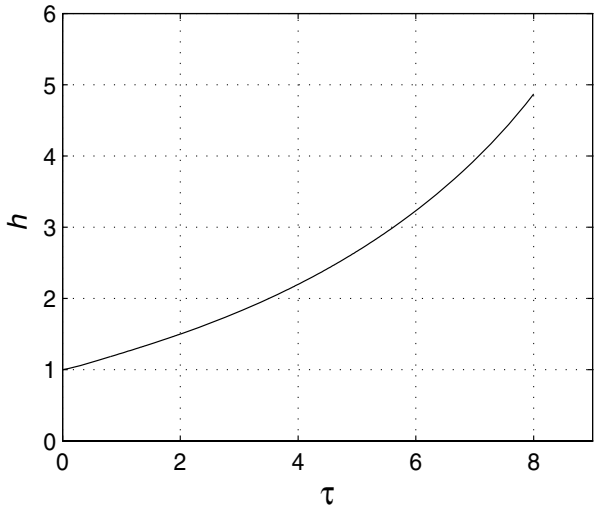


FIG. 6. *Propagation of the turbulent front in the model* (35)–(38).

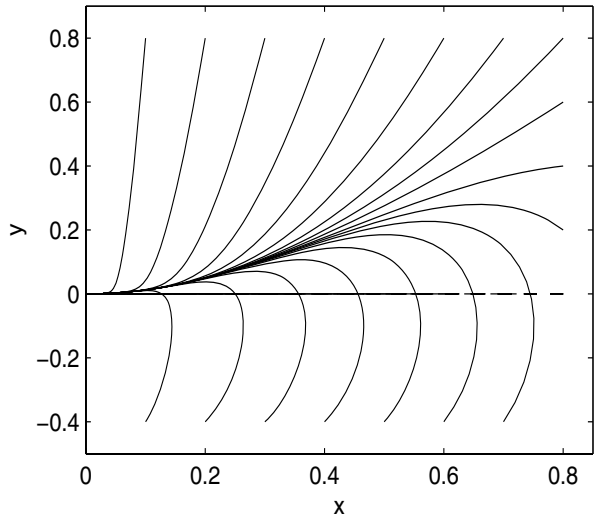


FIG. 7. *Center manifold in model* (39).

$$
\begin{aligned}
&T' = -\frac{4N_2T}{B_2} + 2T,\\
&B_2' = -4B_2 - \frac{4N_2^2T}{B_2} + \frac{12B_4}{B_2},\\
&N_2' = -6N_2 + \frac{2N_2^2}{B_2} + \frac{12N_4}{B_2},\\
&B_4' = -28B_4 + 4N_2^2T - \frac{16N_2N_4T}{B_2},\\
&1 - B_2h^2 - B_4h^4 = 0,\\
&1 - N_2h^2 - N_4h^4 = 0.
\end{aligned}
\tag{45}
$$

System (45) contains six equations with respect to six unknown functions: T, B_2, N_2, B_4, N_4, and h. We solved (45) numerically, making sure that the initial positions of the energy and velocity fronts coincide.

Using the data from the numerical experiments we can deduce results in analytical form. According to the data, at large times some terms in (45) become negligible, namely, $(-16N_2N_4T/B_2)$ in the equation $B_4' = \ldots$, $(12B_4/B_2)$ and $(-4N_2^2T/B_2)$ in the equation $B_2' = \ldots$, and $(12N_4/B_2)$ in the equation $N_2' = \ldots$. Also it is important to note $N_2 \to B_2$. Therefore from (45) we get *asymptotically*

$$T' = -2T, \quad N_2' = -4N_2, \quad B_2 = N_2, \tag{46}$$

from which we get

$$T = T_0 e^{-2(\tau-\tau_0)}, \quad B_2 = N_2 = N_0 e^{-4(\tau-\tau_0)}, \tag{47}$$

where τ_0, T_0, and N_0 are some reference values. Substituting (47) into (45), we get

$$B_4' = -28B_4 + 4N_0^2 T_0 e^{-10(\tau-\tau_0)}. \tag{48}$$

The solution of the homogeneous part of (48), $\sim \exp[-28(\tau-\tau_0)]$, expresses the decay caused by the linear term $(-28B_4)$. This part of the solution is negligible compared to the solution of the nonhomogeneous equation,

$$B_4 = Ce^{-10(\tau-\tau_0)}, \tag{49}$$

expressing the forced dynamics of B_4.

Here we recognize the center manifold mechanism: a rapid decay of a function to a level where the linear term becomes comparable to the nonlinear term. It is easy to find the constant by substituting (49) into (48),

$$C = \frac{2}{9}N_0^2 T_0.$$

Hence, the variable B_4 is attracted to the manifold described by

$$B_4(\tau) = \frac{2}{9}N_0^2 T_0 \exp[-10(\tau-\tau_0)] = \frac{2}{9}N_2^2 T$$

or, using (43),

$$B_4 = \frac{2N_2^2 M^2}{9A}. \tag{50}$$

Let us use the numerical data directly to show that B_4 is indeed attracted to (50). There are many ways to demonstrate the attraction, and below we implement just one of them. A graph B_4 versus N_2^2 and T would be a curved surface. We go over to new variables, in terms of which the surface would be a plane,

$$N_2^2 = \xi + \eta, \quad T = \xi - \eta. \tag{51}$$

The product $N_2^2 T$ is represented by a plane in terms of ξ^2 and η^2:

$$N_2^2 T = \xi^2 - \eta^2. \tag{52}$$

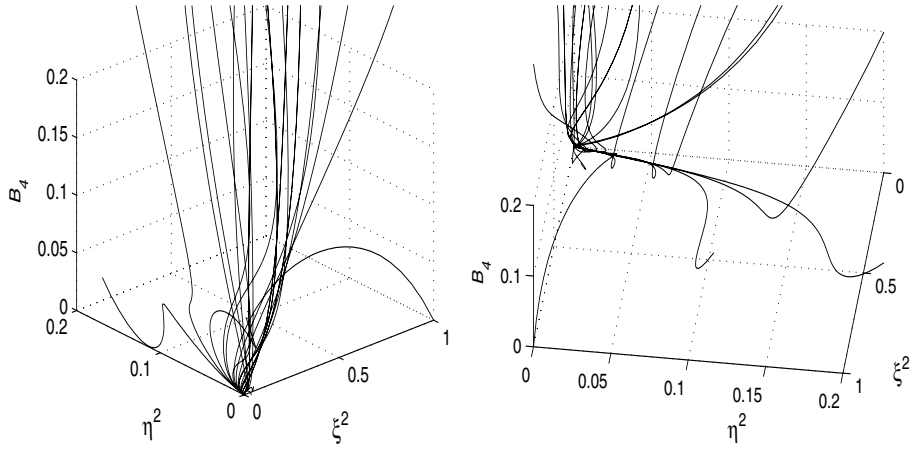


FIG. 8. *Trajectories for the reduced model* (45).

The new variables are defined by (51),

$$\xi = \frac{1}{2}\left(N_2^2 + T\right), \quad \eta = \frac{1}{2}\left(N_2^2 - T\right).$$

Figure 8 gives two views of a set of trajectories in the space (ξ^2, η^2). The right-hand view shows that all the trajectories converge to a surface which, from this particular angle, appears as a straight line. Clearly the surface is a plane.

In order to obtain a closed system from (35)–(38), we removed the term with B_6 from the equation $B_4' = \ldots$ and obtained the dynamics of N_4 from the front equation $1 - N_2h^2 - N_4h^4 = 0$ rather than from the respective dynamic equation.

To let N_4 evolve according to the dynamic law, we add the equation $N_4' = \ldots$. Adding the extra equation makes it necessary to add another unknown to the system, say B_6 (or alternatively, N_6; however, this is not of principle importance):

$$\begin{aligned}
T' &= -\frac{4N_2T}{B_2} + 2T,\\
B_2' &= -4B_2 - \frac{4N_2^2T}{B_2} + \frac{12B_4}{B_2},\\
N_2' &= -6N_2 + \frac{2N_2^2}{B_2} + \frac{12N_4}{B_2},\\
B_4' &= -28B_4 + 4N_2^2T - \frac{16N_2N_4T}{B_2} + \frac{30B_6}{B_2},\\
N_4' &= -20N_4 + \frac{2N_2N_4}{B_2} - \frac{10N_2B_4}{B_2},\\
1 &- B_2h^2 - B_4h^4 - B_6h^6 = 0,\\
1 &- N_2h^2 - N_4h^4 = 0.
\end{aligned} \tag{53}$$

Trajectories for the system (53) are shown in Figure 9. We see the same attractor as for the shorter version (45). In the dynamic equation $B_4' = \ldots$ the new term $(30B_6/B_2)$ and the old term $(-16N_2N_4T/B_2)$ are smaller than the other terms by two orders of magnitude.

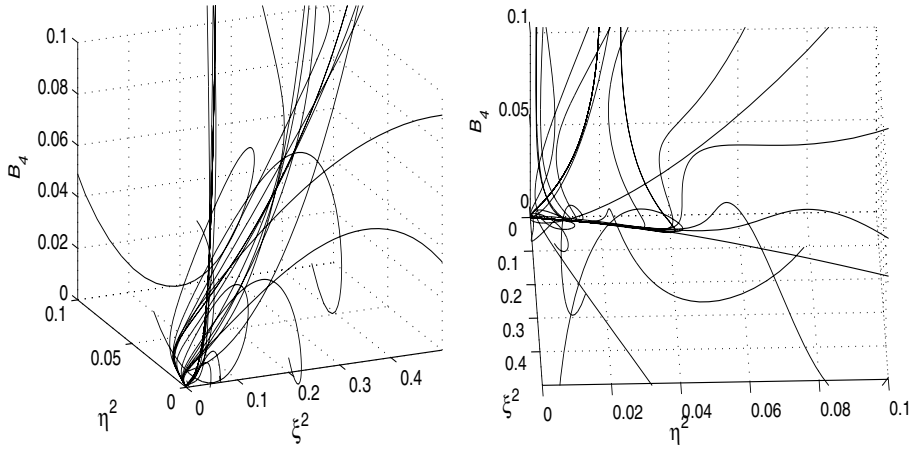


FIG. 9. *Trajectories for the enhanced version* (53) *of the reduced model.*

Now we look for an attractor for N_4. Substituting $N_2 = B_2$ (we emphasize that this relation is asymptotic, not exact) and the expression for B_4, (50), into the equation $N_4' = \dots$ in (53) we get

$$N_4' = -18N_4 - \frac{20}{9}N_0^2 T_0 e^{-10(\tau-\tau_0)}. \tag{54}$$

As in the case for B_4, the solution of the homogeneous part of this equation rapidly decays as $\exp[-18(\tau - \tau_0)]$ so that the solution is virtually the forced one,

$$N_4 = De^{-10(\tau-\tau_0)}. \tag{55}$$

Substituting (55) into (54) gives $D = -5N_0^2 T_0/18$, and therefore the attractor is

$$N_4 = -\frac{5N_2^2 M^2}{18A}. \tag{56}$$

A graph N_4 against ξ^2 and η^2 is again a plane, as is evident from Figure 10.

Remarkably, dividing (50) by (56), we find

$$\frac{B_4}{N_4} = -\frac{4}{5} \tag{57}$$

on the attractor. The dependence B_4/N_4 versus time in the numerical experiments is given in Figure 11. It clearly shows the attraction to the predicted value $(-4/5) = -0.8$.

Further extension of the model can be done by involving the equation $B_6' = \dots$ (without the term with B_8). Accordingly we need to add another unknown, for instance, N_6, to make the front equation $1 - N_2h^2 - N_4h^4 - N_6h^6 = 0$. This process can be continued.

More equations would give a more accurate description; however, the result about the existence of the attractors (50) and (56) holds.

In summary, the solutions of the confined-source problem for the quasi–fluid-dynamical system $\partial_t K = \partial_x (K\partial_x K) + K(\partial_x u)^2$, $\partial_t u = \partial_x (K\partial_x u)$ converge to the

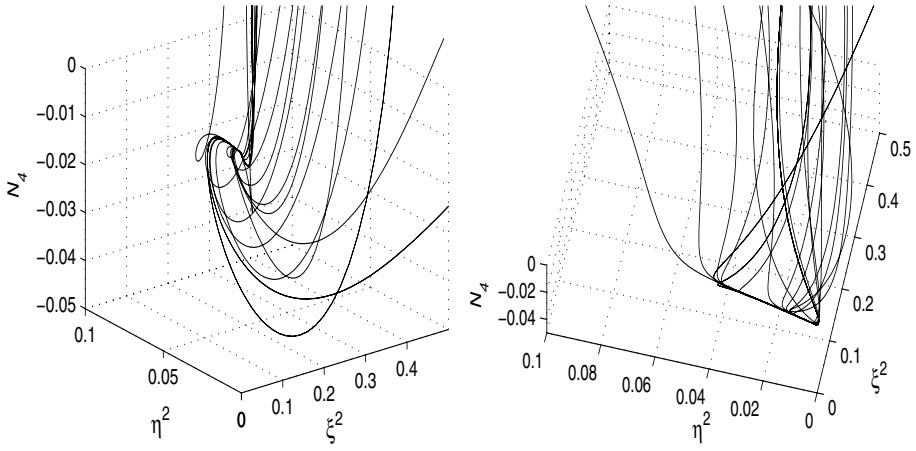


FIG. 10. *Trajectories for the enhanced version* (53) *of the reduced model.*

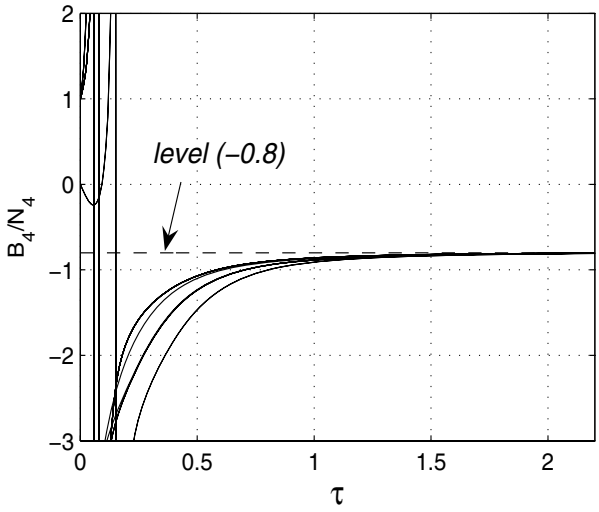


FIG. 11. *The ratio B_4/N_4 versus time.*

attractor

$$K = A\,(1 - B_2x^2 - B_4x^4 - \cdots),$$
$$u = M\,(1 - N_2x^2 - N_4x^4 - \cdots),$$

where

$$B_4 = \frac{2N_2^2M^2}{9A}, \quad N_4 = -\frac{5N_2^2M^2}{18A} \tag{58}$$

and the variables A, M, B_2, and N_2 evolve according to (43), (53). (Note that one should not substitute the asymptotic result (58) into (53) and then solve for A, M, B_2, and N_2. This would break stability, similarly to example (34). If in (34) one replaces $12(B_2 - N_2)$ by its asymptotic value zero, the system becomes unstable; see (31).)

5. Conclusions. We considered a process of an expansion of a turbulent jet driven by turbulent diffusion. The mathematical model essentially involves coupling between the turbulent energy, dissipation, and momentum. Looking for solutions in the form of power-series in a spatial coordinate, we derived dynamical systems with respect to time-dependent series coefficients. The system is essentially nonlinear; however, modifying time allowed us to create linear terms, which dominate during the early dynamics. We analyzed in detail a simplified version of the model with a radically reduced number of variables. The numerical and analytical analyses allowed us to find an attractor in exact form.

REFERENCES

[1] M. Muskat, *The Flow of Homogeneous Fluids through Porous Media*, McGraw-Hill, New York, 1937.

[2] D. G. Aronson, *The porous medium equation*, in Nonlinear Diffusion Problems, Lecture Notes in Math. 1224, A. Fasano and M. Primicerio, eds., Springer, Berlin, 1986, pp. 1–46.

[3] J. Buckmaster, *Viscous sheets advancing over dry beds*, J. Fluid Mech., 81 (1977), pp. 735–756.

[4] J. R. King, *The isolation oxidation of silicon: The reaction-controlled case*, SIAM J. Appl. Math., 49 (1989), pp. 1064–1080.

[5] J. L. Vazquez, *Hyperbolic aspects in the theory of the porous media equation*, in Metastability and Incompletely Posed Problems, S. Antman et al., eds., Springer, New York, 1987, pp. 325–342.

[6] L. A. Peletier, *The porous media equation*, in Applications of Nonlinear Analysis in the Physical Sciences, H. Amann et al., eds., Pitman, London, 1981, pp. 229–241.

[7] D. G. Aronson, *Nonlinear diffusion problems*, in Free Boundary Problems: Theory and Applications I, Res. Notes in Math. 78, A. Fasano and M. Primicerio, eds., Pitman, London, 1983, pp. 135–149.

[8] F. Bernis, *Viscous flows, fourth order nonlinear degenerate parabolic equations and singular elliptic problems*, in Free Boundary Problems: Theory and Applications, Pitman Res. Notes Math. Ser. 323, J. I. Diaz, M. A. Herrero, A. Linan, and J. L. Vazquez, eds., Longman Sci. Tech., Harlow, 1995, pp. 40–56.

[9] J. W. Barrett, J. F. Blowey, and H. Garcke, *Finite element approximation of a fourth order nonlinear degenerate parabolic equation*, Numer. Math., 80 (1998), pp. 525–556.

[10] J. D. Evans, M. Vynnycky, and S. P. Ferro, *Oxidation-induced stresses in the isolation oxidation of silicon*, J. Engrg. Math., 38 (2000), pp. 191–218.

[11] J. W. Barrett, S. Langdon, and R. Nurnberg, *Finite element approximation of sixth order nonlinear degenerate parabolic equation*, Numer. Math., 96 (2004), pp. 401–434.

[12] Ya. B. Zel'dovich and A. S. Kompaneets, *On the theory of heat propagation for temperature dependent thermal conductivity*, in Collection Commemorating the 70th Anniversary of A. F. Joffe, Izv. Akad. Nauk SSSR, 1950, pp. 61–71.

[13] A. D. Polyanin and V. F. Zaitsev, *Handbook of Nonlinear Partial Differential Equations*, Chapman and Hall/CRC, Boca Raton, FL, 2004.

[14] N. F. Smyth and J. M. Hill, *High-order nonlinear diffusion*, IMA J. Appl. Math., 40 (1988), pp. 73–86.

[15] D. V. Strunin and A. J. Roberts, *Self-similarity of decaying turbulent layer*, in Proceedings of the 5th Biennial Engineering Mathematics and Applications Conference (EMAC-2002), The Institution of Engineers, Brisbane, Australia, pp. 205–210.

[16] C. G. Speziale, *Analytical methods for the development of Reynolds-stress closures in turbulence*, in Annual Review of Fluid Mechanics, Vol. 23, Annual Reviews, Palo Alto, CA, 1991, pp. 107–157.

[17] B. E. Launder, G. J. Reece, and W. Rodi, *Progress in the development of a Reynolds-stress turbulence closure*, J. Fluid Mech., 68 (1975), pp. 537–566.

[18] K. Hanjalic and B. E. Launder, *A Reynolds stress model of turbulence and its applications to thin shear flows*, J. Fluid Mech., 52 (1972), pp. 609–638.

[19] A. J. Roberts, *Tony Roberts' Home Page Links to Information*, Department of Mathematics and Computing, University of Southern Queensland, Toowoomba, Australia, http://www.sci.usq.edu.au/staff/robertsa/ (2005).

[20] A. J. Roberts, *Appropriate initial conditions for asymptotic descriptions of the long term evolution of dynamical systems*, J. Austral. Math. Soc. Ser. B, 31 (1989), pp. 48–75.

SIAM J. APPL. MATH.
Vol. 67, No. 6, pp. 1675–1692

NUMERICAL INVESTIGATION OF CAVITATION IN MULTIDIMENSIONAL COMPRESSIBLE FLOWS*

KRISTEN J. DEVAULT†, PIERRE A. GREMAUD†, AND HELGE KRISTIAN JENSSEN‡

Abstract. The compressible Navier–Stokes equations for an ideal polytropic gas are considered in $\mathbb{R}^n$, $n = 2, 3$. The question of possible vacuum formation, an open theoretical problem, is investigated numerically using highly accurate computational methods. The flow is assumed to be symmetric about the origin with a purely radial velocity field. The numerical results indicate that there are weak solutions to the Navier–Stokes system in two and three space dimensions, which display formation of vacuum when the initial data are discontinuous and sufficiently large. The initial density is constant, while the initial velocity field is symmetric, points radially away from the origin, and belongs to H^s_{loc} for all $s < n/2$. In addition, in the one-dimensional case, the numerical solutions are in agreement with known theoretical results.

Key words. Navier–Stokes, Euler, compressible flow, cavitation, pseudospectral

AMS subject classifications. 35Q30, 65M70, 76N99

DOI. 10.1137/060652713

1. Introduction.

A long-standing open problem in the mathematical theory for fluid dynamics is the question of vacuum formation in compressible flow. Roughly stated, the issue is: Do there exist solutions of the Navier–Stokes system (3.1)–(3.3) for viscous compressible flow that exhibit vacuum (vanishing density) in finite time when the initial density is strictly bounded away from zero?

This problem is relevant from a modeling perspective as well as for theoretical results. The underlying assumption in the derivation of the Navier–Stokes equations from physical principles is that the fluid is nondilute and can be described as a continuum. A negative answer to the question above would thus provide self-consistency of the continuum assumption for the Navier–Stokes model. On the other hand, it is known that an a priori estimate of the form

$$\text{for a constant } C = C(T): \quad C^{-1} \le \rho(x,t) \le C \quad \text{for all } (x,t) \in \mathbb{R}^n \times [0,T] \tag{1.1}$$

(where ρ denotes density) implies further estimates and would greatly facilitate existence proofs. For a discussion of this point, see Chapter 3 of [11].

Estimates of the above type are available for one-dimensional (1D) flow [26], even in the case of large and discontinuous data [15]. As reviewed in section 2, large efforts have been invested in searching for similar bounds in the multidimensional (multi-D) case. However, no such results seem to be currently known for the standard Navier–Stokes model with constant transport coefficients.

*Received by the editors February 22, 2006; accepted for publication (in revised form) April 27, 2007; published electronically September 26, 2007.
http://www.siam.org/journals/siap/67-6/65271.html

†Department of Mathematics and Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695-8205 (kjdevaul@ncsu.edu, gremaud@ncsu.edu). The first author's research was partially supported by the National Science Foundation (NSF) through grant DMS-0410561. The second author's research was partially supported by the NSF through grants DMS-0244488 and DMS-0410561.

‡Department of Mathematics, Penn State University, University Park, State College, PA 16802 (jenssen@math.psu.edu). This author's research was partially supported by the NSF through grants DMS-0206631 and DMS-0539549 (CAREER).

To gain some insight, it is natural first to make a careful numerical study of the simplest possible scenario where one could expect cavitation in several space dimensions. This is the subject of the present work. In section 3, we consider 2D and 3D flows with symmetry for the full Navier–Stokes equations, as well as for the barotropic case (where pressure is assumed to be a function of density alone). The equations are written in a proper nondimensionalized form, and the problem is completed with appropriate initial and boundary conditions. More is known in the formal limit of infinitely large Reynolds numbers, i.e., in the case of inviscid fluids. The corresponding Euler equations and their solutions are used as a benchmark for the numerics. Various considerations regarding the inviscid case are discussed in section 4.

The above difficulties are mirrored on the numerical side in the form of various challenges regarding stability. The case of 1D flows illustrates this issue. Riemann data with large jumps may lead to vacuum formation for the Euler equations while, at least for Hoff's solutions [15], no vacuum occurs for solutions to the corresponding Navier–Stokes equations. Minimizing the amount of numerical diffusion is thus paramount. A splitting method is used between the convective part, corresponding to the Euler equations, and the diffusive part. The hyperbolic part is solved by computing *local* similarity solutions. Those solutions are then "diffused," and the process is repeated at the next time step. A highly accurate pseudospectral spatial discretization is used; see section 5. The stiffness [1] of the discretized-in-time system increases as the density ρ goes to zero; for $\rho = 0$, the system is infinitely stiff and has degenerated from being purely differential to being differential algebraic. Several numerical schemes, none of them explicit, are known to handle this kind of difficulty [1, 12] (at least for some problems presenting this type of structure). A backward difference formula (BDF)-type method is considered here. This approach has been successfully tested with respect to mass conservation and energy balance. Finally, a test over the magnitude of the density has to be done to determine whether "vacuum" has been reached, an arduous task in finite precision computations! The solution is analyzed both for its magnitude and its behavior in phase space.

The numerical experiments are set up so that vacuum, if any, appears first at a known point (the origin). The calculations do not attempt to track the solution past vacuum formation. The detection of the vacuum itself requires density to be less than 10^{-14} (see criterion (6.1)), i.e., just slightly above what corresponds to machine epsilon for the IEEE double precision machines used in the experiments.

Our detailed numerical study is described in section 6. It indicates that vacuum formation indeed occurs for multi-D symmetric flows for sufficiently large and discontinuous initial data, i.e., in the regime of high Mach number and high Reynolds number. Section 7 contains a brief summary and a conclusion of our findings.

2. Challenges and relation to other works.

2.1. Theoretical issues.

There is a voluminous literature on compressible flow. A short review of work relevant to vacuum formation and a priori estimates follows.

2.1.1. 1D flows.

For 1D flow, i.e., multi-D flow with planar symmetry, much stronger results are known than what is currently available for higher dimensions. A seminal work of Kazhikhov and Shelukhin [26] considers the full 1D Navier–Stokes system for an ideal polytropic gas. Building on earlier work by Kanel [21] and Kazhikhov [25], the global existence and uniqueness of a smooth ($W^{1,2}$) solution is proved in [26] for arbitrarily large and smooth data. In particular, a priori bounds of the type (1.1) are established. Similar results for more general gases can be found

in [4], [5], [24].

Highly relevant to the present work is an extension to large and rough (possibly discontinuous) data established by Hoff [15]. This result pertains to isentropic or isothermal flow with data (ρ_0, u_0) satisfying

$$\rho_0 \in L^\infty(\mathbb{R})\,, \qquad \operatorname{ess\,inf}_{\mathbb{R}} \rho_0 > 0\,,$$

and

$$\rho_0 - \bar{\rho}\,, \quad u_0 - \bar{u} \ \in L^2(\mathbb{R})\,,$$

where $\bar{\rho}, \bar{u}$ are monotone functions that agree outside a bounded interval with the limiting values of ρ and u at $\pm\infty$. Hoff [15] proves that there exists a global weak solution which satisfies (1.1). This result shows that there is at least one solution of the Navier–Stokes equations that does not exhibit cavitation, even for Riemann-type data with arbitrarily large jumps. (For an extension of this result to certain flows for the full Navier–Stokes system see [20].) This is in contrast to the 1D inviscid Euler equations, for which it is well known that vacuum formation can occur [36]. The present numerical study does not contradict Hoff's result. Indeed, while our results clearly indicate the possibility of vacuum formation in higher dimensions, we have not numerically observed cavitation in solutions of the 1D Navier–Stokes system.

A significant result concerning cavitation in 1D flow is given by Hoff and Smoller [19]. They demonstrate that any, everywhere defined, weak solution of the Navier–Stokes (barotropic or full) system which satisfies some natural weak integrability assumptions cannot contain a vacuum in a nonempty open set unless the initial data do so. (For a refinement of this result, see [9].) Xin and Yuan [38] have recently performed a corresponding analysis for spherically symmetric solutions in $\mathbb{R}^2$ and $\mathbb{R}^3$. For everywhere defined solutions they give detailed information on the behavior of vacuum regions (if any), and they also provide sufficient conditions to rule out cavitation.

2.1.2. Multi-D flows. Much less is known about compressible flow in higher dimensions. Global existence of weak solutions is a formidable problem, and the theory is far from complete. Roughly speaking, currently known results are of two types: (A) for large and rough data that possibly contain vacuum states, and (B) for small, rough (discontinuous) data with $\operatorname{ess\,inf} \rho_0 > 0$.

In the former case, Lions [28] has established existence of global weak solutions in the case of compressible barotropic flow. For recent extensions, including results for the full Navier–Stokes system, see [10], [11] and references therein. It is not known if a bound of the type (1.1) holds for these solutions when the data satisfy $\operatorname{ess\,inf} \rho_0 > 0$.

More is known for "small" data, i.e., data close to a constant state in a suitable norm. In particular, small and sufficiently smooth data generate global smooth solutions without cavitation for the full Navier–Stokes system [31], [22], [23]. For a representative result in the case of barotropic flow, see Chapter 9 in [33]. In a series of papers [6], [7], [8], Danchin has established global existence, and also uniqueness, of compressible flows in several space dimensions for solutions in so-called critical spaces. For flows with even less regularity, possibly with discontinuities across hypersurfaces, Hoff [14] has shown that if the data are sufficiently close to a constant state, in a suitable norm, and with initial density and temperature bounded away from zero, then there exists a global weak solution with the same properties at all later times. No corresponding result seems to be known for large data in several space dimensions.

There are somewhat stronger results available for the type of symmetric (quasi-1D) flows that we consider in this paper. For isothermal flow with spherical symmetry, Hoff [13] establishes existence of a global weak solution for large symmetric data. The solution is obtained as the limit of solutions in shells $\{0 < a \leq r \leq b\}$ as $a \downarrow 0$. By rewriting the equations in Lagrangian coordinates and exploiting the energy estimate, certain a priori bounds are obtained that are independent of the inner radius a. However, while guaranteeing existence of a weak solution, the available a priori bounds do not seem strong enough to determine whether the constructed solution contains a vacuum at the center of motion. (For an extension of this result to the full Navier–Stokes system, see [17].)

We note that the higher the dimension of the space, the easier it should be to generate a vacuum, as the fluid is free to move in more directions. This can be quantified for the corresponding inviscid system. Consider the isentropic Euler equations with spherically symmetric Riemann-type data. Let the initial velocity field have constant magnitude $\bar{u}$ and be directed radially away from the origin. In this case, there is a threshold value $\hat{u}(n)$ of $\bar{u}$, depending on the dimension n, above which a vacuum is formed immediately [43]. One can verify that $\hat{u}(1) > \hat{u}(2) > \hat{u}(3)$. For the 1D Navier–Stokes system, we know from Hoff's result [13] that there exists a weak solution without vacuum; i.e., "$\hat{u}(1) = \infty$" for these solutions. However, in higher dimensions it may well be that there are solutions with strictly positive density everywhere at time zero, but which develop a vacuum at later times.[1] There are also technical reasons that seem to prevent a priori bounds on the density in higher dimensions. More precisely, in the 1D analysis of [26], [13] the bounds on the density are derived from the a priori bounds one gets "for free" from the equations themselves. These are integral bounds that are strictly stronger in one dimension than in higher dimensions due to the geometrical factor of r^{n-1} in the space integrals (i.e., $dx = const.r^{n-1}\,dr$).

2.2. Additional remarks.

Cavitation and uniqueness. The issue of cavitation is closely related to the question of uniqueness and to the concept of solution that one works with. To illustrate this consider the 1D Navier–Stokes equations with Riemann-type data:

$$\rho_0(x) \equiv \bar{\rho} > 0\,, \qquad u_0(x) = \begin{cases} -\bar{u} & \text{for } x < 0\,, \\ \bar{u} & \text{for } x > 0\,, \end{cases} \tag{2.1}$$

where $\bar{u} > 0$. One weak solution to this problem is provided by Hoff's result [15], and this solution does not exhibit cavitation. However, a different solution can also be constructed, with the same data, by piecing together solutions of two disjoint flows into surrounding vacuum. More precisely, consider the two sets of initial data

$$\rho_0^-(x) = \begin{cases} \bar{\rho} & \text{for } x < 0\,, \\ 0 & \text{for } x > 0\,, \end{cases} \qquad u_0^-(x) = \begin{cases} -\bar{u} & \text{for } x < 0\,, \\ \emptyset & \text{for } x > 0\,, \end{cases} \tag{2.2}$$

and

$$\rho_0^+(x) = \begin{cases} 0 & \text{for } x < 0\,, \\ \bar{\rho} & \text{for } x > 0\,, \end{cases} \qquad u_0^+(x) = \begin{cases} \emptyset & \text{for } x < 0\,, \\ \bar{u} & \text{for } x > 0\,. \end{cases} \tag{2.3}$$

(Here $\emptyset$ indicates that the velocity is left undefined where there is no matter.) One can now construct solutions (ρ^-, u^-) and (ρ^+, u^+) corresponding to these data, which

[1] Furthermore, the possibility remains that there are other weak solutions exhibiting vacuum even in one dimension (see below for further comments on uniqueness).

in addition satisfy a physical no-traction boundary condition along a vacuum-fluid interface; see [4], [5], [24], [25]. By concatenating the two solutions, a solution to the original problem (2.1) is obtained, and in this solution an open vacuum region is present from time $t = 0+$ (and staying at least for a short time). In the region between the two solutions, one may simply consider the flow to be undefined. Without going into the discussion of which solution is more relevant, we note that Hoff's solution [15] is defined *everywhere* on $\mathbb{R} \times \mathbb{R}_+$, while the second solution is defined only on the support of its density. The issue of nonuniqueness for 1D compressible Navier–Stokes in connection with vacuums is treated in detail by Hoff and Serre [18].

For flow in several space dimensions, the picture is less clear since we do not even know if there is a solution without cavitation in this case. The present work indicates that there is at least one solution where a vacuum forms. Uniqueness of *general* weak solutions, i.e., without any regularity assumptions beyond what is necessary to make sense of a weak formulation, is not known. On the other hand, sufficiently smooth and small solutions are unique (see [31]), as are flows belonging to critical spaces (see Danchin [6], [7], [8]). To the best of our knowledge, the only uniqueness result for flows with possible discontinuities in the density field is given in a recent work by Hoff [16].

Continuum assumption and physical boundary conditions. In view of the 1D examples above, one should be cautious in making claims about the "physicality" of constructed or computed solutions to the standard Navier–Stokes model (3.1)–(3.3) in the low-density regime. Of course, from a modeling point of view, this is not surprising. The Navier–Stokes system is derived under the assumption that the fluid can be described as a continuum with everywhere strictly positive mass density. Issues related to nonuniqueness in the presence of vacuum are therefore not surprising. For a discussion of this point, see [11]. It is also known [37] that the lifespan of smooth, everywhere defined solutions to the Navier–Stokes system (with vanishing heat conductivity) is finite whenever the initial density is compactly supported.

Once a vacuum has developed, one should impose the physical boundary condition of vanishing traction at the vacuum-fluid interface. In other words, the vacuum should not exert a force on the fluid. Note that in the present work we track the *onset* of vacuum formation, not its subsequent evolution.

Well-posedness. In view of both physical arguments as well as the apparent lack of good a priori estimates for the standard Navier–Stokes model, it is natural to ask whether more accurate models would lead to stronger results. In particular, models where the transport coefficients λ, μ, κ depend on the thermodynamical state have been considered.[2] Recent results for such models are given in [2] and [32]. Several issues pertaining to well-posedness in the presence of vacuum have been analyzed in [29], [30], [39], [40], [41], [42]. These results show that the nonuniqueness observed in [18] can be attributed to the unphysical assumption of constant viscosity coefficient. The corresponding problem for the full system appears to require new methods. For results in this direction for the full 1D system, see [4], [5], [24], and see the recent monograph by Feireisl [11] for the multi-D case.

2.3. Precise formulation. After the above remarks, our original question can be formulated precisely as follows. Given the standard multidimensional Navier–Stokes model with constant transport coefficients, let the pressure be that of an ideal

[2] In the isentropic case a relevant assumption is that the viscosity μ depends on the density. In accordance with kinetic theory it is natural to consider the case with $\mu \sim \rho^k$, for a constant $k > 0$.

polytropic gas or, in the case of barotropic flow, of the form $A\rho^\gamma$ with $\gamma \geq 1$. Consider the initial-boundary value problem in a ball centered at the origin, with initial density strictly bounded away from zero and with a possibly discontinuous initial velocity field. Then: *Does there exist an everywhere defined weak solution of the equations with the property that its density reaches zero in finite time?*

3. The full compressible Navier–Stokes equations. Consider the compressible Navier–Stokes equations for a Newtonian fluid in $\mathbb{R}^n$, $n = 1$, $n = 2$, or $n = 3$, with no external forces or heat sources. The invariant form of the equations in spatial (Eulerian) formulation is

$$\rho_t + \operatorname{div}(\rho\vec{u}) = 0, \tag{3.1}$$

$$(\rho\vec{u})_t + \operatorname{div}(\rho\vec{u}\otimes\vec{u}) = \operatorname{grad}(-p + \lambda \operatorname{div}\vec{u}) + \operatorname{div}(2\mu D), \tag{3.2}$$

$$\mathcal{E}_t + \operatorname{div}\big((\mathcal{E}+p)\vec{u}\big) = \operatorname{div}\big(\lambda(\operatorname{div}\vec{u})\vec{u} + 2\mu D\cdot\vec{u} - \vec{q}\big), \tag{3.3}$$

where ρ is the density, $\vec{u} = (u_1, \ldots, u_n)^T$ is the fluid velocity, p is the pressure, $\mathcal{E}$ is the total energy, D is the deformation rate tensor, $\vec{q}$ is the heat flux vector, and λ and μ are the viscosity coefficients. Equations (3.1)–(3.3) are often referred to as the continuity equation, the conservation of momentum equation, and the conservation of energy equation, respectively. We also have

$$\mathcal{E} = \rho(e + |\vec{u}|^2/2), \qquad D_{ij} = (\partial_i u_j + \partial_j u_i)/2, \qquad \vec{q} = -\kappa\nabla\theta,$$

where e stands for the internal energy, κ is the coefficient of heat conductivity, and θ is the temperature. In what follows, we restrict ourselves to the study of ideal and polytropic (perfect) gases such that

$$p = \mathcal{R}\rho\theta, \qquad e = c_v\theta, \tag{3.4}$$

where $\mathcal{R}$ is the gas constant and c_v is the specific heat at constant volume. The local sound speed c is then given by

$$c = \sqrt{\frac{\gamma p}{\rho}},$$

where $\gamma = 1 + \mathcal{R}/c_v$ is the adiabatic exponent. All the transport coefficients $c_v, \lambda, \mu, \kappa$ are assumed to be constant. For the derivation of the equations, see, e.g., [33], [34].

3.1. Equations for symmetric flow. We next consider the case of flow with symmetry; i.e., the velocity is directed (radially when $n = 2, 3$) away from the origin, and all quantities are functions only of the distance to the origin and of time. Let x denote a point in space, and set $r = |x|$. Setting

$$\rho(r,t) = \rho(x,t), \qquad \vec{u}(x,t) = u(r,t)\frac{x}{r}, \quad \text{etc.},$$

leads to the following system of equations:

$$\rho_t + (\rho u)_\xi = 0, \tag{3.5}$$

$$\rho(u_t + uu_r) + p_r = \nu u_{\xi r}, \tag{3.6}$$

$$c_v\rho(\theta_t + u\theta_r) + pu_\xi = \kappa\theta_{r\xi} + \nu(u_\xi)^2 - \frac{2m\mu}{r^m}(r^{m-1}u^2)_r\,, \tag{3.7}$$

where we have used the notation

$$m = n - 1, \qquad \partial_\xi = \partial_r + \frac{m}{r}, \qquad \nu = \mu + 2\lambda\,.$$

(Note that $\partial_{\xi r} \neq \partial_{r\xi}$ for $n = 2$ and $n = 3$.)

We consider spherically symmetric flows (3.5)–(3.7) in the interior of the interval/disk/ball B_b of fixed outer radius b:

$$\rho(r,0)=\rho_0(r), \qquad u(r,0)=u_0(r), \qquad \theta(r,0)=\theta_0(r) \qquad \text{for } r\le b. \tag{3.8}$$

In the 1D case we require that (3.8) hold for all $|r|\le b$. Throughout, we consider only the case where the gas is set in motion in the outward direction; i.e., we assume that u_0 is a nonnegative function in two and three dimensions and is odd with positive values for $r>0$ in one dimension. Suitable boundary conditions are discussed in 3.4.

3.2. Nondimensional form of symmetric equations. Using the initial data, characteristic length, velocity, density, and temperature can be defined as follows:

$$\begin{aligned}\bar r &:= b\,,\\ \bar u &:= \max_{0\le r\le b}|u_0(r)|\,,\\ \bar\rho &:= \max_{0\le r\le b}\rho_0(r)\,,\\ \bar\theta &:= \max_{0\le r\le b}\theta_0(r)\,.\end{aligned}$$

From these we define characteristic time and pressure by

$$\begin{aligned}\bar t &:= \frac{\bar r}{\bar u}\,,\\ \bar p &:= p(\bar\rho,\bar\theta)\,.\end{aligned}$$

The dimensionless independent variables are then

$$R:=\frac{r}{\bar r}, \qquad T:=\frac{t}{\bar t},$$

and the dimensionless dependent variables are

$$D:=\frac{\rho}{\bar\rho}, \qquad U:=\frac{u}{\bar u}, \qquad \Theta:=\frac{\theta}{\bar\theta}, \qquad P:=\frac{p}{\bar p}\,.$$

Regarding D, U, Θ, P as functions of R and T, we obtain the nondimensionalized system

$$\rho_t+(\rho u)_\xi=0, \tag{3.9}$$

$$\rho(u_t+uu_r)+\frac{1}{\gamma M^2}(\rho\theta)_r=\frac{1}{\mathrm{Re}}u_{\xi r}, \tag{3.10}$$

$$\begin{aligned}\rho(\theta_t+u\theta_r)+(\gamma-1)\rho\theta u_\xi &= \frac{1}{\mathrm{Pr}\,\mathrm{Re}}\theta_{r\xi}\\ &\quad+\gamma(\gamma-1)\frac{\mathrm{M}^2}{\mathrm{Re}}\left((u_\xi)^2-\frac{2m\mu}{\nu}\frac{(r^{m-1}u^2)_r}{r^m}\right),\end{aligned} \tag{3.11}$$

where we have reverted to the original symbols and where

$$\begin{aligned}\mathrm{M}&:=\frac{|\bar u|}{\bar c}=\text{ Mach number}, \qquad \bar c=\text{ sound speed }=\sqrt{\frac{\gamma\bar p}{\bar\rho}}\,,\\ \mathrm{Re}&:=\frac{\bar r\bar\rho\bar u}{\nu}=\text{ Reynolds number},\\ \mathrm{Pr}&:=\frac{\nu c_v}{\kappa}=\text{ Prandtl number}.\end{aligned}$$

These equations are valid also for $n=1$ ($m=0$), provided that r is interpreted as the position along the x-axis; in this case $\partial_\xi=\partial_r=\partial_x$.

3.3. Barotropic equations. If the pressure p is considered as a function of the density ρ only, the energy equation (3.7) decouples from the mass and momentum conservation equations (3.5), (3.6). Here, we consider exclusively the isentropic and isothermal cases where

$$p(\rho) = a\rho^\gamma, \qquad \gamma \geq 1\,, \tag{3.12}$$

and where $a > 0$ is a constant. A nondimensional version of the barotropic equations can be derived in very much the same way as (3.9)–(3.11) were derived, except for the characteristic pressure, which is now taken as

$$\bar{p} = p(\bar{\rho}) = a\bar{\rho}^\gamma.$$

This leads to the nondimensionalized system for barotropic flow

$$\rho_t + (\rho u)_\xi = 0, \tag{3.13}$$

$$\rho(u_t + uu_r) + \frac{1}{\gamma \mathrm{M}^2}(\rho^\gamma)_r = \frac{1}{\mathrm{Re}} u_{\xi r}\,, \tag{3.14}$$

where the Mach number at the reference state is now given by

$$M := \frac{|\bar{u}|}{\bar{c}} = \frac{|\bar{u}|}{a\gamma\bar{\rho}^{\gamma-1}}\,.$$

3.4. Initial and boundary conditions: Balance relations. The following initial and boundary conditions are considered throughout:

$$\rho(r,0) = \rho_0(r) = 1 \quad \text{for } r \geq 0, \qquad u(r,0) = u_0(r) = \begin{cases} 1 & \text{if } r > 0, \\ 0 & \text{if } r = 0, \end{cases} \tag{3.15}$$

$$u(1,t) = 1, \qquad t > 0, \tag{3.16}$$

where $r = 1$ corresponds now to the outer boundary of the computational domain. Furthermore, by symmetry, one has

$$u(0,t) = 0, \qquad t > 0. \tag{3.17}$$

A calculation shows that the initial velocity field is in $H^s_{loc}(\mathbb{R}^n)$ for all $s < n/2$.

It may seem more natural to consider a homogeneous condition of the type $u(1,t) = 0$, instead of (3.16), since existence of weak solutions has been established in the former case. However, such a condition leads to steep gradients and unwanted numerical boundary layer effects on the outer boundary of the domain. A vanishing boundary condition at $x = 1$ would thus complicate the numerical resolution of the problem by making it more susceptible to spurious oscillations. In any case, under the above type of initial conditions (3.15), vacuum formation (if any) is expected to be initiated at the origin (see section 6). For short time intervals one would therefore expect that the outer boundary condition does not significantly influence the behavior of the solution near the origin.

4. The isentropic Euler equations. The isentropic Euler equations are easily obtained from (3.13), (3.14) by formally taking the limit $\mathrm{Re} \to \infty$, i.e.,

$$\rho_t + (\rho u)_\xi = 0, \tag{4.1}$$

$$\rho(u_t + uu_r) + \frac{1}{\gamma \mathrm{M}^2}(\rho^\gamma)_r = 0. \tag{4.2}$$

4.1. The 1D case. Let us consider (4.1), (4.2) together with the Riemann data

$$\rho(r,0)=1 \quad \text{for all } r, \qquad u(r,0)=\begin{cases} -1 & \text{if } r<0,\\ 1 & \text{if } r>0. \end{cases} \tag{4.3}$$

As is well known, the above Riemann problem can easily be solved. Interestingly, the isothermal case, $\gamma=1$, and the general isentropic case, $\gamma>1$, are quite different.

The isothermal case $\gamma=1$. For the "symmetric data" (4.3), the solution is found to consist of two rarefaction waves:

$$\begin{bmatrix} \rho \\ u \end{bmatrix}(r,t)=\begin{cases} \begin{bmatrix} 1 \\ -1 \end{bmatrix} & \text{if } \frac{r}{t}<-1-\frac{1}{\mathrm{M}},\\ \begin{bmatrix} e^{-(\mathrm{M}\frac{r}{t}+\mathrm{M}+1)} \\ \frac{r}{t}+\frac{1}{\mathrm{M}} \end{bmatrix} & \text{if } -1-\frac{1}{\mathrm{M}}<\frac{r}{t}<-\frac{1}{\mathrm{M}},\\ \begin{bmatrix} e^{-\mathrm{M}} \\ 0 \end{bmatrix} & \text{if } -\frac{1}{\mathrm{M}}<\frac{r}{t}<\frac{1}{\mathrm{M}},\\ \begin{bmatrix} e^{\mathrm{M}\frac{r}{t}-\mathrm{M}-1} \\ \frac{r}{t}-\frac{1}{\mathrm{M}} \end{bmatrix} & \text{if } \frac{1}{\mathrm{M}}<\frac{r}{t}<1+\frac{1}{\mathrm{M}},\\ \begin{bmatrix} 1 \\ 1 \end{bmatrix} & \text{if } 1+\frac{1}{\mathrm{M}}<\frac{r}{t}. \end{cases}$$

Note that, regardless of the values of the Mach number, M, the above solution does not lead to cavitation. Indeed, the smallest value of the density is found to be $e^{-\mathrm{M}}$.

The case $\gamma>1$. Again, the solution is found to consist of two rarefaction waves for the data (4.3). However, in the present case, cavitation can occur. More precisely, if $M>\frac{2}{\gamma-1}$, the solution $\left[\begin{smallmatrix}\rho\\ u\end{smallmatrix}\right](r,t)$ is given by

$$\begin{cases} \begin{bmatrix} 1 \\ -1 \end{bmatrix} & \text{if } \frac{r}{t}<-1-\frac{1}{\mathrm{M}},\\ \begin{bmatrix} \left(\frac{2}{\gamma+1}-\mathrm{M}\frac{\gamma-1}{\gamma+1}(1+\frac{r}{t})\right)^{2/(\gamma-1)} \\ \frac{1}{\mathrm{M}(\gamma+1)}\left(2+(1-\gamma)\mathrm{M}+2\mathrm{M}\frac{r}{t}\right) \end{bmatrix} & \text{if } -1-\frac{1}{\mathrm{M}}<\frac{r}{t}<-1+\frac{2}{\gamma-1}\frac{1}{\mathrm{M}},\\ \begin{bmatrix} 0 \\ \emptyset \end{bmatrix} & \text{if } -1+\frac{2}{\gamma-1}\frac{1}{\mathrm{M}}<\frac{r}{t}<1-\frac{2}{\gamma-1}\frac{1}{\mathrm{M}},\\ \begin{bmatrix} \left(\frac{2}{\gamma+1}+\mathrm{M}\frac{\gamma-1}{\gamma+1}(-1+\frac{r}{t})\right)^{2/(\gamma-1)} \\ \frac{1}{\mathrm{M}(\gamma+1)}\left(-2+(-1+\gamma)\mathrm{M}+2\mathrm{M}\frac{r}{t}\right) \end{bmatrix} & \text{if } 1-\frac{2}{\gamma-1}\frac{1}{\mathrm{M}}<\frac{r}{t}<1+\frac{1}{\mathrm{M}},\\ \begin{bmatrix} 1 \\ 1 \end{bmatrix} & \text{if } 1+\frac{1}{\mathrm{M}}<\frac{r}{t}. \end{cases}$$

Note that, for this latter solution, no velocity u is specified in the vacuum. If, on the other hand, the fluid is not sheared as hard, i.e., if $0<\mathrm{M}<\frac{2}{\gamma-1}$, then no cavitation

takes place and the solution $\left[\begin{smallmatrix}\rho\\ u\end{smallmatrix}\right](r,t)$ is found to be

$$\begin{cases} \begin{bmatrix} 1 \\ -1 \end{bmatrix} & \text{if } \frac{r}{t} < -1 - \frac{1}{\mathrm{M}}, \\ \begin{bmatrix} \left(\frac{2}{\gamma+1} - \mathrm{M}\,\frac{\gamma-1}{\gamma+1}(1+\frac{r}{t})\right)^{2/(\gamma-1)} \\ \frac{1}{\mathrm{M}(\gamma+1)}\left(2 + (1-\gamma)\mathrm{M} + 2\mathrm{M}\frac{r}{t}\right) \end{bmatrix} & \text{if } -1 - \frac{1}{\mathrm{M}} < \frac{r}{t} < -\frac{1}{\mathrm{M}} + \frac{\gamma-1}{2}, \\ \begin{bmatrix} \left(1 - \frac{\mathrm{M}}{2}(\gamma-1)\right)^{\frac{2}{\gamma-1}} \\ 0 \end{bmatrix} & \text{if } -\frac{1}{\mathrm{M}} + \frac{\gamma-1}{2} < \frac{r}{t} < \frac{1}{\mathrm{M}} - \frac{\gamma-1}{2}, \\ \begin{bmatrix} \left(\frac{2}{\gamma+1} + \mathrm{M}\,\frac{\gamma-1}{\gamma+1}(-1+\frac{r}{t})\right)^{2/(\gamma-1)} \\ \frac{1}{\mathrm{M}(\gamma+1)}\left(-2 + (-1+\gamma)\mathrm{M} + 2\mathrm{M}\frac{r}{t}\right) \end{bmatrix} & \text{if } \frac{1}{\mathrm{M}} - \frac{\gamma-1}{2} < \frac{r}{t} < 1 + \frac{1}{\mathrm{M}}, \\ \begin{bmatrix} 1 \\ 1 \end{bmatrix} & \text{if } 1 + \frac{1}{\mathrm{M}} < \frac{r}{t}. \end{cases}$$

4.2. The multidimensional axisymmetric case. Similarity solutions to the Euler equations have also been considered for the 2D and 3D axisymmetric problems that are studied in this paper. Even though no closed form solutions can be found, the analysis reveals that solutions without swirls may exhibit cavitation if $\gamma > 1$ but not if $\gamma = 1$; see [43, section 7.4].

More precisely, let us concentrate on the case $\gamma > 1$. Following [43], self-similar solutions to (4.1), (4.2) are sought in the form

$$\rho = \rho(s), \qquad u = u(s), \qquad \text{where } s = \frac{t}{r}.$$

This leads to the ODE system

$$\rho_s = m\frac{\rho u(1-su)}{s^2c^2 - (1-su)^2}, \tag{4.4}$$

$$u_s = m\frac{sc^2 u}{s^2c^2 - (1-su)^2}, \tag{4.5}$$

$$\rho(0) = 1, \quad u(0) = 1, \tag{4.6}$$

where $c = \frac{1}{M}\rho^{\frac{\gamma-1}{2}}$. The generalization of the 2D results of [43] to 3D problems is straightforward. It is presented here for the sake of completeness. Introducing the variables

$$I = su \quad \text{and} \quad K = sc,$$

relations (4.4), (4.5) can be written

$$\frac{dI}{d\tau} = I\left((1-I)^2 - (1+m)K^2\right) \equiv I\,\mathcal{F}(I,K), \tag{4.7}$$

$$\frac{dK}{d\tau} = K\left((1-I)^2 - K^2 - \frac{m}{2}(\gamma-1)I(1-I)\right) \equiv K\,\mathcal{G}(I,K), \tag{4.8}$$

$$\frac{ds}{d\tau} = s\left((1-I)^2 - K^2\right), \tag{4.9}$$

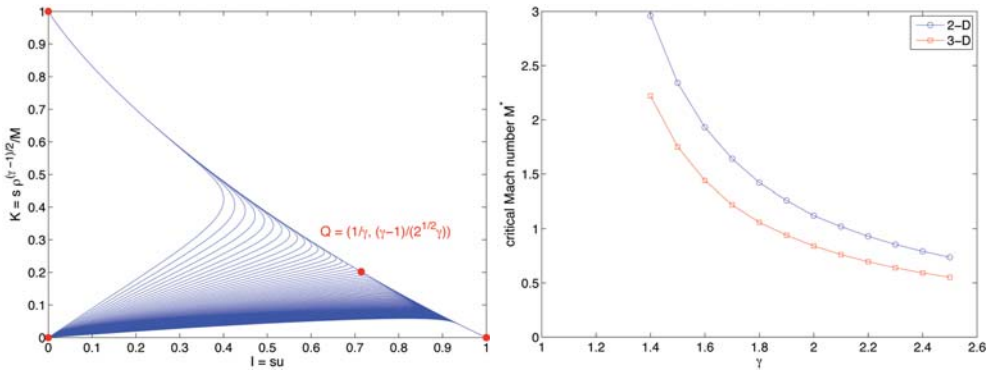


FIG. 4.1. *Left: Phase diagram* ($I = su$, $K = sc$) *of the solutions to* (4.7), (4.8), (4.9) *for* $m = 1$ *(2D) and* $\gamma = 1.4$. *Solutions "leaving"* $(0,0)$ *are represented for values of the Mach number* $M = 1.1, 1.2, 1.3, \ldots, 10$. *Right: Dependency of the critical Mach number* $M^\star$ *on* γ *for* $2D$ *and* $3D$ *problems as obtained through numerical integration of* (4.4), (4.5), (4.6). *Vacuum takes place if and only if* $M > M^\star$.

with the obvious definitions for $\mathcal{F}$ and $\mathcal{G}$ and where τ is a new independent variable defined by (4.9), which is introduced to make (4.7), (4.8) an autonomous system. In the (I, K)-phase plane, the stationary points of (4.7), (4.8) are found to be

$$(0,0), (1,0), (0,1) \quad \text{and} \quad Q = \left(\frac{2}{\gamma(1+m)+1-m}, \frac{1}{\sqrt{1+m}} \frac{\gamma(1+m)-1-m}{\gamma(1+m)+1-m} \right).$$

Let us consider the region $\Omega \subset \mathbb{R}^2$ defined by

$$\Omega = \Big\{ (I,K); I > 0, K > 0, \\ \mathcal{G} > 0 \text{ for } 0 < I \leq \frac{2}{\gamma(1+m)+1-m}, \\ \mathcal{F} > 0 \text{ for } \frac{2}{\gamma(1+m)+1-m} \leq I < 1 \Big\}.$$

A simple sign study along $\partial\Omega$ shows that Ω is invariant under (4.7), (4.8). The system (4.7), (4.8), (4.9) then has integral curves from $(0,0)$ to either $(0,1)$, $(1,0)$ or Q.

LEMMA 1. *Let* $m = 1$ *(two space dimensions) or* 2 *(three space dimensions) and let* $\gamma > 1$. *Then there exists* $M^\star = M^\star(m,\gamma) > 0$ *such that*

- *if* $M < M^\star$, *the solution* (ρ, u) *of* (4.4), (4.5), (4.6) *is defined for* $0 < s < \bar{s} < \infty$, *where* $\bar{s}$ *is such that* $(\rho(\bar{s}), u(\bar{s})) = (\bar{\rho}, 0)$ *(convergence to* $(I,K) = (0,1)$ *at a finite* s*-value, no vacuum)*;
- *if* $M > M^\star$, *the solution* (ρ, u) *of* (4.4), (4.5), (4.6) *is defined for* $0 < s < \bar{s} < \infty$, *where* $\bar{s}$ *is such that* $(\rho(\bar{s}), u(\bar{s})) = (0, \bar{u})$ *(convergence to* $(I,K) = (1,0)$ *at a finite* s*-value, vacuum)*;
- *if* $M = M^\star$, *the solution* (ρ, u) *of* (4.4), (4.5), (4.6) *is defined for* $0 < s < \infty$ *and converges to* $(I,K) = Q$ *as* $s \to \infty$ *(critical case).*

Proof. The proof follows in a straightforward way from [43, section 7.4] ∎

There does not appear to be an explicit formula for the critical value $M^\star$. However, equations (4.4), (4.5), (4.6) can be solved numerically and an approximate value of $M^\star$ inferred from the corresponding results. The feasibility of this approach is illustrated in Figure 4.1(left), which is in exact agreement with the above lemma.

The above observations provide a relatively easy way to numerically investigate vacuum formation for the multidimensional Euler equations. A standard ODE solver can be used to integrate the above equations. It is worth noting that the points $(I, K) = (0, 1)$ or $(1, 0)$ are reached for finite values of s; the solution can then be continuously extended toward $r = 0$. We omit the details. Figure 4.1(right) illustrates the dependency of $M^\star$ on γ for 2D and 3D problems as found through numerical investigation.

5. Discretization and numerical analysis. The numerical approach for solving the Navier–Stokes equations (3.9), (3.10), (3.11) under the specific assumptions considered here is based on a splitting between the Euler equations, on the one hand, and a diffusive equation, on the other hand. This way, one can take advantage of the similarity solutions considered in the previous section (which can be solved to a high degree of accuracy by ODE solvers; see below). At each space/time node, such a solution is locally constructed. This process is akin to numerically solving families of local Riemann problems, as is routinely done in many numerical schemes for hyperbolic conservation laws; see, e.g., [27].

The splitting algorithm is illustrated in the case of the barotropic equations (4.1), (4.2). Let (ρ_n, u_n) be the solution at some time $t_n = n\,\Delta t$, where Δt is the time step. To obtain the solution (ρ_{n+1}, u_{n+1}) at a later time $t_{n+1} = (n+1)\Delta t$, an "Euler step" is first taken; i.e., one solves (4.1), (4.2) from t_n to t_{n+1} with the initial condition

$$\rho(\cdot, t_n) = \rho_n, \qquad u(\cdot, t_n) = u_n.$$

The resulting solution at time t_{n+1} is denoted $(\rho^\star, u^\star)$. The following diffusive step is then taken:

$$\rho^\star u_t = \frac{1}{\mathrm{Re}} u_{\xi r}, \qquad t \in (t_n, t_{n+1}), \tag{5.1}$$

$$u(\cdot, t_n) = u^\star. \tag{5.2}$$

As both steps have to be solved numerically, their respective discretization is now described. For notational convenience, the diffusive step is described first.

5.1. The diffusive step. Equation (5.1) is discretized in space using Chebyshev collocation methods [3]. Such methods deliver high accuracy with a low number of nodes for smooth solutions (which are expected here for $t > 0$). To circumvent the coordinate singularity at $r = 0$ of the 2D and 3D problems, Chebyshev–Gauss–Radau nodes are used instead of the more common Chebyshev–Gauss–Lobatto nodes. In the spatial domain $(0, 1)$ those nodes have location

$$r_j = \frac{1}{2}\left(1 + \cos\left(\frac{2\pi j}{2N-1}\right)\right), \qquad j = 0, \ldots, N-1.$$

In each case, N stands for the number of nodes. For $r \in (0, 1)$ and $t > 0$, we seek an approximation u_N of u of the form

$$u_N(r, t) = \sum_{i=0}^{N-1} U_i(t)\psi_i(r),$$

where $\{\psi_i\}_{i=0}^{N-1}$ are the Lagrange interpolation polynomials at the Chebyshev–Gauss–Lobatto/Radau nodes on $[0, 1]$, i.e., $\psi_i(x_j) = \delta_{ij}$.

Interpolation at one of the above sets of nodes of a function $v = v(r,t)$ simply takes the form

$$I_N v(r,t) = \sum_{j=0}^{N-1} v(r_j,t)\psi_j(r).$$

By definition, the Chebyshev collocation derivative of v with respect to r at those nodes is then

$$\frac{\partial}{\partial r}(I_N v)(r_l,t) = \sum_{j=0}^{N-1} v(r_j,t)\psi_j'(r_l) = \sum_{j=0}^{N-1} D_{lj} v(r_j,t),$$

with $D_{lj} = \psi_j'(r_l)$. The collocation derivative at the nodes can then be obtained through matrix multiplication.

The discrete velocity u_N takes the form

$$u_N(r,t) = \sum_{i=0}^{N-1} U_i(t)\psi_i(r).$$

The semidiscretized in space problem (5.1) has the form

$$Z_U\left((I_N\rho^\star).*U'\right) = \left(D^2 + m\,\mathrm{diag}(1./R)D - m\,\mathrm{diag}\,1./(R.\hat{}\,2)\right)U + B_U, \tag{5.3}$$

with the obvious notation for U and D; the vector R is the node vector, $R_j = r_j$, $j = 0,\ldots,N-1$. The matrix Z_U zeroes the first and/or the last entry(ies) of a vector, and further B_U is a vector related to the boundary conditions (to be specified below). In the above equations, a "dotted operation" (for instance $.*$) refers to that operation being performed elementwise (for instance, $U.*V$ is the vector of ith component U_iV_i, $i = 0,\ldots,N-1$).

The computations are carried out with the boundary condition (3.16). In case the density vanishes (or becomes very small) at some node r_i at time t, i.e., $\mathcal{I}_N\rho^\star(r_i,t) = 0$, then the differential equation for $\mathcal{U}_i(t)$ degenerates into an algebraic equation; in other words, (5.3) becomes differential algebraic. How much of a numerical problem this is depends on the index of the system. The minimum number of times one has to differentiate all or part of (5.3) to recover an ODE system is the index of the differential algebraic equation (DAE) [1]. Here the index is easily found to be equal to 1. Indeed, differentiating the algebraic equation for $\mathcal{U}_i$ leads to an ODE, provided that the operator $D^2 + m\,\mathrm{diag}(1./R)D - m\,\mathrm{diag}\,1./(R.\hat{}\,2)$ with proper side conditions applied to U' is nonsingular. The side conditions are that the velocity is fixed on the outer boundary and thus $U_1'(t) = 0$ and that nonsingular solutions are sought. The above operator can be checked to be nonsingular by direct inspection of the matrices involved. Alternatively, at the continuous level, one can check that the only nonsingular solution to $\partial_{rr}\dot{u} + m/r\partial_r\dot{u} - m\dot{u}/r^2 = 0$ with $\dot{u}(1) = 0$ is the trivial solution $\dot{u} \equiv 0$ (where $\dot{u} = u_t$).

In the presence of vacuum, the system (5.3) is a DAE that is semiexplicit of index 1; as such it is amenable to relatively simple time discretization such as BDF [1]. Here, we used the MATLAB routine ODE15s, which implements a variant of BDF [35].

5.2. The Euler step. Let $(\rho_n(r_j), u_n(r_j))$ be given values for the density and velocity at time t^n at the node r_j, $j = 0,\ldots,N-1$. A family of N similarity solutions

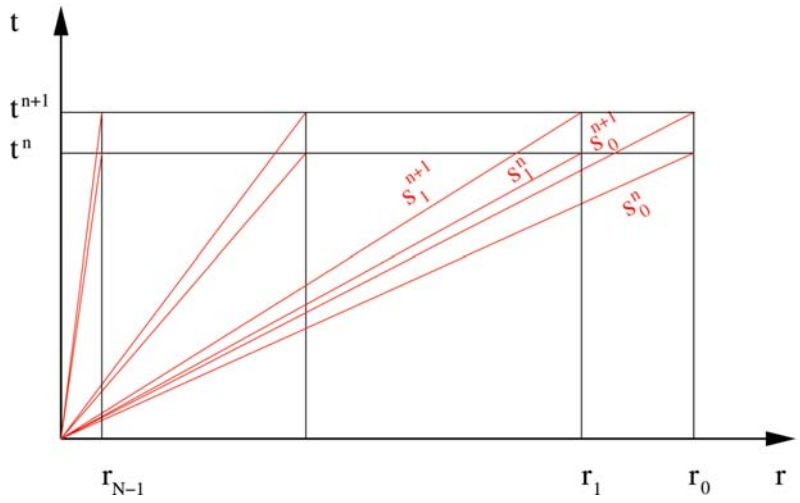


FIG. 5.1. *Meshes used for the diffusive step* $(\{r_j, t^n\}_{j=0,n=0}^{N-1,N_t})$ *and Euler step* $(\{s_j^n\}_{j=0,n=0}^{N-1,N_t})$.

is now defined in the same way as in section 4.1. More precisely, let $s_j^n = \frac{t^n}{r_j}$ be the slopes corresponding to the above data points, and let $s_j^{n+1} = \frac{t^{n+1}}{r_j}$ be the slopes at the next time step; see Figure 5.1. Then, for each $j = 0, \ldots, N-1$, equations (4.4), (4.5) are solved from $s = s_j^n$ to s_j^{n+1} with initial conditions

$$\rho(s_j^n) = \rho_n(r_j), \qquad u(s_j^n) = u_n(r_j). \tag{5.4}$$

5.3. The splitting algorithm. For given Mach and Reynolds numbers, and given spatial and temporal resolutions, i.e., N and Δt being chosen, the following steps are taken for N_t time steps:

- initialize ρ_0 and u_0 according to (3.15),
- for $n = 0$ to $N_t - 1$
  - for $j = N-1$ to 0 by -1
    - EULER: solve (4.4), (4.5), (5.4)
    - set $\rho^\star(r_j) = \rho(s_j^{n+1}) = \rho_{n+1}(r_j)$ and $u^\star(r_j) = u(s_j^{n+1})$
  - end
  - DIFFUSION: solve (5.1), (5.2) through (5.3)
  - set $u_{n+1}(r_j) = u(r_j, t_{n+1})$, $j = 0, \ldots, N-1$
- end

6. Numerical results. The problem (3.13), (3.14) has been solved using the method described in the previous section with various values of the physical parameters (Mach number M, Reynolds number Re, and adiabatic coefficient γ). The initial and boundary conditions are given by (3.15) and (3.16), respectively. The mesh size is fixed at $N = 32$ for all the results given below.

For all our examples, the numerical density ρ_N is an increasing function of r, and thus the numerical solution is probed at the node closest to the origin for vacuum detection. Vacuum formation, if it occurs, is expected to take place during a fast initial transient phase. A comparison with the inviscid case is instructive. If cavitation takes place for the Euler equations, it does so instantaneously (at the origin), i.e., for $t = 0+$. For the 1D case, this fact is obvious from the explicit self-similar solutions given in

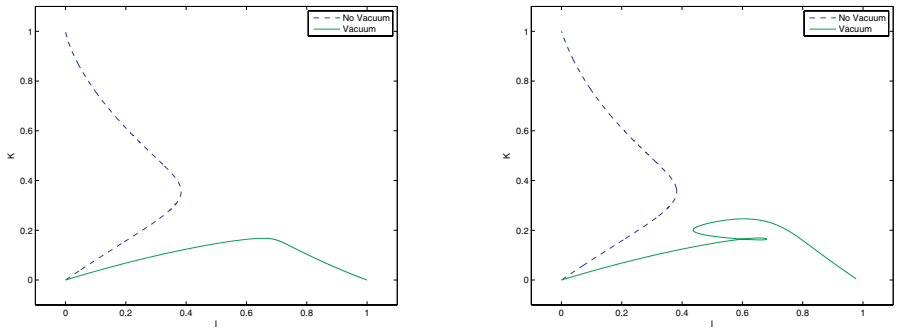


FIG. 6.1. *Phase diagram corresponding to the evolution of the solutions, at the node closest to the origin, of the Euler solution (left) and the Navier–Stokes solution (right) (see section* 4 *for definition of* I *and* K*). The parameters are taken as* $m = 2$ *(3D) and* $\mathrm{M} = 1.2$ *(no vacuum) and* $\mathrm{M} = 2.7$ *(vacuum); for the Navier–Stokes solution (right), the Reynolds number* Re *is* 10^6.

section 4. Regarding the Navier–Stokes equations, our efforts are concentrated on the multidimensional cases for which there is no node at the origin. For the node closest to the origin, phase space trajectories can be used in a way similar to what was done in Figure 4.1(left), giving a clear picture of the evolution of the density there. Using the same notation as in section 4, Figure 6.1 illustrates the difference between the evolution of the Euler solutions (left panel) and Navier–Stokes solutions (right panel).

Based on the above remarks, a specific calculation is said to lead to vacuum formation if for some time t, $0 < t < .005$,

$$\rho_N(r_{N-1}, t) < tol = 10^{-14}, \tag{6.1}$$

where $r_{N-1} = \frac{1}{2}(1 + \cos(\frac{2N-1}{2N-2}\pi))$ is the Chebyshev–Gauss–Radau node closest to the origin. In some cases, the time asymptotic behavior of the solution at r_{N-1} was not clear based on a phase plane analysis. Those cases are reported below as inconclusive, even if the density itself satisfied (6.1).

6.1. The 1D case. The above method cannot be used directly in the 1D case. Using an adapted method (details are omitted), no vacuum formation was numerically observed for the 1D Navier–Stokes solution as is shown in Figure 6.2. The absence of vacuum in the solutions to the 1D Navier–Stokes system is consistent with the solutions found by Hoff [15], who considers discontinuous data of the same type as in the present paper.

6.2. The multi-D barotropic case. The 2D and 3D barotropic flows (3.13), (3.14) with initial and boundary conditions (3.15), (3.16) are solved, for fixed values of the adiabatic coefficient γ, on grids in "Reynolds and Mach number space," i.e., for a collection of values of those two parameters.

Figure 6.3 corresponds to the value $\gamma = 1.4$ for the 2D and 3D cases, in the left and right panels, respectively. As explained above, the gray area corresponds to values of the parameters for which the numerical calculations were inconclusive, for instance due to the absence of a clear asymptotic behavior in phase space during the allotted computational time. It clearly illustrates that for large enough values of M there appears to be vacuum formation. The displayed results are relatively insensitive to

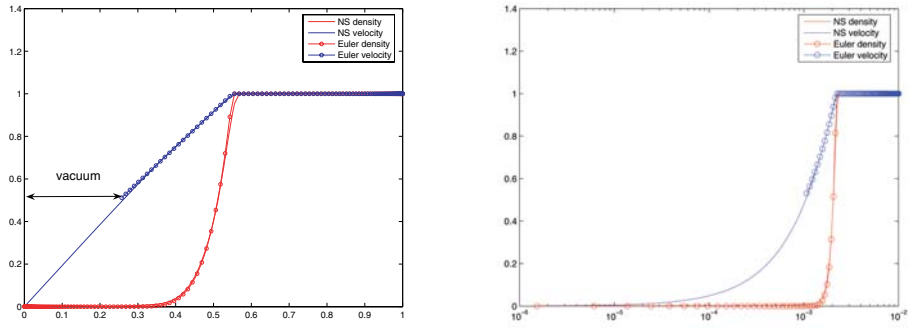


FIG. 6.2. *1D Navier–Stokes and Euler solutions for* $\gamma = 1.4$, $\mathrm{M} = 10$. *Left: Solutions at time* $t = 0.5$ *with* $\mathrm{Re} = 10{,}000$ *(for the Navier–Stokes flow). Right: Solutions at time* $t = .002$ *with* $\mathrm{Re} = 1{,}000{,}000$ *(for the Navier–Stokes flow). For these values, both the 1D Euler solution and the multi-D Navier–Stokes solution exhibit cavitation (see Figure* 6.3 *below), while the 1D Navier–Stokes solution does not.*

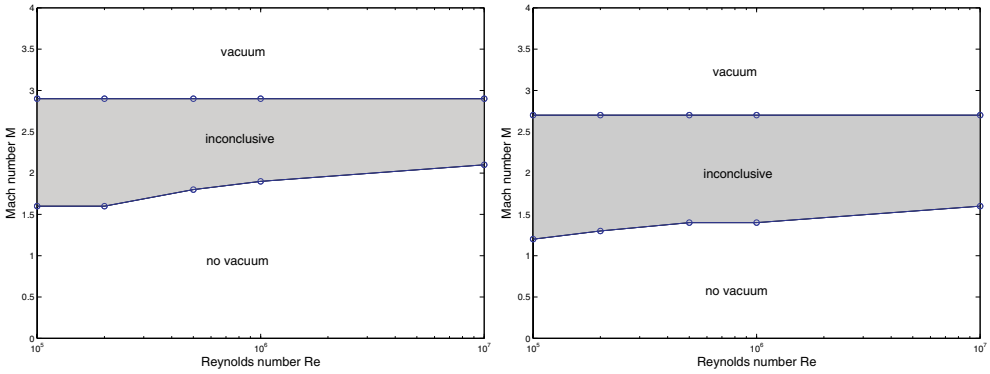


FIG. 6.3. *Vacuum formation for 2D (left) and 3D (right) barotropic flows* (3.13), (3.14) *with initial and boundary conditions* (3.15), (3.16) *and* $\gamma = 1.4$.

the discretization parameters. It is also observed that vacuum is more easily formed in three than in two dimensions, in agreement with the remarks in section 2.1.2.

Similar results were observed for larger values of γ: it gets easier to create vacuum as γ increases.

6.3. The full system. As mentioned in section 4.2, the construction of self-similar solutions for the nonisentropic Euler equations (3.1)–(3.3) appears to be open. Therefore, we do not have the benefit of using this tool as part of the numerical algorithm. While more research is needed, preliminary calculations based on an unsplit algorithm indicate vacuum formation here as well.

7. Conclusion. Our numerical results indicate that vacuum formation is possible in solutions of the multidimensional compressible Navier–Stokes equations. This applies to discontinuous and sufficiently large data where the initial density is uniformly bounded away from zero. The same numerical code gives results for one-dimensional flow that are in agreement with the known analytical results. The conclusions do not contradict the currently known results for multi-D flow.

No attempt was made to follow the solutions *past* vacuum formation. The present study also leaves unanswered the issue of whether vacuum formation is instantaneous, as is the case for the corresponding solutions to the Euler equations.

Acknowledgments. The last author is indebted to David Hoff for several discussions about the vacuum problem in compressible flow. The clarity and accuracy of this article were significantly improved by the criticism of an anonymous referee.

REFERENCES

[1] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Classics in Applied Mathematics 14, SIAM, Philadelphia, 1995.

[2] D. Bresch and B. Desjardins, *Stabilité de solutions faibles globales pour les équations de Navier-Stokes compressible avec température* [Stability of global weak solutions for the Navier-Stokes equations modelling compressible and heat-conducting fluids], C. R. Math. Acad. Sci. Paris, 343 (2006), pp. 219–224 (in French).

[3] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer Ser. Comput. Dynam., Springer, Berlin, 1987.

[4] C. M. Dafermos, *Global smooth solutions to the initial-boundary value problem for the equations of one-dimensional nonlinear thermoviscoelasticity*, SIAM J. Math. Anal., 13 (1982), pp. 397–408.

[5] C. M. Dafermos and L. Hsiao, *Global smooth thermomechanical processes in one-dimensional nonlinear thermoviscoelasticity*, Nonlinear Anal., 6 (1982), pp. 435–454.

[6] R. Danchin, *Global existence in critical spaces for compressible Navier-Stokes equations*, Invent. Math., 141 (2000), pp. 579–614.

[7] R. Danchin, *Global existence in critical spaces for compressible viscous and heat conductive gases*, Arch. Ration. Mech. Anal., 160 (2001), pp. 1–39.

[8] R. Danchin, *On the uniqueness in critical spaces for compressible Navier-Stokes equations*, NoDEA, Nonlinear Differential Equations Appl., 12 (2005), pp. 111–128.

[9] R. Duan and Y. Zhao, *A note on the non-formation of vacuum states for compressible Navier-Stokes equations*, J. Math. Anal. Appl., 311 (2005), pp. 744–754.

[10] E. Feireisl, A. Novotný, and H. Petzeltová, *On the existence of globally defined weak solutions to the Navier-Stokes equations*, J. Math. Fluid Mech., 3 (2001), pp. 358–392.

[11] E. Feireisl, *Dynamics of Viscous Compressible Fluids*, Oxford University Press, London, 2004.

[12] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II: *Stiff and Differential Algebraic Problems*, Springer Ser. Comput. Math. 14, Springer, Berlin, 1996.

[13] D. Hoff, *Spherically symmetric solutions of the Navier-Stokes equations for compressible, isothermal flow with large, discontinuous initial data*, Indiana Univ. Math. J., 41 (1992), pp. 1225–1302.

[14] D. Hoff, *Discontinuous solutions of the Navier-Stokes equations for multidimensional flows of heat-conducting fluids*, Arch. Ration. Mech. Anal., 139 (1997), pp. 303–354.

[15] D. Hoff, *Global solutions of the equations of one-dimensional, compressible flow with large data and forces, and with differing end states*, Z. Angew. Math. Phys., 49 (1998), pp. 774–785.

[16] D. Hoff, *Uniqueness of weak solutions of the Navier–Stokes equations of multidimensional, compressible flow*, SIAM J. Math. Anal., 37 (2006), pp. 1742–1760.

[17] D. Hoff and H. K. Jenssen, *Symmetric nonbarotropic flows with large data and forces*, Arch. Ration. Mech. Anal., 173 (2004), pp. 297–343.

[18] D. Hoff and D. Serre, *The failure of continuous dependence on initial data for the Navier–Stokes equations of compressible flow*, SIAM J. Appl. Math., 51 (1991), pp. 887–898.

[19] D. Hoff and J. Smoller, *Non-formation of vacuum states for compressible Navier-Stokes equations*, Comm. Math. Phys., 216 (2001), pp. 255–276.

[20] S. Jiang and P. Zhang, *Global weak solutions to the Navier-Stokes equations for a* 1*D viscous polytropic ideal gas*, Quart. Appl. Math., 61 (2003), pp. 435–449.

[21] Ya. I. Kanel, *A model system of equations for the one-dimensional motion of a gas*, Differ. Uravn., 4 (1968), pp. 721–734.

[22] S. Kawashima, *Systems of Hyperbolic-Parabolic Composite Type, with Applications to the Equations of Magnetohydrodynamics*, Ph.D. thesis, Department of Mathematics, Kyoto University, Kyoto, Japan, 1983.

[23] S. Kawashima and T. Nishida, *Initial-boundary value problems for the equations of motion of compressible viscous and heat-conductive fluids*, Comm. Math. Phys., 89 (1983), pp. 445–464.

[24] B. Kawohl, *Global existence of large solutions to initial boundary value problems for a viscous, heat-conducting, one-dimensional real gas*, J. Differential Equations, 58 (1985), pp. 76–103.

[25] A. V. Kazhikhov, *Sur la solubilité globale des problèmes monodimensionnels aux valeurs initiales-limitées pour les équations d'un gaz visqueux et calorifère*, C. R. Acad. Sci. Paris, Sér. A, 284 (1977), pp. 317–320 (in French).

[26] A. V. Kazhikhov and V. V. Shelukhin, *Unique global solution with respect to time of initial-boundary value problems for one-dimensional equations of a viscous gas*, J. Appl. Math. Mech., 41 (1977), pp. 273–282; translated from Prikl. Mat. Meh., 41 (1977), pp. 282–291 (in Russian).

[27] R. J. Leveque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, UK, 2002.

[28] P. L. Lions, *Mathematical Topics in Fluid Mechanics, Vol.* 2*, Compressible Models*, Oxford University Press, London, 1998.

[29] T. P. Liu, Z. Xin, and T. Yang, *Vacuum states for compressible flow*, Discrete Contin. Dynam. Systems, 4 (1998), pp. 1–32.

[30] T. Luo, Z. Xin, and T. Yang, *Interface behavior of compressible Navier-Stokes equations with vacuum*, SIAM J. Math. Anal., 31 (2000), pp. 1175–1191.

[31] A. Matsumura and T. Nishida, *The initial value problem for the equations of motion of viscous and heat-conductive gases*, J. Math. Kyoto U., 20 (1980), pp. 67–104.

[32] A. Mellet and A. Vasseur, *Existence and Uniqueness of Global Strong Solutions for One-Dimensional Compressible Navier-Stokes Equations*, preprint, Mathematics Department, University of Texas at Austin, 2006.

[33] A. Novotný and I. Straškraba, *Introduction to the Mathematical Theory of Compressible Flow*, Oxford University Press, London, 2004.

[34] J. Serrin, *Mathematical principles of classical fluid mechanics*, in Handbuch der Physik, Bd. 8/1, Strömungsmechanik I, Springer-Verlag, Berlin, 1959, pp. 125–263.

[35] L. F. Shampine, M. W. Reichelt, and J. A. Kierzenka, *Solving index-*1 *DAEs in MATLAB and Simulink*, SIAM Rev., 41 (1999), pp. 538–552.

[36] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Grundlehren Math. Wiss. 258, Springer-Verlag, New York, 1994.

[37] Z. Xin, *Blowup of smooth solutions to the compressible Navier-Stokes equation with compact density*, Comm. Pure Appl. Math., 51 (1998), pp. 229–240.

[38] Z. Xin and H. Yuan, *Vacuum state for spherically symmetric solutions of the compressible Navier-Stokes equations*, J. Hyperbolic Differential Equations, 3 (2006), pp. 403–442.

[39] T. Yang, Z. Yao, and C. Zhu, *Compressible Navier-Stokes equations with density-dependent viscosity and vacuum*, Comm. Partial Differential Equations, 26 (2001), pp. 965–981.

[40] T. Yang and H. Zhao, *A vacuum problem for the one-dimensional compressible Navier-Stokes equations with density-dependent viscosity*, J. Differential Equations, 184 (2002), pp. 163–184.

[41] T. Yang and C. Zhu, *Compressible Navier-Stokes equations with degenerate viscosity coefficient and vacuum*, Comm. Math. Phys., 230 (2002), pp. 329–363.

[42] S.-W. Vong, T. Yang, and C. Zhu, *Compressible Navier-Stokes equations with degenerate viscosity coefficient and vacuum.* II, J. Differential Equations, 192 (2003), pp. 475–501.

[43] Y. Zheng, *Systems of Conservation Laws: Two-Dimensional Riemann Problems*, Birkhäuser Boston, Cambridge, MA, 2001.

SIAM J. APPL. MATH.
Vol. 67, No. 6, pp. 1693-1703

HOPF BIFURCATION IN DIFFERENTIAL EQUATIONS WITH DELAY FOR TUMOR–IMMUNE SYSTEM COMPETITION MODEL*

RADOUANE YAFIA†

Abstract. This paper deals with the qualitative analysis of the solutions to a model that refers to the competition between the immune system and an aggressive host such as a tumor. The model which describes this competition is governed by a system of differential equations with one delay. It is shown that the dynamics depends crucially on the time delay parameter. By using the time delay as a parameter of bifurcation, the analysis is focused on the Hopf bifurcation problem to predict the occurrence of a limit cycle bifurcating from the nontrivial steady state. The obtained results depict the oscillations, given by simulations (see [M. Galach, *Int. J. Appl. Math. Comput. Sci.*, 13 (2003), pp. 395–406]), which are observed in reality (see [D. Kirschner and J. C. Panetta, *J. Math. Biol.*, 37 (1998), pp. 235–252]). It is suggested to examine by laboratory experiments how to employ these results for control of tumor growth.

Key words. tumor–immune system competition, delayed differential equations, Hopf bifurcation, periodic solutions

AMS subject classification. 43K18

DOI. 10.1137/060657947

1. Introduction.

We consider a model concerning the competition of tumor cells with the immune system. The modeling approach, proposed by many authors, uses ordinary and delayed differential equations; see [14, 16, 19, 20, 24, 27]. Other authors use kinetic equations that give a complex description, at the cellular scale, in comparison with other, simpler models. Kinetic models are needed to describe heterogeneity of virulence; see [1, 2, 3, 4, 9, 26].

Modeling in other fields of biology also uses kinetic equations; for instance, [11] develops a kinetic theory approach to describe population dynamics, while [3] deals with the development of suitable general mathematical structures including a large variety of Boltzmann-type models. Other authors use models based on partial differential equations corresponding to population dynamics of cells with internal structure [22] (however, not heterogeneous) or models based on interacting agents [21].

The reader interested in a more complete bibliography about the evolution of a cell, and the pertinent role of cellular phenomena in directing the body toward recovery or toward illness, is referred to [13, 17]. A detailed description of virus, antivirus, and body dynamics can be found in [8, 12, 23, 25].

The mathematical model under consideration was proposed in a recent paper by Galach [16], who proposed a simple model, with one delay, of tumor–immune system competition. The idea is inspired from [20]. He also refers to numerical results in [20] to compare them with those obtained in his paper [16].

The mathematical analysis of this present paper is motivated by experimental and numerical results; see [19, 16], respectively. These results give evidence that the oscillating state of the tumor is more desirable, from the medical point of view, than the monotonically growing state, presumably because oscillations prolong the nonterminal phase of disease. We can ask what are the conditions and feedback loops

*Received by the editors April 22, 2006; accepted for publication (in revised form) May 23, 2007; published electronically September 26, 2007.
http://www.siam.org/journals/siap/67-6/65794.html
†Faculté Polydisciplinaire, Université Ibn Zohr, B.P:638, Ouarzazate, Morocco (yafia1@yahoo.fr).

which make oscillations possible and whether the tumor can be preserved in such an oscillating state by therapeutic means for an indefinite time. This leads to study of the model in terms of a time delay ordinary differential equation system. We believe that it makes sense to ask qualitative questions about the behavior of the system locally with respect to time, such as questioning the existence and stability of fixed points and the existence of Hopf bifurcation. Being interested in periodic or quasi-periodic behavior of the tumor, we investigate the model with respect to existence of the Hopf point.

This paper is organized as follows. In section 2, we introduce the model. In section 3, we establish some results on the stability of the possible steady states (trivial and nontrivial) of the delayed system (2.4). The existence of a critical value of the delay in which the nontrivial steady state changes stability is investigated. The main result of this paper is given in section 4. Based on the Hopf bifurcation theorem, we show the occurrence of Hopf bifurcation when the delay crosses some critical value. Section 5 is devoted to a numerical application for the parameter values of system (2.4). In section 6, we give short discussions.

2. Mathematical model. When unknown tissues, organisms, or tumor cells appear in a body the immune system tries to identify them and, if successful, attempts to eliminate them. The immune system response consists of two different interacting responses: the cellular response and the humoral response. The cellular response is carried by T lymphocytes. The humoral response is related to the class of cells called B lymphocytes. A dynamics of the antitumor immune response in vivo is complicated and not well understood.

The immune response begins when tumor cells are identified. Then tumor cells are caught by macrophages, which are found in all tissues in the body and circulate in the blood stream. Macrophages absorb tumor cells, destroy them, and release a series of cytokines which activate T helper cells (i.e., a subpopulation of T lymphocytes). These latter cells coordinate the counterattack. T helper cells can also be directly stimulated to interact with antigens. These helper cells cannot kill tumor cells, but they send urgent biochemical signals to a special type of T lymphocytes called natural killers (NKs). T cells begin to multiply and release other cytokines that further stimulate more T cells, B cells, and NK cells. As the number of B cells increases, T helper cells send a signal to start the production of antibodies. Antibodies circulate in the blood and are attached to tumor cells, which implies that the tumor cells are more quickly engulfed by macrophages or killed by NK cells. Like all T cells, NK cells are programmed to recognize one specific type of infected cell or cancer cell. NK cells are lethal and constitute a critical line of the defense.

The model proposed in [20] describes the response of effector cells (ECs) to the growth of tumor cells (abbreviated TCs from here on). This model differs from others because it takes into account the penetration of TCs by ECs, which simultaneously causes the inactivation of ECs. It is assumed that interactions between ECs and TCs in vitro can be described by the kinetic scheme shown in Figure 1, where E, T, C,

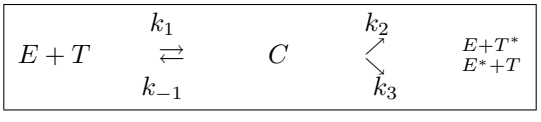


FIG. 1. *Kinetic scheme describing interactions between ECs and TCs* (*see* [16]).

E^*, and T^* are the local concentrations of ECs, TCs, EC–TC complexes, inactivated ECs, and "lethally hit" TCs, respectively. k_1 and k_{-1} denote the rates of bindings of ECs to TCs and the detachment of ECs from TCs without damaging TCs, k_2 is the rate at which EC–TC interactions program TCs for lysis, and k_3 is the rate at which EC–TC interactions inactivate ECs.

Kuznetsov and Taylor's model [20] is as follows:

$$\begin{cases} \frac{dE}{dt} = s + F(C,T) - d_1 E - k_1 ET + (k_{-1} + k_2)C, \\ \frac{dT}{dt} = aT(1 - bT) - k_1 ET + (k_{-1} + k_3)C, \\ \frac{dC}{dt} = k_1 ET - (k_{-1} + k_2 + k_3)C, \\ \frac{dE^*}{dt} = k_3 C - d_2 E^*, \\ \frac{dT^*}{dt} = k_2 C - d_3 T^*, \end{cases} \tag{2.1}$$

where s is the normal (i.e., not increased by the presence of the tumor) rate of the flow of adult ECs into the tumor site. $F(C,T)$ describes the accumulation of ECs in the tumor site; d_1, d_2, and d_3 are the coefficients of the processes of destruction and migration for E, E^*, and T^*, respectively; a is the coefficient of the maximal growth of tumor; and b is the environment capacity.

It is claimed in [20] that experimental observations motivate the approximation $\frac{dC}{dt} \approx 0$. Therefore, it is assumed that $C \approx KET$, where $K = \frac{k_1}{k_2 + k_3 + k_{-1}}$, and the model can be reduced to two equations which describe the behavior of ECs and TCs only. Moreover, in [16] it is suggested that the function F should be in the form $F(C,T) = F(E,T) = \theta ET$. Therefore, the model (2.1) takes the form

$$\begin{cases} \frac{dE}{dt} = s + \alpha_1 ET - dE, \\ \frac{dT}{dt} = aT(1 - bT) - nET, \end{cases} \tag{2.2}$$

where $\alpha_1 = \theta - m$, and a, b, s have the same meanings as $n = Kk_2$, $m = Kk_3$, and $d = d_1$, respectively, in (2.1). All coefficients except α_1 are positive. The sign of α_1 depends on the relation between θ and m. If the stimulation coefficient of the immune system exceeds the neutralization coefficient of ECs during the formation of EC–TC complexes, then $\alpha_1 > 0$. We use the dimensionless form of model (2.2),

$$\begin{cases} \frac{dx}{dt} = \sigma + \omega xy - \delta x, \\ \frac{dy}{dt} = \alpha y(1 - \beta y) - xy, \end{cases} \tag{2.3}$$

where x denotes the dimensionless density of ECs; y stands for the dimensionless density of the population of TCs; $\alpha = \frac{a}{nT_0}$, $\beta = bT_0$, $\delta = \frac{d}{nT_0}$, $\sigma = \frac{s}{nE_0T_0}$, and $\omega = \frac{\alpha_1}{n}$ represent the immune response to the appearance of the tumor cells; and E_0 and T_0 are the initial conditions. The existence, uniqueness, and nonnegativity of solutions are analyzed in [16], and the nonexistence of a nonnegative periodic solution of system (2.3) is proved. The existence and stability of a periodic solution of system (2.3) are studied in [28].

For $\omega > 0$ and $\alpha\delta < \sigma$, system (2.3) has one nonnegative steady state P_0, which is stable, and for $\omega > 0$ and $\alpha\delta > \sigma$ (2.3) has two possible nonnegative steady states P_0 and P_2, where the first is unstable and the second is stable (see [16]).

The delayed mathematical model corresponding to (2.3) is given by the following system [16]:

$$\begin{cases} \frac{dx}{dt} = \sigma + \omega x(t-\tau)y(t-\tau) - \delta x, \\ \frac{dy}{dt} = \alpha y(1-\beta y) - xy, \end{cases} \tag{2.4}$$

where the parameter τ is the time delay which the immune system needs to develop a suitable response after the recognition of nonself cells (see [16]). Time delays in connection with tumor growth also appear in [5, 6, 7, 15].

The existence and uniqueness of solutions of system (2.4) for every $t > 0$ are established in [16], and in the same paper it is shown that

(1) if $\omega \geq 0$, these solutions are nonnegative for any nonnegative initial conditions (biologically realistic case);

(2) if $\omega < 0$, there exist nonnegative initial conditions such that the solution becomes negative in a finite time interval.

The existence and stability of periodic solution of system (2.4) when $\omega < 0$ are studied in [29, 30], respectively.

Our goal in this paper is to consider case (1) when $\omega > 0$ (this case corresponds to the existence of a nonnegative solution), which is the most biologically meaningful one. We study the asymptotic behavior of the possible steady states P_0 and P_2 with respect to the delay τ. We establish that the Hopf bifurcation may occur by using the delay as a parameter of bifurcation. This result is depicted numerically.

3. Steady states and stability for positive delays. Consider the system (2.4), and suppose that $\omega > 0$. We distinguish between two cases, $\alpha\delta < \sigma$ and $\alpha\delta > \sigma$.

Let the following hypotheses hold:

(A_1) $\omega > 0$ and $\alpha\delta < \sigma$.

(A_2) $\omega > 0$ and $\alpha\delta > \sigma$.

Under hypothesis (A_1), system (2.4) has a unique positive equilibrium P_0 given by $P_0 = (\frac{\sigma}{\delta}, 0)$, and the linearized system around P_0 takes the form

$$\begin{cases} \frac{dx}{dt} = \omega\frac{\sigma}{\delta}y(t-\tau) - \delta x, \\ \frac{dy}{dt} = (\alpha - \frac{\sigma}{\delta})y, \end{cases} \tag{3.1}$$

which leads to the characteristic equation

$$W(\lambda) = \left(\lambda + \frac{\sigma}{\delta} - \alpha\right)(\lambda + \delta). \tag{3.2}$$

Therefore, we have the following result.

PROPOSITION 3.1. *Assume* (A_1). *The equilibrium point* P_0 *is asymptotically stable for all* $\tau > 0$.

Proof. The characteristic equation (3.2) has two roots $\lambda_1 = -\frac{\sigma}{\delta} + \alpha$ and $\lambda_2 = -\delta$ which are independent of τ and are negative. From [18], the equilibrium point P_0 is asymptotically stable for all $\tau > 0$. □

Under hypothesis (A_2), system (2.4) has two equilibrium points $P_0 = (\frac{\sigma}{\delta}, 0)$ and $P_2 = (x_2, y_2)$, where

$$x_2 = \frac{-\alpha(\beta\delta - \omega) + \sqrt{\Delta}}{2\omega}, \qquad y_2 = \frac{\alpha(\beta\delta + \omega) - \sqrt{\Delta}}{2\alpha\beta\omega},$$

with $\Delta = \alpha^2(\beta\delta - \omega)^2 + 4\alpha\beta\sigma\omega$.

From the characteristic equation (3.2), we state the following result.

PROPOSITION 3.2. *Assume* (A_2). *Then, the equilibrium point P_0 is unstable for all $\tau > 0$.*

Proof. The characteristic equation (3.2) has two roots $\lambda_1 = -\frac{\sigma}{\delta} + \alpha$ and $\lambda_2 = -\delta$ which are independent of τ. As $\alpha\delta > \sigma$, we have $\lambda_1 > 0$. From [18], the equilibrium point P_0 is unstable for all $\tau > 0$. □

In the next sections, we shall study the stability of the nontrivial equilibrium point P_2.

Let $z(t) = (u(t), v(t)) = (x(t), y(t)) - (x_2, y_2)$; then system (2.4) is written as a functional differential equation in $C := C([-\tau, 0], \mathbb{R}^2)$,

$$\frac{dz(t)}{dt} = L(\tau)z_t + f(z_t, \tau), \tag{3.3}$$

where $L(\tau) : C \longrightarrow \mathbb{R}^2$ (a linear operator) and $f : C \times \mathbb{R} \longrightarrow \mathbb{R}^2$ are given, respectively, by

$$L(\tau)\varphi = \begin{pmatrix} \omega y_2\varphi_1(-\tau) + \omega x_2\varphi_2(-\tau) - \delta\varphi_1(0) \\ -y_2\varphi_1(0) + (\alpha - 2\alpha\beta y_2 - x_2)\varphi_2(0) \end{pmatrix}$$

and

$$f(\varphi, \tau) = \begin{pmatrix} \sigma + \omega\varphi_1(-\tau)\varphi_2(-\tau) + \omega x_2 y_2 - \delta x_2 \\ -\alpha\beta\varphi_2^2(0) + \alpha y_2 - \alpha\beta y_2^2 - \varphi_1(0)\varphi_2(0) - x_2 y_2 \end{pmatrix}$$

for $\varphi = (\varphi_1, \varphi_2) \in C$ and $z_t \in C$, defined by $z_t(\theta) = z(t + \theta)$, for all $\theta \in [-\tau, 0]$.

The characteristic equation of linearized equation

$$\frac{dz(t)}{dt} = L(\tau)z_t \tag{3.4}$$

has the form

$$W(\lambda, \tau) = \lambda^2 + p\lambda + r + (s\lambda + q)e^{-\lambda\tau} = 0, \tag{3.5}$$

where $p = \delta + \alpha\beta y_2 > 0$, $r = \delta\alpha\beta y_2 > 0$, $s = -\omega y_2 < 0$, and $q = \alpha\omega y_2(1 - 2\beta y_2) > 0$.

The stability of the equilibrium point P_2 is a result of the localization of the roots of (3.5); then we have the following theorem.

THEOREM 3.3. *Assume* (A_2) *and $\alpha > \sup(\frac{\omega}{\beta}, \frac{\sigma}{\delta})$. Then, there exist $\beta_1 > 0$ and $\tau_l > 0$ such that, for all $0 < \beta < \beta_1$, the nontrivial steady state P_2 is asymptotically stable for $\tau < \tau_l$ and unstable for $\tau > \tau_l$, where*

$$\tau_l = \frac{1}{\zeta_l} \arccos\left\{ \frac{q(\zeta_l^2 - r) - ps\zeta_l^2}{s^2\zeta_l^2 + q^2} \right\}, \tag{3.6}$$

$$\zeta_l^2 = \frac{1}{2}(s^2 - p^2 + 2r) + \frac{1}{2}[(s^2 - p^2 + 2r)^2 - 4(r^2 - q^2)]^{\frac{1}{2}}, \tag{3.7}$$

$$\beta_1 = \frac{2\omega(2\sigma - \alpha\delta) - \sqrt{\Delta_1}}{2\alpha^2\delta^2} > 0, \tag{3.8}$$

and

$$\Delta_1 = 4\alpha^2\omega^2(4\sigma^2 - 4\alpha\delta\sigma + 2\alpha^2\delta^2) > 0. \tag{3.9}$$

For the proof of Theorem 3.3, we need Lemma 3.4.

Consider the equation

$$\lambda^2 + p\lambda + r + (s\lambda + q)e^{-\lambda\tau} = 0, \tag{3.10}$$

where p, r, q, and s are real numbers.

Let the following hypotheses hold:

(H_1) $p + s > 0$.

(H_2) $q + r > 0$.

(H_3) $r^2 - q^2 < 0$ or ($s^2 - p^2 + 2r > 0$ and $(s^2 - p^2 + 2r)^2 = 4(r^2 - q^2)$).

LEMMA 3.4 (see [10]). *If* (H_1)–(H_3) *hold, then when* $\tau \in [0, \tau_l)$ *all roots of* (3.10) *have negative real parts; when* $\tau = \tau_l$ (3.10) *has a pair of purely imaginary roots* $\pm i\zeta_l$*; and when* $\tau > \tau_l$ (3.10) *has at least one root with positive real part, where* τ_l *and* ζ_l *are defined as in Theorem* 3.3.

Proof of Theorem 3.3. From the hypothesis $\alpha > \sup(\frac{\omega}{\beta}, \frac{\sigma}{\delta})$ and the expressions of p, q, s, and r, we have $p + s > 0$ and $q + r > 0$. Therefore, hypotheses (H_1) and (H_2) are satisfied.

Then, all roots of the characteristic equation (3.5) have negative real parts for $\tau = 0$, and the steady state P_2 is asymptotically stable for $\tau = 0$.

By Rouche's theorem, it follows that the roots of (3.5) have negative real parts for some critical value of the delay τ.

We want to determine if the real part of some root increases to reach zero and eventually becomes positive as τ varies.

If $i\zeta$ is a root of (3.5), then

$$-\zeta^2 + ip\zeta + is\zeta(\cos(\tau\zeta) + i\sin(\tau\zeta)) + r + q(\cos(\tau\zeta) + i\sin(\tau\zeta)) = 0. \tag{3.11}$$

Separating the real and imaginary parts, we have

$$\begin{cases} -\zeta^2 + r = -q\cos(\tau\zeta) + s\zeta\sin(\tau\zeta), \\ p\zeta = -s\zeta\cos(\tau\zeta) - q\sin(\tau\zeta). \end{cases} \tag{3.12}$$

It follows that ζ satisfies

$$\zeta^4 - (s^2 - p^2 + 2r)\zeta^2 + (r^2 - q^2) = 0. \tag{3.13}$$

The two roots of the above equation can be expressed as follows:

$$\zeta^2 = \frac{1}{2}(s^2 - p^2 + 2r) \pm \frac{1}{2}[(s^2 - p^2 + 2r)^2 - 4(r^2 - q^2)]^{\frac{1}{2}}. \tag{3.14}$$

As

$$r^2 - q^2 = \alpha^2 y_2^2(\delta^2\beta^2 - \omega^2(1 - 2\beta y_2)^2),$$

the sign of $r^2 - q^2$ is deduced from the sign of

$$(\delta\beta - \omega(1 - 2\beta y_2)) = \frac{2\alpha\beta\delta - \sqrt{\Delta}}{\alpha}, \tag{3.15}$$

which is negative for $0 < \beta < \beta_1$, where β_1 is given by (3.8). Because the discriminant of (3.15) is positive for all $\alpha > 0$, we have $\delta, \sigma > 0$.

Therefore, we have $r^2 - q^2 < 0$, and hypothesis (H_3) is satisfied, which concludes the proof of Theorem 3.3. □

4. Hopf bifurcation occurrence. We apply the Hopf bifurcation theorem to show the existence of a nontrivial periodic solution of system (3.3), for suitable values of parameter delay, and use the Hopf bifurcation as a bifurcation parameter. Therefore, the periodicity is a result of changing the type of stability from a stable stationary solution to a limit cycle.

In what follows, we recall the formulation of the Hopf bifurcation theorem for retarded differential equations. Let the equation

$$\frac{dx(t)}{dt} = F(\alpha, x_t) \tag{4.1}$$

hold with $F : \mathbb{R} \times C \longrightarrow \mathbb{R}^n$, F of class $\mathcal{C}^k$, $k \geq 2$, $F(\alpha, 0) = 0$ for all $\alpha \in \mathbb{R}$, and $C = C([-r, 0], \mathbb{R}^n)$ the space of continuous functions from $[-r, 0]$ into $\mathbb{R}^n$. As usual, x_t is the function defined from $[-r, 0]$ as $\mathbb{R}^n$ by $x_t(\theta) = x(t + \theta)$, $r \geq 0$, and $n \in \mathbb{N}^*$.

The following assumptions are stated:

(M_0) F is of class $\mathcal{C}^k$, $k \geq 2$, $F(\alpha, 0) = 0$ for all $\alpha \in \mathbb{R}$, and the map $(\alpha, \varphi) \longrightarrow D^k_\varphi F(\alpha, \varphi)$ sends bounded sets into bounded sets.

(M_1) The characteristic equation

$$\Delta(\alpha, \lambda) = \det(\lambda Id - D_\varphi F(\alpha, 0) \exp(\lambda(.)Id)) \tag{4.2}$$

of the linearized equation of (4.1) around the equilibrium $v = 0$,

$$\frac{dv(t)}{dt} = D_\varphi F(\alpha, 0) v_t, \tag{4.3}$$

has in $\alpha = \alpha_0$ a simple imaginary root $\lambda_0 = \lambda(\alpha_0) = i$; all other roots λ satisfy $\lambda \neq m\lambda_0$ for $m \in \mathbb{Z}$.

(M_2) $\lambda(\alpha)$ is the branch of roots passing through λ_0, and we have

$$\frac{\partial}{\partial \alpha} \operatorname{Re} \lambda(\alpha)_{/\alpha=\alpha_0} \neq 0. \tag{4.4}$$

THEOREM 4.1 (see [18]). *Under assumptions* (M_0), (M_1), *and* (M_2), *there exist constants* $\varepsilon_0 > 0$ *and* δ_0, *functions* $\alpha(\varepsilon)$, $T(\varepsilon)$, *and a* $T(\varepsilon)$*-periodic function* $x^*(\varepsilon)$ *such that*

(a) *all of these functions are of class* $\mathcal{C}^{k-1}$ *with respect to* ε, *for* $\varepsilon \in [0, \varepsilon_0[$, $\alpha(0) = \alpha_0$, $T(0) = 2\pi$, $x^*(0) = 0$;

(b) $x^*(\varepsilon)$ *is a* $T(\varepsilon)$*-periodic solution of* (4.1), *for the parameter values equal to* $\alpha(\varepsilon)$;

(c) *for* $|\alpha - \alpha_0| < \delta_0$ *and* $|T - 2\pi| < \delta_0$, *any* T*-periodic solution* p, *with* $\|p\| < \delta_0$, *of* (4.1) *for the parameter value* α, *there exists* $\varepsilon \in [0, \varepsilon_0[$ *such that* $\alpha = \alpha(\varepsilon)$, $T = T(\varepsilon)$, *and* p *is up to a phase shift equal to* $x^*(\varepsilon)$.

We notice that (M_1) implies that the root λ_0 lies on a branch of roots $\lambda = \lambda(\alpha)$ of (4.2) of class $\mathcal{C}^{k-1}$.

The following theorem gives the main result of this paper.

THEOREM 4.2. *Assume* (A_2) *and* $\alpha > \sup(\frac{\omega}{\beta}, \frac{\sigma}{\delta})$. *There exists* $\varepsilon_1 > 0$ *such that, for each* $0 \leq \varepsilon < \varepsilon_1$, (3.3) *has a family of periodic solutions* $p_l(\varepsilon)$, *with period* $T_l = T_l(\varepsilon)$, *for the parameter values* $\tau = \tau(\varepsilon)$ *such that* $p_l(0) = P_2$, $T_l(0) = \frac{2\pi}{\zeta_l}$, *and* $\tau(0) = \tau_l$, *where* τ_l *and* ζ_l *are given, respectively, as in* (3.6) *and* (3.7).

Proof. For the proof of Theorem 4.2, we apply the Hopf bifurcation theorem, Theorem 4.1.

From the expression of f in (3.3), we have

$$f(0,\tau)=0 \text{ for all } \tau>0$$

and

$$\frac{\partial f(0,\tau)}{\partial\varphi}=0 \text{ for all } \tau>0.$$

Then hypothesis (M_0) is satisfied. From (3.5) and Theorem 3.3, the characteristic equation (3.5) has a pair of simple imaginary roots $\lambda_l=i\zeta_l$ and $\overline{\lambda}_l=-i\zeta_l$ at $\tau=\tau_l$.

From (3.5), we have

$$W(\lambda_l,\tau_l)=0,$$

and, by differentiation, we find

$$\frac{\partial}{\partial\lambda}W(\lambda_l,\tau_l)=2i\zeta_l+p+(s-\tau(is\zeta_l+q))e^{-i\zeta_l\tau_l}\neq 0.$$

According to the implicit function theorem, there exist neighborhoods $\mathcal{U}\subset\mathbb{R}$ and $\mathcal{V}\subset\mathbb{C}$ of τ_l and λ_l, respectively, and a complex function $\lambda=\lambda(\tau)$ defined from $\mathcal{U}$ to $\mathcal{V}$, such that

$$\lambda(\tau_l)=\lambda_l,$$

and we have

$$W(\lambda(\tau),\tau)=0 \text{ for } \tau\in\mathcal{U}; \tag{4.5}$$

thus, we find

$$\lambda'(\tau)=-\frac{\partial W(\lambda,\tau)/\partial\tau}{\partial W(\lambda,\tau)/\partial\lambda} \text{ for } \tau\in\mathcal{U}. \tag{4.6}$$

From (3.5), we have

$$\begin{aligned}\frac{\partial}{\partial\tau}W(\lambda,\tau)&=-\lambda(s\lambda+q)e^{-\lambda\tau}\\&=\lambda(\lambda^2+p\lambda+r)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial\lambda}W(\lambda,\tau)&=2\lambda+p+(s-\tau s\lambda-\tau q)e^{-\lambda\tau}\\&=2\lambda+p+\tau(\lambda^2+p\lambda+r)-s\left(\frac{\lambda^2+p\lambda+r}{s\lambda+q}\right).\end{aligned}$$

From these expressions, a simple calculation leads to

$$\lambda'(\tau)=\frac{\lambda(s\lambda+q)e^{-\lambda\tau}}{2\lambda+p+(s-\tau s\lambda-\tau q)e^{-\lambda\tau}}. \tag{4.7}$$

From (3.5), (4.5), and (4.7), we obtain the following expression of $\lambda'(\tau)$ for $\tau\in\mathcal{U}$:

$$\lambda'(\tau)=-\lambda\frac{A(\lambda)}{B(\tau,\lambda)}, \tag{4.8}$$

where

$$A(\lambda) = s\lambda^3 + (sp+q)\lambda^2 + (sr+pq)\lambda + qr$$

and

$$B(\tau,\lambda) = \tau s\lambda^3 + (s+\tau(sp+q))\lambda^2 + (2q+\tau(sr+pq))\lambda + pq - sr + \tau qr.$$

Let $\lambda(\tau) = \kappa(\tau) + i\zeta(\tau)$ (where κ and ζ are the real and imaginary parts of λ such that $\kappa(\tau_l) = 0$ and $\zeta(\tau_l) = \zeta_l$, respectively).

From (4.8) and for $\tau = \tau_l$, we have

$$\lambda'(\tau)_{/\tau=\tau_l} = -i\zeta_l \frac{A(i\zeta_l)}{B(\tau_l, i\zeta_l)}. \tag{4.9}$$

By separating the real and imaginary parts in (4.9), we deduce that

$$\begin{aligned}\kappa'(\tau)_{/\tau=\tau_l} &= \frac{\partial}{\partial\tau}\operatorname{Re}\lambda(\tau)_{/\tau=\tau_l} \\ &= \zeta_l^2 \frac{s^2\zeta_l^4 + (sqr(\tau_l - 1) + 2q^2)\zeta_l^2 + sr^2(q - sr)}{A^2 + B^2} \\ &\quad + \frac{pq^2(p+r) - qr(2q + \tau_l(sr+pq))}{A^2 + B^2},\end{aligned}$$

where

$$A = -(s + \tau_l(sp+q))\zeta_l^2 + pq - sr + \tau qr$$

and

$$B = -\tau_l s\zeta_l^3 + (2q + \tau_l(sr+pq))\zeta_l.$$

As ζ_l (which exists for $0 < \beta < \beta_1$) is a root of (3.13), we conclude that

$$\kappa'(\tau)_{/\tau=\tau_l} \neq 0.$$

Then hypotheses (M_1) and (M_2) are satisfied, which completes the proof of Theorem 4.2. ∎

5. Application. Let us fix the parameters of the model as follows: $\alpha = 1.636$, $\beta = 0.002$, $\sigma = 0.1181$, $\delta = 0.3747$, and $0.00184 < \omega < 0.01185$. Moreover, assume $\omega = 0.04$ according to medical experiments (Kuznetsov and Taylor [20]). Therefore, $\alpha\beta = 0.003272 > \omega$ for $\omega = 0.04$. $\alpha\delta = 0.6160372 > \sigma$, $\beta < \beta_1 = 0.04654$.

The nontrivial steady state $P_2 = (x_2, y_2)$ is given by $x_2 = 1.6113$, $y_2 = 7.5352$, $p = 0.3994$, $r = 0.0092$, $s = -0.3014$, and $q = 0.4782$.

It is easy to verify that the conditions of Theorem 4.2 are fulfilled. A limit cycle bifurcation occurs when the time delay parameter τ passes through $\tau = \tau_l = 2.3581$, where the relative eigenvalues are $\lambda_l = i0.2010$. Moreover, we can compute the approximate period of the bifurcating periodic solution by

$$T_l = \frac{2\pi}{|\lambda_l|} = 9.2681 \text{ days}.$$

The bifurcating periodic solution of (2.4) is visualized with the aid of MATLAB software.

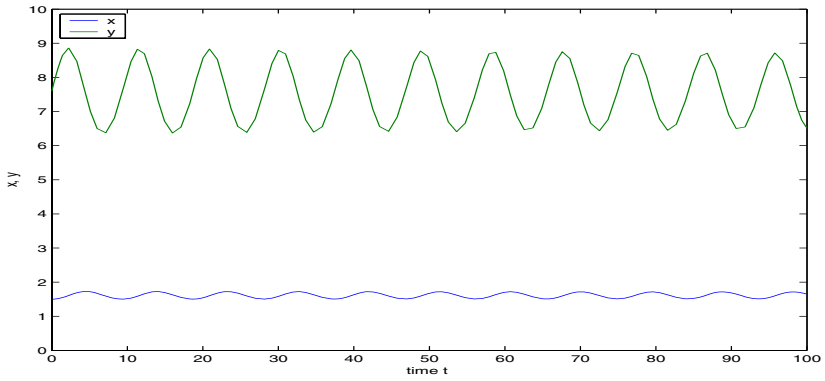


FIG. 2. *The bifurcating periodic solution of* (2.4) *in* (t, x) *and* (t, y) *planes for the delay parameter value* $\tau = \tau_l$.

We can deduce, according to Theorem 4.2, that the nontrivial stationary point P_2 is asymptotically stable when $\tau < \tau_l$. In this case the model allows for the possibility of a small tumor mass. When $\tau > \tau_l$, P_2 becomes unstable and a large tumor mass appears.

After extension of the time delay τ, the stable equilibrium is lost and the tumor starts to oscillate periodically (period equal to 9.2681 days). See Figure 2. Because of those oscillations the tumor can disappear (if the periodic oscillations are stable) or the patient can die (if the periodic oscillations are unstable).

6. Discussions. A numerical investigation, given in [16], shows that the characteristic equation (3.5) of the linearized system (2.4) around the nontrivial steady state P_2 has a purely imaginary root for some $\tau = \tau_0$, and the switching of stability may occur by using the Mikhailov hodograph.

In this paper, we gave an analytical study of stability (with respect to the time delay τ) of the possible steady states P_0 and P_2 for the positive values of the parameter ω.

In section 4 we prove that system (2.4) has a family of periodic solutions bifurcating from the nontrivial steady state.

We hope the results proposed in this paper improve the understanding of the qualitative properties of the description delivered by model (2.4). So far we now have a description of stability properties and Hopf bifurcation with a detailed analysis of the influence of delays terms. As demonstrated, whenever time delay was introduced into the tumor immune model, a Hopf point was found, leading to oscillatory behavior. It seems to us worthwhile to examine these results in the clinical context, that is, to check whether or not one can contain tumor growth by imposing (by use of certain drugs) time delay in tumor immune system competition.

Acknowledgments. I am very grateful to Professor N. Bellomo for improving this paper and I thank the two referees for their valuable suggestions.

REFERENCES

[1] N. BELLOMO AND L. PREZIOSI, *Modeling and mathematical problems related to tumor immune system interactions*, Math. Comput. Modelling, 31 (2000), pp. 413–452.

[2] N. BELLOMO AND G. FORNI, *Looking for new paradigms towards a biological-mathematical theory of complex multicellular systems*, Math. Models Methods Appl. Sci., 16 (2006), pp. 1001–1029.

[3] A. Bellouquid and M. Delitala, *Mathematical methods and tools of kinetic theory towards modelling complex biological systems*, Math. Models Methods Appl. Sci., 15 (2005), pp. 1639–1666.
[4] A. Bellouquid and M. Delitala, *Modelling Complex Biological Systems—A Kinetic Theory Approach*, Birkhäuser Boston, Boston, MA, 2006.
[5] M. Bodnar and U. Foryś, *Behaviour of solutions to Marchuk's model depending on a time delay*, Int. J. Appl. Math. Comput. Sci., 10 (2000), pp. 97–112.
[6] M. Bodnar and U. Foryś, *Periodic dynamics in the model of immune system*, Appl. Math. (Warsaw), 27 (2000), pp. 113–126.
[7] H. M. Byrne, *The effect of time delay on the dynamics of avascular tumor growth*, Math. Biosci., 144 (1997), pp. 83–117.
[8] R. Bürger, *The Mathematical Theory of Selection, Recombination, and Mutation*, Wiley, New York, 2000.
[9] N. Bellomo and P. Maini, *Preface to special issue on cancer modeling* II, Math. Models Methods Appl. Sci., 16, suppl. (1999), pp. iii–vii.
[10] K. L. Cooke and Z. Grossman, *Discrete delay, distributed delay, and stability switches*, J. Math. Anal. Appl., 86 (1982), pp. 592–627.
[11] L. Desvillettes and C. Prévots, *Modelling in Population Dynamics through Kinetic-Like Equations*, Preprint 99/19, Mathematics Department, The University of Orléans, Orléans, France, 1999.
[12] O. Dieckmann and J. P. Heesterbeek, *Mathematical Epidemiology of Infectious Diseases*, Wiley, New York, 2000.
[13] G. Forni, R. Fao, A. Santoni, and L. Frati, eds., *Cytokine Induced Tumor Immunogeneticity*, Academic Press, New York, 1994.
[14] U. Foryś, *Marchuk's model of immune system dynamics with application to tumor growth*, J. Theor. Med., 4 (2002), pp. 85–93.
[15] U. Foryś and M. Kolev, *Time delays in proliferation and apoptosis for solid avascular tumor*, in Mathematical Modelling of Population Dynamics, Banach Center Publ. 63, Polish Academy of Sciences, Warsaw, Poland, 2004, pp. 187–196.
[16] M. Galach, *Dynamics of the tumor-immune system competition—the effect of time delay*, Int. J. Appl. Math. Comput. Sci., 13 (2003), pp. 395–406.
[17] L. Greller, F. Tobin, and G. Poste, *Tumor heterogenity and progression: Conceptual foundation for modeling*, Invasion and Metastasis, 16 (1996), pp. 177–208.
[18] J. K. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.
[19] D. Kirschner and J. C. Panetta, *Modeling immunotherapy of the tumor-immune interaction*, J. Math. Biol., 37 (1998), pp. 235–252.
[20] V. A. Kuznetsov and M. A. Taylor, *Nonlinear dynamics of immunogenic tumors: Parameter estimation and global bifurcation analysis*, Bull. Math. Biol., 56 (1994), pp. 295–321.
[21] P. L. Lollini, S. Motta, and F. Pappalardo, *Modelling the immune competition*, Math. Models Methods Appl. Sci., 16, suppl. (2006), pp. 1091–1124.
[22] P. Michel, *Existence of solutions to the cell division eigenproblem*, Math. Models Methods Appl. Sci., 16, suppl. (2006), pp. 1125–1154.
[23] R. M. May and M. A. Nowak, *Virus Dynamics (Mathematical Principles of Immunology and Virology)*, Oxford University Press, Oxford, UK, 2000.
[24] H. Mayer, K. S. Zänker, and U. der Heiden, *A basic mathematical model of the immune response*, Chaos, 5 (1995), pp. 155–161.
[25] A. S. Perelson and G. Weisbuch, *Immunology for physicists*, Rev. Modern Phys., 69 (1997), pp. 1219–1268.
[26] L. Preziosi, ed., *Cancer Modelling and Simulation*, CRC Press/Chapman & Hall, Boca Raton, FL, 2003.
[27] J. Waniewski and P. Zhivkov, *A simple mathematical model for tumor–immune system interactions*, in Proceedings of the 8th National Conference on Application of Mathematics in Biology and Medicine, 2002, pp. 149–154.
[28] R. Yafia, *Hopf bifurcation analysis and numerical simulations in an ODE model of the immune system with positive immune response*, Nonlinear Anal. Real World Appl., to appear.
[29] R. Yafia, *Hopf bifurcation in a delayed model for tumor–immune system competition with negative immune response*, Discrete Dynamics in Nature and Society, Volume 2006, Article ID 95296, pp. 1–9.
[30] R. Yafia, *Stability of limit cycle in a delayed model for tumor immune system competition with negative immune response*, Discrete Dynamics in Nature and Society Volume 2006, Article ID 58463, pp. 1–13.

SIAM J. APPL. MATH.
Vol. 67, No. 6, pp. 1704–1717

A STOCHASTIC MODEL AND ASSOCIATED FOKKER–PLANCK EQUATION FOR THE FIBER LAY-DOWN PROCESS IN NONWOVEN PRODUCTION PROCESSES*

T. GÖTZ†, A. KLAR‡, N. MARHEINEKE†, AND R. WEGENER‡

Abstract. In this paper we present and investigate a stochastic model and its associated Fokker–Planck equation for the lay-down of fibers on a conveyor belt in the production process of nonwoven materials. The model is based on a stochastic differential equation taking into account the motion of the fiber under the influence of turbulence. A reformulation as a stochastic Hamiltonian system and an application of the stochastic averaging theorem lead to further simplifications of the model. Finally, the model is used to compute the distribution of functionals of the process that are important for the quality assessment of industrial fabrics.

Key words. fiber dynamics, stochastic Hamiltonian system, Fokker–Planck equations, stochastic averaging

AMS subject classifications. 37H10, 60H30, 70H05

DOI. 10.1137/06067715X

1. Introduction.

The understanding of the forms generated by the lay-down of flexible fibers onto a moving conveyor belt is of great interest in the production process of nonwovens that find their applications in, e.g., composite materials (filters), textiles, and the hygiene industry [3]. In the melt-spinning process of nonwoven materials, hundreds of individual endless fibers obtained by the continuous extrusion of a melted polymer are stretched and entangled by highly turbulent air flows to finally form a web on the conveyor belt. The quality of this web and the resulting nonwoven material—in terms of homogeneity and load capacity—depends essentially on the dynamics and the deposition of the fibers.

A mathematical model and numerical simulations for the nonwoven production process are presented in [12]. The paper focuses on the fiber spinning and lay-down, where the fiber dynamics in the deposition region close to the conveyor belt is dominated by the turbulent air flow. For the description of the interaction between fibers and turbulent flow, a stochastic force model is derived and analyzed in [18] as well as experimentally validated in [19]. Applying this concept, the fiber fabric can in principle be numerically generated and its quality investigated in the spirit of the multiscale image analysis of [22]. However, these or similar simulations usually lead to excessively large computation times, when all physical details of the production process are considered. Thus, simplified models for the lay-down process are needed. In particular, this is true for optimization and control procedures where many different simulations are needed. Experimental studies on the forms of threads laid on a moving belt as well as a simplified general theory for the buckling of the fibers can be found in [11]. The coiling behavior of flexible rods is investigated in [17].

*Received by the editors December 8, 2006; accepted for publication (in revised form) June 4, 2007; published electronically September 26, 2007. This work has been supported by the Kaiserslautern Excellence Cluster *Dependable Adaptive Systems and Mathematical Modeling*.
http://www.siam.org/journals/siap/67-6/67715.html
†Fachbereich Mathematik, Technische Universität Kaiserslautern, Kaiserslautern, Germany (goetz@mathematik.uni-kl.de, marheineke@mathematik.uni-kl.de).
‡Fraunhofer ITWM, Kaiserslautern, Germany (klar@itwm.fhg.de, wegener@itwm.fhg.de).

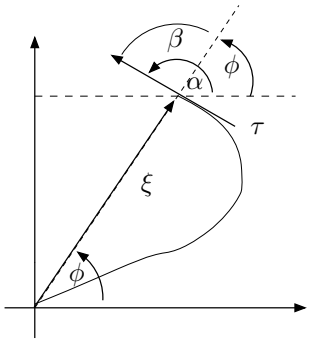


FIG. 1. *Fiber scenario on the conveyor belt.*

Motivated by the research work done in the field of woven textile composites [7], e.g., modeling [15], numerical and asymptotical stress analysis [23], and stiffness/load capacity investigations [6], we focus in this paper on the modeling of nonwoven textiles and the determination of textile properties, e.g., weight distribution, that are important for the quality assessment of industrial nonwoven fabrics. In particular, we present a new simplified stochastic model for the fiber lay-down process, i.e., for the generation of the fiber web on the conveyor belt that is assumed for simplicity to be nonmoving. Taking into account the fiber motion under the influence of turbulence, the process is described by a stochastic differential system in section 2. Its associated Kolmogoroff equation and stationary solution are investigated in section 3. Moreover, we include remarks on the identification of the process parameters. Section 4 contains an investigation of the scaled stochastic Hamiltonian system using stochastic averaging. In section 5 we conclude with the computation of the probability distribution of process functionals.

2. Stochastic model for fiber lay-down. In the industrial application a nonwoven material is generated by the superposition of many elastic, slender, and inextensible fibers. The nonwoven quality is measured by the homogeneity of the fiber mass distribution. Neglecting the interaction of neighboring fibers, we focus here on a single fiber in an isotropic lay-down process. Instead of describing this lay-down process in all physical details, we directly model the resulting fiber on the conveyor belt as an arc-length parameterized curve $\xi : \mathbb{R}_0^+ \to \mathbb{R}^2$ satisfying $\|\partial_t \xi\| = 1$; see Figure 1. Assuming constant cross sections, the curve carries the full information of the mass contribution of this fiber to the mass distribution of the nonwoven material. We prescribe ξ by a dynamical system with respect to the arc-length t that contains the crucial physical characteristics of the lay-down process, i.e., buckling/coiling of the fiber as well as the influence of the turbulent air flow in the deposition region on the fiber. Assuming for simplicity a nonmoving conveyor belt, the dynamical system is modeled by the following stochastic differential equations:

$$\begin{aligned} \mathrm{d}\xi_t &= \tau(\alpha_t)\,\mathrm{d}t, \\ \mathrm{d}\alpha_t &= -b(\|\xi\|)\frac{\xi_t}{\|\xi_t\|}\cdot\tau^{\perp}(\alpha_t)\,\mathrm{d}t + A\,\mathrm{d}W_t, \end{aligned}$$

where $\tau(\alpha) = (\cos\alpha, \sin\alpha)^T$ denotes the normalized tangent on the fiber. Since a curved fiber returns to the reference point of the spinning process determined by the nozzle position according to its coiling behavior, the change of the angle α is assumed

to be proportional to $\xi \cdot \tau^{\perp}(\alpha)$ with $\tau^{\perp}(\alpha) = (-\sin\alpha, \cos\alpha)^T$. The amplitude of this drive is prescribed by a continuously differentiable function $b : \mathbb{R}_0^+ \to \mathbb{R}$ with $b(r) > 0$ for $r > 0$ and $b'(r) \geq 0$ for $r \geq 0$. Let r_0 be the argument that satisfies $b(r_0) = 1/r_0$. The effect of the turbulent flow on the fiber in the deposition region close to the conveyor belt yields a stochastic force in the fiber lay-down process that is modeled here by a Wiener process W_t in $\mathbb{R}$ with amplitude A.

The parameters A and b depend on the specific industrial application and need to be adapted to experimental data. In case of an anisotropic lay-down process the scalar-valued function b has to be replaced by an appropriate matrix-valued one, i.e.,

$$\mathrm{d}\alpha_t = -\frac{\xi_t}{\|\xi_t\|} M(\|\xi_t\|)\tau^{\perp}(\alpha_t)\,\mathrm{d}t + A\,\mathrm{d}W_t,$$

where $M(r) \in \mathbb{R}^{2\times 2}$ denotes a positive definite matrix.

Since length, i.e., fiber (arc-)length t and position ξ, is the only dimension in the lay-down process, the system can be nondimensionalized by scaling it with the typical deposition radius r_0. Then, we have $b^\star(r^\star) = r_0\, b(r_0 r^\star)$ and $A^\star = \sqrt{r_0}A$. Consequently, $b^\star(1) = 1$ holds, and the dimensionless noise amplitude $A^\star$ characterizes the relation between stochastic and deterministic rates in the behavior of the system. Figure 2 shows examples for the pathwise behavior of the solution for varying noise A with constant drive $b(r) = 1$ and fixed fiber length. Physically relevant scenarios typically include parameters ranging from $A = 0.1$ to $A = 5$ depending on the size of the turbulent force exerted on the fiber. Note that for convenience we have skipped the superscript star * denoting the dimensionless quantities. Moreover, we embed our model in the context of dynamical systems and stochastic processes and use in the following the notation and interpretation of time for the arc-length parameter.

Introducing polar coordinates $\xi = (r\cos\phi, r\sin\phi)^T$, $r = \|\xi\|$ and the angle $\beta = \alpha - \phi$ with ϕ in ξ-space (Figure 1), the given stochastic differential system can be rewritten in terms of $(r, \beta) \in [0, \infty) \times [0, 2\pi]$:

$$\mathrm{d}r_t = \cos\beta_t\,\mathrm{d}t, \tag{2.1a}$$

$$\mathrm{d}\beta_t = \left(b(r_t) - \frac{1}{r_t}\right)\sin\beta_t\,\mathrm{d}t + A\,\mathrm{d}W_t, \tag{2.1b}$$

$$\mathrm{d}\phi_t = \frac{\sin\beta_t}{r_t}\,\mathrm{d}t. \tag{2.1c}$$

Due to symmetry, we can restrict ourselves to $\beta \in [0, \pi]$. Since the equation for ϕ_t is decoupled from the remaining (r_t, β_t)-process, this transformation leads to a dimension reduction of the problem. The deterministic (r, β)-system with $A = 0$ moves on closed orbits in the (r, β)-plane, as illustrated in Figure 3. Its fixpoint is $(r, \beta) = (1, \pi/2)$. The periodic orbits are given by level sets $H(r, \beta) = h \in [0, \infty)$ of the Hamiltonian

$$H(r, \beta) = B(r) - \ln r - \ln\sin\beta,$$

where $B(r) = \int_1^r b(r')\,\mathrm{d}r'$. For fixed energy h, the radius r takes values in $[r_{\min}(h), r_{\max}(h)]$ with $0 < r_{\min}(h) < 1 < r_{\max}(h)$.

We can rewrite (2.1) in Hamiltonian coordinates $(r, z) \in [0, \infty) \times (-\infty, \infty)$ with $z = \ln\tan(\beta/2)$, i.e., $\beta = 2\arctan(e^z)$, $z' = 1/\sin\beta$, $\beta' = \sin\beta$. Using Ito's formula

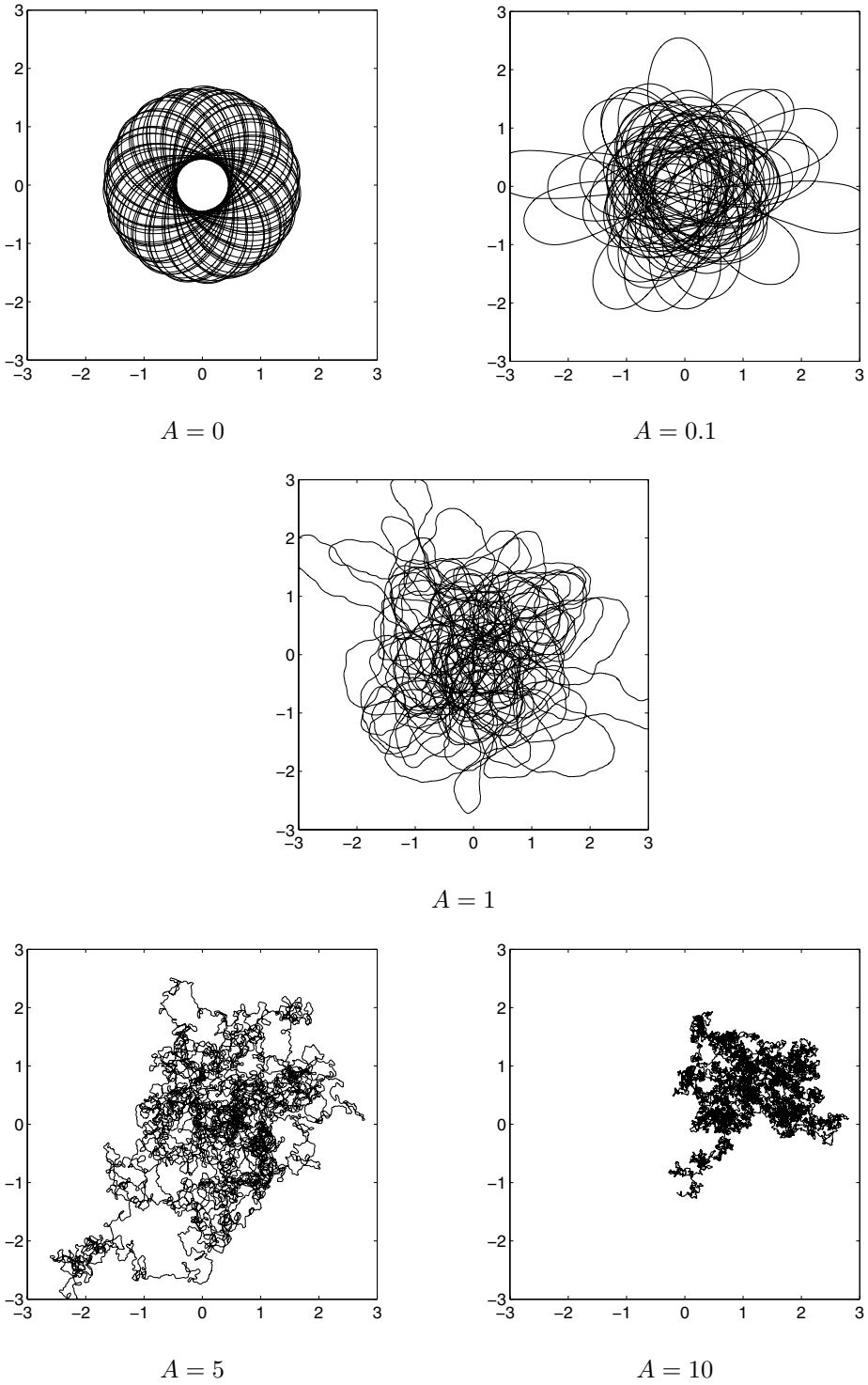


FIG. 2. *Representative path behavior for balanced ($A = 1$) as well as deterministic ($A < 1$) and stochastic ($A > 1$) dominated (ξ, α)-systems.*

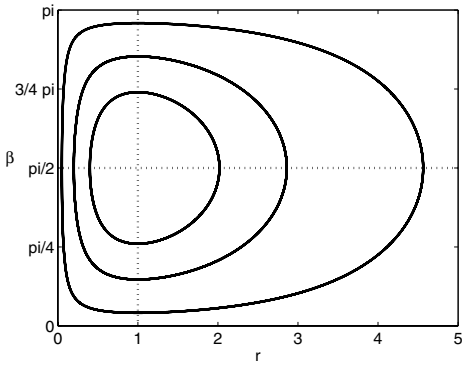


FIG. 3. *Orbits of the deterministic (r, β)-system for different values of h.*

then gives

$$\mathrm{d}r_t = \cos\beta(z_t)\,\mathrm{d}t, \tag{2.2a}$$

$$\mathrm{d}z_t = \left(b(r_t) - \frac{1}{r_t}\right)\mathrm{d}t - \frac{A^2\cos\beta(z_t)}{2\sin^2\beta(z_t)}\,\mathrm{d}t + \frac{A}{\sin\beta(z_t)}\,\mathrm{d}W_t \tag{2.2b}$$

and consequently, with $H(r,z) = B(r) - \ln(r) - \ln\sin(2\arctan(e^z))$,

$$\mathrm{d}r_t = -\partial_z H(r_t, z_t)\,\mathrm{d}t,$$

$$\mathrm{d}z_t = \partial_r H(r_t, z_t)\,\mathrm{d}t - \frac{A^2\cos\beta(z_t)}{2\sin^2\beta(z_t)}\,\mathrm{d}t + \frac{A}{\sin\beta(z_t)}\,\mathrm{d}W_t.$$

Remark 1. Linearizing the system in (2.2) around the fixpoint $(r,z) = (1,0)$, we obtain a Hamiltonian system with the Hamiltonian function

$$H_{lin}(r,z) = \frac{1}{2}\Big(b'(1)+1\Big)(r-1)^2 + \frac{1}{2}z^2$$

being a harmonic oscillator. Its period of motion [4] is $T_{lin} = 2\pi/\sqrt{b'(1)+1}$.

In general, as in the nonlinear case considered above, the period of motion T_H is not constant but depends on the energy h. An integral representation for T_H stated in (3.4) can be derived analytically; see below and [4] for further investigations. For small h, the nonlinear period of motion tends obviously to the linearized one.

Example 1. Considering the linearized system, the period of motion is $T_{lin} = 2\pi$ for $b(r) = 1$ and $T_{lin} = 2\pi/\sqrt{2}$ for $b(r) = r$. For the corresponding nonlinear cases, numerical evaluations of T_H are presented in Figure 4.

3. Kolmogoroff equation and stationary solution. We start the investigation of the fiber lay-down model by considering the associated Fokker–Planck equation and determining its stationary distribution as $t \to \infty$. We use the term "associated Fokker–Planck equation" for the Kolmogoroff forward equation of the stochastic process. The solution of this equation gives the probability density of the stochastic process for a given initial distribution; see, for example, [5].

The Fokker–Planck equation for the density $p_1 : (t, r, \beta) \mapsto p_1(t, r, \beta)$ of system (2.1) with $r \in [0, \infty)$, $\beta \in [0, 2\pi)$ is given by

$$\partial_t p_1 + \cos\beta\,\partial_r p_1 + \left(b(r) - \frac{1}{r}\right)\partial_\beta(\sin\beta\, p_1) = \frac{A^2}{2}\partial_{\beta\beta} p_1, \tag{3.1}$$

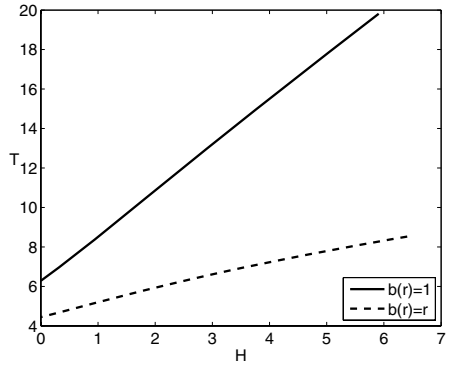


FIG. 4. *Period of motion $T_H(h)$ for $b(r) = 1$ (–) and $b(r) = r$ (- -).*

where the initial condition $p_1(t = 0, r, \beta)$ is prescribed. Moreover, periodic boundary conditions for β and the additional condition

$$\int_0^{2\pi} \int_0^{\infty} p_1(t, r, \beta)\, dr\, d\beta = 1$$

are prescribed. A stationary solution is obviously

$$p_{S_1}(r, \beta) = Cre^{-B(r)} \tag{3.2}$$

with the normalization constant C. In Hamiltonian coordinates (r, $z = \ln \tan(\beta/2)$) of system (2.2) we have

$$p_{S_2}(r, z) = p_{S_1}(r, \beta(z))\, \beta'(z) = Cre^{-B(r)} \sin \beta(z) = Ce^{-H(r,z)}$$

due to the transformation of measures. Note that the subscript of the stationary solution indicates the corresponding system.

Remark 2. In Hamiltonian coordinates the Fokker–Planck equation reads with the Hamiltonian function $H(r, z) = V(r) - W(z)$, where $V(r) = B(r) - \ln(r)$ and $W(z) = \ln \sin \beta(z)$:

$$\partial_t p_2 - \partial_r(p_2 \partial_z H) + \partial_z(p_2 \partial_r H) = \frac{A^2}{2} \partial_z(e^{-2W}(\partial_z p_2 + p_2 \partial_z H)).$$

Obviously, a solution is e^{-H}, as one would expect from physical considerations; see, e.g., Risken [21].

In the energy variable H the stationary distribution reads

$$\begin{aligned} p_{S_H}(h) &= \frac{\mathrm{d}}{\mathrm{d}h} \int_{H(r,z)<h} p_{S_2}(r, z)\, \mathrm{d}r\, \mathrm{d}z = C\, e^{-h} \frac{\mathrm{d}}{\mathrm{d}h} \int_{H(r,z)<h} \mathrm{d}r\, \mathrm{d}z \\ &= C\, e^{-h}\, T_H(h) \end{aligned} \tag{3.3}$$

because the period of motion of the deterministic system is given by

$$T_H(h) = \frac{\mathrm{d}}{\mathrm{d}h} \int_{H(r,z)<h} \mathrm{d}r\, \mathrm{d}z. \tag{3.4}$$

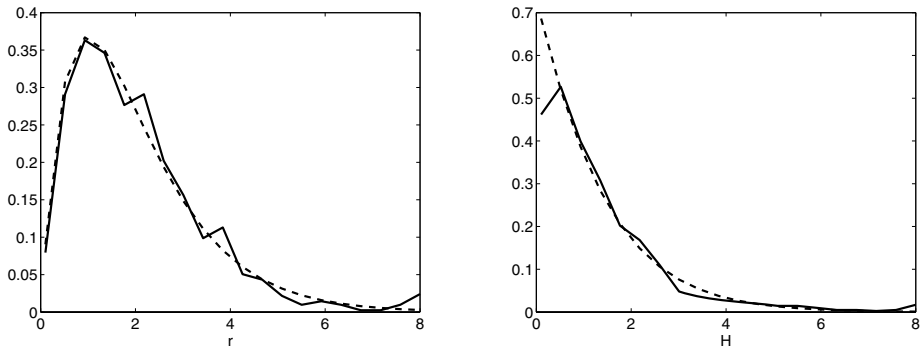


FIG. 5. *Stationary distributions, analytical (- -) versus numerical (–). Left:* p_{S_1}. *Right:* p_{S_H} *for* $b(r) = 1$.

This follows from the following consideration (cf. [4]). Take the reduced one-parametric family of deterministic processes with initial condition $H(r(0), \beta(0)) = h$; then every point (r, z) in the phase-space can be prescribed by h and the first time t to reach this point. As a simple consequence of the conservation of energy and the canonical structure of the Hamiltonian system, the functional determinant of the corresponding transformation is one. Hence, $\int_{H(r,z)<h} \mathrm{d}r\,\mathrm{d}z = \int_0^h \int_0^{T_H(h')} \mathrm{d}t\,\mathrm{d}h'$ which yields (3.4).

A Monte Carlo simulation of the stochastic model (2.1) with $b(r) = 1$ enables the computation of the time development of the solution and the comparison of the analytical stationary distributions p_{S_1} with the stationary distribution of the r_t-process and, respectively, the comparison of p_{S_H} with the numerically computed stationary distribution of the energy process H_t; see Figure 5. For further simulations of the time development of the stochastic model, see the next section.

Remark 3. Alternatively, to obtain the stationary solution and the time evolution of the density one can solve the partial differential equation (3.1) numerically; see [14]. The solutions indicate the convergence to the stationary solution with a rate of convergence depending on A, where the convergence is slow for A small and A large and fast for intermediate A. Compare also Figure 2 showing the paths of the stochastic process for different A.

Remark 4. An analytical investigation of the convergence towards equilibrium and the rate of convergence is complicated due to the degeneracy of the Fokker–Planck equation (3.1) with respect to the variable r. For related problems we refer to [24, 8]. A proof of ergodicity and convergence to the stationary distribution is performed in [10].

Remark 5. The identification of the parameters, i.e., drive b and noise amplitude A, in the lay-down model is important for the realistic description of industrially relevant scenarios. Comparing the stationary distribution p_{S_1} with experimentally available data, we can determine the function B and thus its derivative b. The noise amplitude A can in principle be computed from

$$(\mathrm{d}\alpha_t)^2 = A^2\,\mathrm{d}t$$

or alternatively from

$$\lim_{h\to 0} \frac{\mathbb{E}\left[(\alpha_{t+h} - \alpha_t)^2\right]}{h} = A^2,$$

supposing that the real process is prescribed by white noise. More sophisticated approaches can be found in [16].

4. Stochastic averaging and energy equation. In the following we consider lay-down processes (2.1) with small noise $A = \sqrt{\epsilon}\tilde{A}$ on associated long "time" scales $t = \tilde{t}/\epsilon$ with $0 < \epsilon \ll 1$; see Figure 2 for the pathwise behavior for different noise levels. In this case, a simplified approximation of the dynamics can be given by stochastic averaging. This leads to a reduced system as $\epsilon \to 0$, i.e., a stochastic differential equation for the limit energy process for which we determine the drift and variance coefficients.

Dropping the tildes, the rescaled $(r_t^\epsilon, \beta_t^\epsilon)$-system reads

$$\begin{aligned} \mathrm{d}r_t^\epsilon &= \frac{1}{\epsilon}\cos\beta_t^\epsilon \,\mathrm{d}t, \\ \mathrm{d}\beta_t^\epsilon &= \frac{1}{\epsilon}\left(b(r_t^\epsilon) - \frac{1}{r_t^\epsilon}\right)\sin\beta_t^\epsilon \,\mathrm{d}t + A\,\mathrm{d}W_t. \end{aligned} \tag{4.1}$$

Applying Ito's formula, the resulting energy process $H_t^\epsilon = H(r_t^\epsilon, \beta_t^\epsilon)$ fulfills the equation

$$\begin{aligned} \mathrm{d}H_t^\epsilon &= \partial_r H\,\mathrm{d}r_t^\epsilon + \partial_\beta H\,\mathrm{d}\beta_t^\epsilon + \frac{1}{2}\partial_{\beta\beta}H\,(\mathrm{d}\beta_t^\epsilon)^2 \\ &= \left(b(r_t^\epsilon) - \frac{1}{r_t^\epsilon}\right)\mathrm{d}r_t^\epsilon - \cot\beta_t^\epsilon\,\mathrm{d}\beta_t^\epsilon + \frac{1}{2\sin^2\beta_t^\epsilon}\,(\mathrm{d}\beta_t^\epsilon)^2 \\ &= \frac{A^2}{2\sin^2\beta_t^\epsilon}\,\mathrm{d}t - A\cot\beta_t^\epsilon\,\mathrm{d}W_t. \end{aligned}$$

Using formally the stochastic averaging theorem (see, e.g., [13] or [20] and [1, 2] for an application to stochastic Hamiltonian systems), the limit equation for H_t^ϵ as ϵ tends to 0 is given by

$$\mathrm{d}H_t^0 = a_H(H_t^0)\,\mathrm{d}t + \sigma_H(H_t^0)\,\mathrm{d}W_t$$

with drift and variance

$$a_H(h) = \frac{A^2}{2T_H(h)}\int_0^{T_H(h)} \frac{1}{\sin^2\beta(t)}\,\mathrm{d}t, \quad \sigma_H^2(h) = \frac{A^2}{T_H(h)}\int_0^{T_H(h)} \cot^2\beta(t)\,\mathrm{d}t. \tag{4.2}$$

In these formulas β denotes the solution of the deterministic (r, β)-process for fixed energy h. The expressions for a_H and σ_H^2 can be rewritten in explicit form as

$$a_H(h) = \frac{A^2}{2}\left(1 + \overline{T_H}(h)\right), \qquad \sigma_H^2(h) = A^2\,\overline{T_H}(h), \tag{4.3}$$

with

$$\overline{T_H}(h) = \frac{e^{2h}}{T_H(h)}\int_0^h T_H(h')e^{-2h'}\,\mathrm{d}h'.$$

The derivation of (4.3) is based on two equations for the coefficients a and σ. First, due to their form in (4.2) we have

$$2a_H(h) - \sigma_H^2(h) = A^2. \tag{4.4}$$

The second equation is obtained from the stationary distribution. Consider the Fokker–Planck equation corresponding to H_t^0:

$$\partial_t p_H + \partial_h \left(a_H(h) p_H\right) = \frac{1}{2} \partial_{hh} \left(\sigma_H^2(h) p_H\right).$$

Looking for integrable solutions of the stationary equation, we consider

$$2a_H(h)\, p_{S_H} = \partial_h \left(\sigma_H^2(h)\, p_{S_H}\right).$$

The solution reads

$$p_{S_H}(h) = \tilde{C} \exp\left(-\int^h \frac{(\sigma_H^2)'(h') - 2a_H(h')}{\sigma_H^2(h')}\, dh'\right)$$

with the normalizing constant $\tilde{C}$. On the other hand, the stationary solution for the $(r_t^\epsilon, \beta_t^\epsilon)$-process is independent of ϵ, and according to (3.3),

$$p_{S_H}(h) = Ce^{-h}\, T_H(h)$$

holds for all ϵ and therefore also for the limit process. Hence, from the comparison of the different expressions for p_{S_H}, we obtain

$$\frac{(\sigma_H^2)'(h) - 2a_H(h)}{\sigma_H^2(h)} = 1 - (\ln T_H(h))'. \tag{4.5}$$

Equations (4.4) and (4.5) yield a differential equation for the variance

$$(\sigma_H^2)'(h) = A^2 + 2\sigma_H^2(h) - (\ln T_H(h))'\, \sigma_H^2(h)$$

from which the explicit formulas in (4.3) can be concluded.

Summarizing, the energy equation for the limit process H_t^0 and its associated Fokker–Planck equation read

$$\begin{aligned} \mathrm{d}H_t^0 &= \frac{A^2}{2}(1 + \overline{T_H}(H_t^0))\, \mathrm{d}t + A\sqrt{\overline{T_H}(H_t^0)}\, \mathrm{d}W_t, \\ \partial_t p_H + \frac{A^2}{2} \partial_h \left((1 + \overline{T_H}(h))\, p_H\right) &= \frac{A^2}{2} \partial_{hh} \left(\overline{T_H}(h)\, p_H\right). \end{aligned}$$

The introduction of the alternative energy process

$$G_t^\epsilon = e^{-H_t^\epsilon} = r_t^\epsilon\, e^{-B(r_t^\epsilon)} \sin \beta_t^\epsilon$$

is more suitable for the following numerical simulations since it is restricted on the interval $[0, 1]$. Analogously to the previous averaging procedure or directly from H_t^ϵ by means of the Ito calculus, i.e.,

$$\mathrm{d}G_t^0 = -e^{-H_t^0}\, \mathrm{d}H_t^0 + \frac{1}{2} e^{-H_t^0}\, (\mathrm{d}H_t^0)^2,$$

we determine the limit process G_t^0 as

$$\mathrm{d}G_t^0 = a_G(G_t^0)\, \mathrm{d}t + \sigma_G(G_t^0)\, \mathrm{d}W_t \tag{4.6}$$

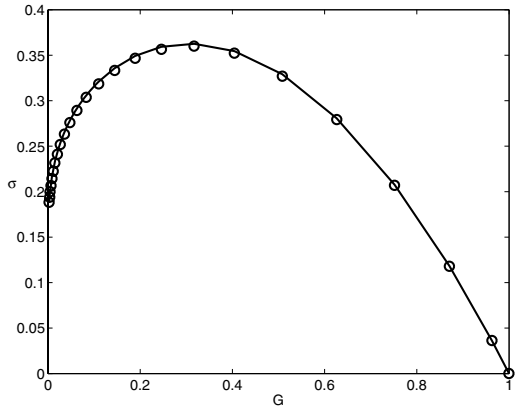


FIG. 6. σ_G *for* $A = 1$, $b(r) = 1$, *computed via* (4.8) *(–), via* (4.7) *(o).*

with drift and variance

$$a_G(g) = -\frac{A^2}{2} g, \qquad \sigma_G^2(g) = \frac{A^2 g^2}{T_G(g)} \int_0^{T_G(g)} \cot^2 \beta(t)\, \mathrm{d}t \tag{4.7}$$

or the explicit version,

$$\begin{aligned} \sigma_G^2(g) &= A^2 \overline{T_G}(g), \\ \overline{T_G}(g) &= \frac{1}{T_G(g)} \int_g^1 T_G(g')\, g'\, \mathrm{d}g'. \end{aligned} \tag{4.8}$$

Here, the period of motion T_G is defined by the associated transformation $T_G(g) = T_H(-\ln g)$. The associated Fokker–Planck equation is

$$\partial_t p_G - \frac{A^2}{2}\, \partial_g (g p_G) = \frac{A^2}{2}\, \partial_{gg} (\overline{T_G}(g) p_G). \tag{4.9}$$

The equation is complemented with an initial condition and conservation of mass

$$\int_0^1 p_G(t, g)\, \mathrm{d}g = 1.$$

The transformed stationary solution of the Fokker–Planck equation is

$$p_{S_G}(g) = C\, T_G(g), \quad g \in [0, 1].$$

Remark 6. Equation (4.9) can be rewritten in the usual form of a Sturm–Liouville problem. For a more detailed treatment of these problems, see, for example, [25]. At least formally it is easily observed that the rate of convergence to equilibrium for this equation is faster as A becomes larger, which is consistent with the statement on the speed of convergence of the full problem (3.1) in Remark 3.

Remark 7. For $T_H(h)$ tending to infinity as $h \to \infty$ (as motivated by Figure 4) and $T_G(g)\, g$ being integrable over $[0, 1]$, we observe from (4.8) that the variance satisfies $\sigma_G^2(g = 0) = 0$ with $\sigma_G^2(g) \sim (T_H(-\ln g))^{-1}$ as $g \to 0$. See Figure 6 for a numerical evaluation of σ_G for $b(r) = 1$ and $A = 1$.

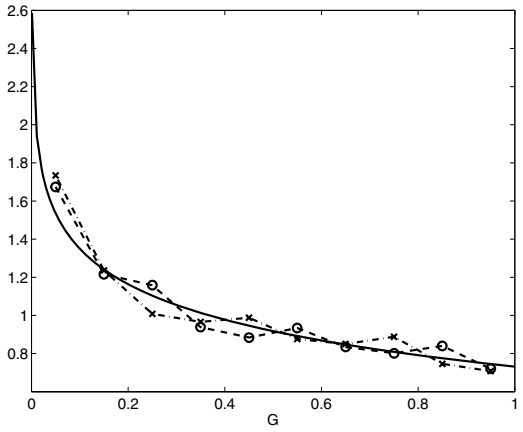


FIG. 7. *Stationary distribution p_{S_G} (–) versus numerical computed distributions for G_t^ϵ, $\epsilon = 1$ (x-.) and G_t^0 (o- -) for $b(r) = 1$.*

A numerical simulation of the corresponding stochastic differential equations (4.1) and (4.6) yields the stationary distributions of the energy process G_t^ϵ, $\epsilon = 1$, and the limit process G_t^0 that are plotted against the theoretical stationary distribution p_{S_G} in Figure 7. The temporal evolution of the mean value and the standard deviation of the two energy processes is visualized in Figure 8, where ϵ is chosen as $\epsilon = 1, 0.1$. The results for smaller ϵ are similar to the case $\epsilon = 0.1$. We observe that for large ϵ the decay in the beginning differs significantly, before the final behavior of the processes is driven by the standard noise of Monte Carlo simulations. For smaller values of ϵ the time evolutions of the full process and the limit process coincide.

5. Distributions of process functionals. An important issue for the quality assessment of fabrics is the distribution of fiber length that lies in a prescribed spatial domain, since this information yields the weight distribution in the physical space and thus gives insight into the structure of the nonwoven material, i.e., holes, thinning, swelling. In the context of stochastic processes the distribution of the length of the fibers laid down in a prescribed domain is associated with the distribution of the time the process stays in this domain. In the following we perform a numerical study of such functionals of the process.

The "time" spent in a domain D of the (r, β)–phase-space is described by the distribution of the random variable $I = \int_{t_0}^T \chi_D(r_t, \beta_t)\,\mathrm{d}t$ for fixed T, where χ_D is the characteristic function of D. Alternatively for domains given by the energy functional $G = e^{-H}$, e.g., $D = D_g = \{(r, \beta) \mid G(r, \beta) < g\}$, we can also use the approximate equations for G_t^0 to determine an approximation I_G for I with $I_G = \int_{t_0}^T \chi_{[0,g]}(G_t^0)\,\mathrm{d}t$.

Remark 8. The distribution of the above functionals can in principle be determined by solving a related partial differential equation to obtain the characteristic function of I. The distribution of I is then computed using the inverse Fourier transform; see [9]. However, for the nonlinear processes considered here there is no explicit solution of these equations, and a direct evaluation of the functionals by a Monte Carlo method is more straightforward.

Figure 9 shows the distribution of the functionals I/T and I_G/T with $D = D_g = \{(r, \beta) \mid G(r, \beta) < g\}$ for initial values $G_0 = 0.53$, $t_0 = 0$, $g = 0.3$ and final time $T = 40, 200, 400$ using the $(r_t^\epsilon, \beta_t^\epsilon)$-process with $\epsilon = 1$ and the G_t^0-process, respectively.

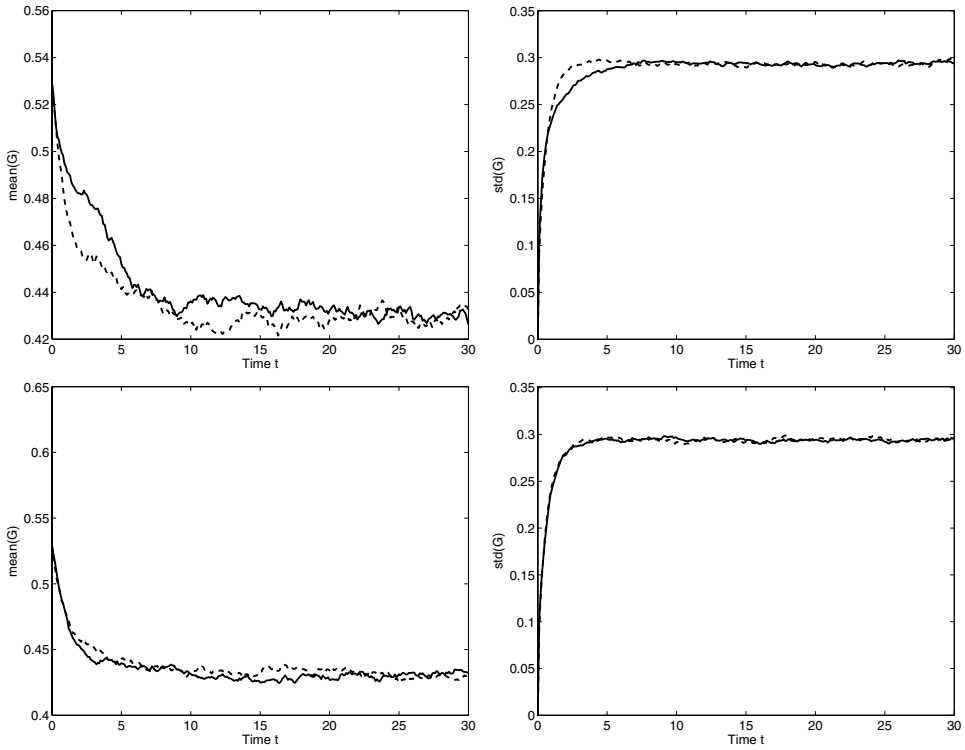


FIG. 8. *Time development of mean values (left) and standard deviations (right) of* G_t^ϵ *(solid) and* G_t^0 *(dashed) for* $\epsilon = 1, 0.1$ *(from top to bottom) for* $b(r) = 1$.

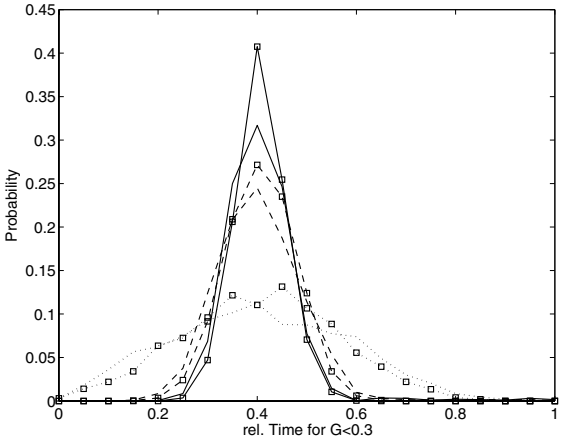


FIG. 9. *Probability distribution of time spent in* D_g, $g = 0.3$, $P(D_g) = 0.39$ *using the* $(r_t^\epsilon, \beta_t^\epsilon)$*-process,* $\epsilon = 1$ *(line), and* G_t^0*-process (line and marker) for* $T = 40$ *(··),* 200 *(- -),* 400 *(–).*

Obviously, in these cases evaluating the functionals with the limit G_t^0-process gives a good approximation of the true value even for $\epsilon = 1$.

Remark 9. For large times T, starting with the stationary distribution, the distribution of I_G tends towards a δ-distribution at the value $P(D_g) := \int_0^g p_{S_G}(g')\,\mathrm{d}g'$:

$$\lim_{T\to\infty} \frac{1}{T} \int_{t_0}^{t_0+T} \chi_{[0,g]}(G_t^0)\,\mathrm{d}t = P(D_g).$$

This means that, at least numerically, the distribution function behaves as predicted by the ergodic theorem.

6. Conclusion and outlook. In this paper we presented a new stochastic model for the lay-down of fibers which can be used as a basis for further investigations of the production process of nonwovens. We have determined the associated Kolmogoroff equation and stationary solution of the model and derived the corresponding energy process with drift and variance coefficients by help of stochastic averaging. The limit energy process enables the simple computation of probability distributions of process functionals, e.g., fiber length distribution, that are helpful for the quality assessment of nonwoven materials. Some points which need to be discussed in further detail are listed below:

- The application of the above results on the industrial process requires the extension of the model with regard to an anisotropic lay-down process and a moving conveyor belt. This generalization will lead to a matrix-valued drive function and an additional drift (transport) term in the (ξ, α)-system. For example, for a moving conveyor belt, we have to distinguish between the fiber curve η on the conveyor belt and the deviation ξ of the fiber from the now moving reference point of the spinning process

$$\eta_t = \xi_t - \kappa t e_1,$$

  where $\kappa \in [0, 1]$ is the speed of the belt moving in direction e_1 relative to the spinning velocity of the fiber. The inextensibility condition holds for η_t; i.e., $\|\partial_t \eta\| = 1$. Hence, in generalization to the presented model ξ_t fulfills

$$\begin{aligned} \mathrm{d}\xi_t &= \tau_t\,\mathrm{d}t + \kappa e_1\,\mathrm{d}t, \\ \mathrm{d}\alpha_t &= -b(\|\xi_t\|)\frac{\xi_t}{\|\xi_t\|}\tau_t^{\perp}\,\mathrm{d}t + A\,\mathrm{d}W_t. \end{aligned}$$

- The practical relevance of the model has to be guaranteed by the identification of the parameters, i.e., drive and noise amplitude. Therefore, appropriate validation data will be generated by the simulation of the complete physical production process as discussed above. Both parameters can be identified from the simulation of the complete production process of nonwovens, according to [12, 18], for different industrially relevant cases.
- The theoretical and numerical analysis of the solution of the degenerate Fokker–Planck equation (3.1) or the associated equation for the (ξ_t, α_t)-process and, in particular, the rates of convergence to equilibrium, gives valuable insight into the behavior of the process. A detailed analytical investigation with explicit rates of convergence for the process is presented in [10]. For a numerical investigation we refer to [14].
- The analytical investigation of the Fokker–Planck equation for the limit energy process G_t^0 is also a point of interest that is left to future work.

REFERENCES

[1] S. Albeverio and A. Klar, *Long time behavior of nonlinear stochastic oscillators: The one-dimensional Hamiltonian case*, J. Math. Phys., 35 (1994), pp. 4005–4027.

[2] S. Albeverio and A. Klar, *Longtime behaviour of stochastic Hamiltonian systems: The multidimensional case*, Potential Anal., 12 (2000), pp. 281–297.

[3] W. Albrecht, H. Fuchs, and W. Kittelmann, *Nonwoven Fabrics*, Wiley, New York, 2003.

[4] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer, New York, 1984.

[5] L. Arnold, *Stochastic Differential Equations*, Krieger, Malabar, FL, 1992.

[6] E. J. Barbero, T. M. Damiani, and J. Trovillion, *Micromechanics of fabric-reinforced composites with periodic microstructure*, Internat. J. Solids Structures, 42 (2005), pp. 2489–2504.

[7] B. N. Cox and G. Flanagan, *Handbook of Analytical Methods for Textile Composites*, Technical Report NASA-97-cr4750, NASA Langley Technical Report Server, 1997.

[8] L. Desvilettes and C. Villani, *On the trend to global equilibrium for spatially inhomogeneous entropy-dissipating systems: The linear Fokker-Planck equation*, Comm. Pure Appl. Math., 54 (2001), pp. 1–42.

[9] I. I. Gikhman and A. V. Skorokhod, *Introduction to the Theory of Random Processes*, W. B. Saunders, Philadelphia, 1969.

[10] M. Grothaus and A. Klar, *Ergodicity and Rate of Convergence for a Non-Sectorial Fiber Lay-Down Process*, preprint, TU Kaiserslautern, Kaiserslautern, Germany, 2007.

[11] J. W. Hearle, M. A. Sultan, and S. Govender, *The form taken by threads laid on a moving belt, Part* I–III, J. Textile Institute, 67 (1976), pp. 373–386.

[12] D. Hietel and N. Marheineke, *Mathematical modeling and numerical simulation of fiber dynamics*, Proc. Appl. Math. Mech., 5 (2005), pp. 667–670.

[13] R. Z. Khasminskii, *The behavior of a conservative system under the action of slight friction and slight random noise*, Prikl. Mat. Meh., 28 (1964), pp. 931–935.

[14] A. Klar, P. Reuterswärd, and M. Seaïd, *A Semi-Lagrangian Method for a Fokker-Planck Equation Describing Fiber Dynamics*, preprint, TU Kaiserslautern, Kaiserslautern, Germany, 2007.

[15] J. Koko, *Modeling of textile composites with warp/weft frictional contact*, J. Engrg. Math., 38 (2000), pp. 297–308.

[16] Y. Kutoyants, *Statistical Inference for Ergodic Diffusion Processes*, Springer, London, 2004.

[17] L. Mahadevan and J. B. Keller, *Coiling of flexible ropes*, Proc. Roy. Soc. London Ser. A, 452 (1996), pp. 1679–1694.

[18] N. Marheineke and R. Wegener, *Fiber dynamics in turbulent flows: General modeling framework*, SIAM J. Appl. Math., 66 (2006), pp. 1703–1726.

[19] N. Marheineke and R. Wegener, *Fiber dynamics in turbulent flows: Specific Taylor drag*, SIAM J. Appl. Math., accepted.

[20] G. Papanicolaou, D. Stroock, and S. Varadhan, *Martingale approach to some limit theorems*, in Statistical Mechanics and Dynamical Systems, Duke Univ. Math. Ser. 3, D. Ruelle, ed., Duke University, Durham, NC, 1978.

[21] H. Risken, *The Fokker-Planck Equation: Methods of Solution and Applications*, Springer, Berlin, 1996.

[22] M. Scholz and B. Claus, *Analysis and simulation of nonwoven textures*, ZAMM Z. Angew. Math. Mech., 79 (1999), pp. 237–240.

[23] S. Sihn, E. V. Iarve, and A. K. Roy, *Three-dimensional stress analysis of textile composites,* I *and* II, Internat. J. Solids Structures, 41 (2004), pp. 1377–1410.

[24] C. Villani, *Hypocoercivity*, preprint, École Normale Supérieure de Lyon, Lyon, France, 2006.

[25] J. Weidmann, *Linear Operators in Hilbert Spaces*, Grad. Texts in Math. 68, Springer, New York, Berlin, 1980.

SIAM J. APPL. MATH.
Vol. 67, No. 6, pp. 1718–1739

PULSE PROPAGATION AND TIME REVERSAL IN RANDOM WAVEGUIDES*

JOSSELIN GARNIER† AND GEORGE PAPANICOLAOU‡

Abstract. Mode coupling in a random waveguide can be analyzed with asymptotic analysis based on separation of scales when the propagation distance is large compared to the size of the random inhomogeneities, which have small variance, and when the wavelength is comparable to the scale of the inhomogeneities. In this paper we study the asymptotic form of the joint distribution of the mode amplitudes at different frequencies. We derive a deterministic system of transport equations that describe the evolution of mode powers. This result is applied to the computations of pulse spreading in a random waveguide. It is also applied to the analysis of time reversal in a random waveguide. We show that randomness enhances spatial refocusing and that diffraction-limited focal spots can be obtained even with small-size time-reversal mirrors. The refocused field is statistically stable for broadband pulses in general. We show here that it is also stable for narrowband pulses, provided that the time-reversal mirror is large enough.

Key words. acoustic waveguides, random media, asymptotic analysis

AMS subject classifications. 76B15, 35Q99, 60F05

DOI. 10.1137/060659235

1. Introduction.

In a perfect waveguide, energy propagates through its guided wave modes, which do not interact with each other. Imperfections in the material or in the geometrical properties of the waveguide induce mode coupling. These imperfections are usually small, but their effects accumulate over large propagation distances and can be significant. In this paper we consider wave propagation in an acoustic waveguide whose bulk modulus is a three-dimensional random function. Using the propagating modes of the unperturbed waveguide, we can reduce the three-dimensional wave propagation problem to the analysis of a system of coupled ordinary differential equations with random coefficients. It is in the frequency domain that the mode amplitudes satisfy these differential equations. We can analyze them as a system of random differential equations in a diffusion approximation, in which the mode amplitudes are a multidimensional diffusion process. This coupled mode asymptotic analysis was considered previously for applications in underwater acoustics [14, 5], fiber optics [17], and quantum mechanics [11].

The main purpose of this paper is to analyze the asymptotic behavior of the coupled mode amplitudes in the time domain. This requires the analysis of the joint distribution of the mode amplitudes at two nearby frequencies, which results in a system of transport equations for the mode powers. This is done in section 6. In section 7 we apply this result to the analysis of pulse spreading in a random waveguide.

In section 8 we consider time reversal in a random waveguide and present the first analysis of the phenomenon of side-lobe suppression, which has been observed in experiments [15] and in numerical simulations [2]. Side-lobe suppression is the main

*Received by the editors May 8, 2006; accepted for publication (in revised form) May 18, 2007; published electronically October 5, 2007.

http://www.siam.org/journals/siap/67-6/65923.html

†Laboratoire de Probabilités et Modèles Aléatoires & Laboratoire Jacques-Louis Lions, Université Paris VII, 2 Place Jussieu, 75251 Paris Cedex 5, France (garnier@math.jussieu.fr).

‡Mathematics Department, Stanford University, Stanford, CA 94305 (papanico@math.stanford.edu). The work of this author was supported by grants ONR N00014-02-1-0088 and NSF DMS-0354674-001.

application of our asymptotic analysis. Time-reversal refocusing has been studied extensively both experimentally and theoretically, in various contexts such as in ultrasound and underwater acoustics, as reviewed in [7, 8]. A time-reversal mirror is an active array of transducers that records a signal, time reverses it, and re-emits it into the medium. The waves generated at the time-reversal mirror propagate back to their source and focus near it, as if wave propagation was run in reverse time. Surprisingly, random inhomogeneities enhance the refocusing of the time-reversed waves near the original source location [4, 8]. Ultrasonic wave propagation and time reversal in homogeneous waveguides is studied experimentally and theoretically in [22]. Time-reversal refocusing was experimentally investigated in underwater acoustics in [15, 23], where the random inhomogeneities in the environment reduce the side-lobes that are seen in refocusing in a homogeneous waveguide.

In addition to enhanced refocusing and side-lobe suppression, statistical stability of the refocused field is critical for applications in communications and detection. Statistical stability means that the refocused field does not depend on the particular realization of the random medium. It has been studied in *broadband* regimes of propagation in one-dimensional random media [3, 10], in three-dimensional randomly layered media [9, 10], and in three-dimensional wave propagation in the paraxial approximation [2, 21]. Stabilization of the refocused field for broadband pulses results from the superposition of its many approximately uncorrelated frequency components. We show in section 8 that in random waveguides we have statistical stability of the refocused field even for *narrowband* pulses, provided that the number of propagating modes and the size of the time-reversal mirror are large enough.

2. Propagation in homogeneous waveguides. In this section we study wave propagation in an acoustic waveguide that supports a finite number of propagating modes. In an ideal waveguide the geometric structure and the medium parameters can have a general form in the transverse directions but must be homogeneous along the waveguide axis. There are two general types of ideal waveguides: those that surround a homogeneous region with a confining boundary, and those in which the confinement is achieved with a transversely varying index of refraction. We will present the analysis of the effects of random perturbations on waveguides of the first type and will illustrate specific results with a planar waveguide. The main difference in working with waveguides of the second type is that the transverse wave mode profiles depend on the frequency, but this does not affect the theory we present here.

2.1. Modeling of the waveguide. We consider linear acoustic waves propagating in three space dimensions modeled by the system of wave equations

$$\rho(\mathbf{r})\frac{\partial \mathbf{u}}{\partial t} + \nabla p = \mathbf{F}\,, \qquad \frac{1}{K(\mathbf{r})}\frac{\partial p}{\partial t} + \nabla \cdot \mathbf{u} = 0\,, \tag{2.1}$$

where p is the acoustic pressure, $\mathbf{u}$ is the acoustic velocity, ρ is the density of the medium, and K is the bulk modulus. The source is modeled by the forcing term $\mathbf{F}(t, \mathbf{r})$. We assume that the transverse profile of the waveguide is a simply connected region $\mathcal{D}$ in two dimensions. The direction of propagation along the waveguide axis is z and the transverse coordinates are denoted by $\mathbf{x} \in \mathcal{D}$. In the interior of the waveguide the medium parameters are homogeneous:

$$\rho(\mathbf{r}) \equiv \bar{\rho}, \qquad K(\mathbf{r}) = \bar{K} \qquad \text{for } \mathbf{x} \in \mathcal{D} \text{ and } z \in \mathbb{R}\,.$$

By differentiating the second equation of (2.1) with respect to time and substituting the first equation into it, we get the standard wave equation for the pressure field,

$$\Delta p - \frac{1}{\bar{c}^2}\frac{\partial^2 p}{\partial t^2} = \nabla \cdot \mathbf{F}\,, \tag{2.2}$$

where $\Delta = \Delta_\perp + \partial_z^2$ and $\Delta_\perp$ is the transverse Laplacian. The sound speed is $\bar{c} = \sqrt{\bar{K}/\bar{\rho}}$. We must now prescribe boundary conditions on the boundary $\partial\mathcal{D}$ of the domain $\mathcal{D}$. In underwater acoustics, or in seismic wave propagation, the density is much smaller outside than inside the waveguide. This means that we must use a pressure release boundary condition since the pressure is very weak outside, and therefore, by continuity, the pressure is zero just inside the waveguide. Motivated by such examples, we will use Dirichlet boundary conditions

$$p(t,\mathbf{x},z) = 0 \qquad \text{for } \mathbf{x} \in \partial\mathcal{D} \text{ and } z \in \mathbb{R}\,. \tag{2.3}$$

We could also consider other types of boundary conditions if, for example, the boundary of the waveguide is a rigid wall, in which case the normal velocity vanishes. By (2.1) we obtain Neumann boundary conditions for the pressure.

2.2. The propagating and evanescent modes. A waveguide mode is a monochromatic wave $p(t,\mathbf{x},z) = \hat{p}(\omega,\mathbf{x},z)e^{-i\omega t}$ with frequency ω, where $\hat{p}(\omega,\mathbf{x},z)$ satisfies the time harmonic form of the wave equation (2.2) without a source term:

$$\partial_z^2\hat{p}(\omega,\mathbf{x},z) + \Delta_\perp\hat{p}(\omega,\mathbf{x},z) + k^2(\omega)\hat{p}(\omega,\mathbf{x},z) = 0\,. \tag{2.4}$$

Here $k = \omega/\bar{c}$ is the wavenumber and we have Dirichlet boundary conditions on $\partial\mathcal{D}$. The transverse Laplacian in $\mathcal{D}$ with Dirichlet boundary conditions on $\partial\mathcal{D}$ is self-adjoint in $L^2(\mathcal{D})$. Its spectrum is an infinite number of discrete eigenvalues

$$-\Delta_\perp\phi_j(\mathbf{x}) = \lambda_j\phi_j(\mathbf{x}),\ \mathbf{x}\in\mathcal{D}, \qquad \phi_j(\mathbf{x}) = 0,\ \mathbf{x}\in\partial\mathcal{D}$$

for $j = 1, 2, \ldots$. The eigenvalues are positive and nondecreasing, and we assume for simplicity that they are simple, so we have $0 < \lambda_1 < \lambda_2 < \cdots$. The eigenmodes are real and form an orthonormal set

$$\int_{\mathcal{D}} \phi_j(\mathbf{x})\phi_l(\mathbf{x})d\mathbf{x} = \delta_{jl}\,,$$

with $\delta_{jl} = 1$ if $j = l$ and 0 otherwise. For a given frequency ω, there exists a unique integer $N(\omega)$ such that $\lambda_{N(\omega)} \le k^2(\omega) < \lambda_{N(\omega)+1}$, with the convention that $N(\omega) = 0$ if $\lambda_1 > k(\omega)$. The modal wavenumbers $\beta_j(\omega) \ge 0$ for $j \le N(\omega)$ are defined by

$$\beta_j^2(\omega) = k^2(\omega) - \lambda_j\,. \tag{2.5}$$

The solutions $\hat{p}_j(\omega,\mathbf{x},z) = \phi_j(\mathbf{x})e^{\pm i\beta_j(\omega)z}$, $j = 1,\ldots,N(\omega)$, of the wave equation (2.4) are the propagating waveguide modes. For $j > N(\omega)$ we define the modal wavenumbers $\beta_j(\omega) > 0$ by $\beta_j^2(\omega) = \lambda_j - k^2(\omega)$, and the corresponding solutions $\hat{q}_j(\omega,\mathbf{x},z) = \phi_j(\mathbf{x})e^{\pm\beta_j(\omega)z}$ of the wave equation (2.4) are the evanescent modes.

In this paper we shall illustrate some results for the planar waveguide. This is the case where $\mathcal{D}$ is $(0,d)\times\mathbb{R}$, and we consider only solutions that depend on $x \in (0,d)$. In this case we have $\lambda_j = \pi^2 j^2/d^2$, $\phi_j(x) = \sqrt{2/d}\sin(\pi jx/d)$, $j \ge 1$, and the number of propagating modes is $N(\omega) = [(\omega d)/(\pi\bar{c})]$, where $[x]$ is the integer part of x.

2.3. Excitation conditions for a source. We consider a point-like source located at $(\mathbf{x}_0, z = 0)$ that emits a signal with orientation in the z-direction:

$$\mathbf{F}(t, \mathbf{x}, z) = f(t)\delta(\mathbf{x} - \mathbf{x}_0)\delta(z)\mathbf{e}_z\,.$$

Here $\mathbf{e}_z$ is the unit vector pointing in the z-direction. By the first equation of (2.1), this source term implies that the pressure satisfies the following jump conditions across the plane $z = 0$:

$$\hat{p}(\omega, \mathbf{x}, z = 0^+) - \hat{p}(\omega, \mathbf{x}, z = 0^-) = \hat{f}(\omega)\delta(\mathbf{x} - \mathbf{x}_0)\,,$$

while the second equation of (2.1) implies that there is no jump in the longitudinal velocity so that the pressure field also satisfies $\partial_z\hat{p}(\omega, \mathbf{x}, z = 0^+) = \partial_z\hat{p}(\omega, \mathbf{x}, z = 0^-)$. Here $\hat{f}$ is the Fourier transform of f with respect to time:

$$\hat{f}(\omega) = \int f(t)e^{i\omega t}dt, \qquad f(t) = \frac{1}{2\pi}\int \hat{f}(\omega)e^{-i\omega t}d\omega\,.$$

The pressure field can be written as a superposition of the complete set of modes,

$$\begin{aligned}\hat{p}(\omega, \mathbf{x}, z) = &\left[\sum_{j=1}^{N} \frac{\hat{a}_j(\omega)}{\sqrt{\beta_j(\omega)}}e^{i\beta_j z}\phi_j(\mathbf{x}) + \sum_{j=N+1}^{\infty} \frac{\hat{c}_j(\omega)}{\sqrt{\beta_j(\omega)}}e^{-\beta_j z}\phi_j(\mathbf{x})\right]\mathbf{1}_{(0,\infty)}(z)\\ &+ \left[\sum_{j=1}^{N} \frac{\hat{b}_j(\omega)}{\sqrt{\beta_j(\omega)}}e^{-i\beta_j z}\phi_j(\mathbf{x}) + \sum_{j=N+1}^{\infty} \frac{\hat{d}_j(\omega)}{\sqrt{\beta_j(\omega)}}e^{\beta_j z}\phi_j(\mathbf{x})\right]\mathbf{1}_{(-\infty,0)}(z)\,,\end{aligned}$$

where $\hat{a}_j$ is the amplitude of the jth right-going mode propagating in the right half-space $z > 0$, $\hat{b}_j$ is the amplitude of the jth left-going mode propagating in the left half-space $z < 0$, and $\hat{c}_j$ (resp., $\hat{d}_j$) is the amplitude of the jth right-going (resp., left-going) evanescent mode. Substituting this expansion into the jump conditions, multiplying by $\phi_j(\mathbf{x})$, integrating with respect to $\mathbf{x}$ over $\mathcal{D}$, and using the orthogonality of the modes, we express the mode amplitudes in terms of the source:

$$\hat{a}_j(\omega) = -\hat{b}_j(\omega) = \frac{\sqrt{\beta_j(\omega)}}{2}\hat{f}(\omega)\phi_j(\mathbf{x}_0)\,, \qquad \hat{c}_j(\omega) = -\hat{d}_j(\omega) = -\frac{\sqrt{\beta_j(\omega)}}{2}\hat{f}(\omega)\phi_j(\mathbf{x}_0)\,.$$

3. Mode coupling in random waveguides. We consider a randomly perturbed waveguide section occupying the region $z \in [0, L/\varepsilon^2]$, with two homogeneous waveguides occupying the two half-spaces $z < 0$ and $z > L/\varepsilon^2$. The bulk modulus and the density have the form

$$\begin{aligned}\frac{1}{K(\mathbf{x}, z)} &= \begin{cases} \frac{1}{\bar{K}}\left(1 + \varepsilon\nu(\mathbf{x}, z)\right) & \text{for} \quad \mathbf{x} \in \mathcal{D}, \quad z \in [0, L/\varepsilon^2],\\ \frac{1}{\bar{K}} & \text{for} \quad \mathbf{x} \in \mathcal{D}, \quad z \in (-\infty, 0) \cup (L/\varepsilon^2, \infty),\end{cases}\\ \rho(\mathbf{x}, z) &= \bar{\rho} \quad \text{for} \quad \mathbf{x} \in \mathcal{D}, \quad z \in (-\infty, \infty)\,,\end{aligned}$$

where ν is a zero-mean, stationary, and ergodic random process with respect to the axis coordinate z. Moreover, it is assumed to possess enough decorrelation; more precisely, it fulfills the condition that "ν is ϕ-mixing, with $\phi \in L^{1/2}(\mathbb{R}^+)$" [16, section 4.6.2].

The perturbed wave equation satisfied by the pressure field is

$$\Delta p - \frac{1+\varepsilon\nu(\mathbf{x},z)}{\bar{c}^2}\frac{\partial^2 p}{\partial t^2} = \nabla\cdot\mathbf{F}\,, \tag{3.1}$$

where the average sound speed is $\bar{c} = \sqrt{\bar{K}/\bar{\rho}}$. The pressure field also satisfies the boundary conditions (2.3). We consider that a point-like source located at $(\mathbf{x}_0, 0)$ emits a pulse $f(t)$ and we denote by $\hat{a}_{j,0}(\omega)$ the initial mode amplitudes as described in the previous section. The weak fluctuations of the medium parameters induce a coupling between the propagating modes, as well as between propagating and evanescent modes, which build up and become of order one after a propagation distance of order ε^{-2}, as expected from the diffusion approximation theory.

3.1. Coupled amplitude equations. We fix the frequency ω and expand the field $\hat{p}$ inside the randomly perturbed waveguide in terms of the transverse eigenmodes,

$$\hat{p}(\omega,\mathbf{x},z) = \sum_{j=1}^{N(\omega)} \phi_j(\mathbf{x})\hat{p}_j(\omega,z) + \sum_{j=N(\omega)+1}^{\infty} \phi_j(\mathbf{x})\hat{q}_j(\omega,z)\,, \tag{3.2}$$

where $\hat{p}_j$ is the amplitude of the jth propagating mode and $\hat{q}_j$ is the amplitude of the jth evanescent mode. We introduce the right-going and left-going mode amplitudes $\hat{a}_j(\omega,z)$ and $\hat{b}_j(\omega,z)$, defined by

$$\hat{p}_j = \frac{1}{\sqrt{\beta_j}}\left(\hat{a}_j e^{i\beta_j z} + \hat{b}_j e^{-i\beta_j z}\right)\,, \qquad \frac{d\hat{p}_j}{dz} = i\sqrt{\beta_j}\left(\hat{a}_j e^{i\beta_j z} - \hat{b}_j e^{-i\beta_j z}\right)$$

for $j \le N(\omega)$. The total field $\hat{p}$ satisfies the time harmonic wave equation

$$\Delta\hat{p}(\omega,\mathbf{x},z) + k^2(\omega)(1+\varepsilon\nu(\mathbf{x},z))\hat{p}(\omega,\mathbf{x},z) = 0\,. \tag{3.3}$$

Using (3.2) in this equation, multiplying it by $\phi_l(\mathbf{x})$, and integrating over $\mathbf{x}\in\mathcal{D}$, we deduce from the orthogonality of the eigenmodes $(\phi_j)_{j\ge1}$ the following system of coupled differential equations for the mode amplitudes:

$$\frac{d\hat{a}_j}{dz} = \frac{i\varepsilon k^2}{2}\sum_{1\le l\le N}\frac{C_{jl}(z)}{\sqrt{\beta_j\beta_l}}\left(\hat{a}_l e^{i(\beta_l-\beta_j)z} + \hat{b}_l e^{-i(\beta_l+\beta_j)z}\right)\,, \tag{3.4}$$

$$\frac{d\hat{b}_j}{dz} = -\frac{i\varepsilon k^2}{2}\sum_{1\le l\le N}\frac{C_{jl}(z)}{\sqrt{\beta_j\beta_l}}\left(\hat{a}_l e^{i(\beta_l+\beta_j)z} + \hat{b}_l e^{i(\beta_j-\beta_l)z}\right)\,, \tag{3.5}$$

where

$$C_{jl}(z) = \int_{\mathcal{D}}\phi_j(\mathbf{x})\phi_l(\mathbf{x})\nu(\mathbf{x},z)d\mathbf{x}\,, \tag{3.6}$$

and we have neglected the evanescent modes. The system (3.4)–(3.5) is complemented with the boundary conditions

$$\hat{a}_j(\omega,0) = \hat{a}_{j,0}(\omega), \qquad \hat{b}_j\left(\omega,\frac{L}{\varepsilon^2}\right) = 0 \tag{3.7}$$

for the propagating modes. The second condition indicates that no wave is incoming from the right.

It is possible to carry out a complete asymptotic analysis, including the coupling with the evanescent modes, in a general context of random ordinary differential equations [20], and specifically for random waveguide problems [13, 12] and shallow water waves [18]. The evanescent modes do affect the propagation statistics in the asymptotic regime considered here. A detailed asymptotic analysis shows, however, that the coupling with evanescent modes does not remove energy from the propagating modes. This coupling changes only the frequency-dependent phases of the propagating mode amplitudes. The overall effect for these modes is an additional dispersive term, which is given in terms of the two-point statistics of the random process ν, in the same asymptotic regime as the one considered here. This effective dispersion affects the operator $\mathcal{L}$ for the statistics of the complex mode amplitudes in (4.1) but not the operator $\mathcal{L}_P$ for their square modulus in (4.10). Therefore, all results that involve the propagation of energy, which includes nearly all results presented here, are not affected by the evanescent modes. We indicate in the following where the results are affected and refer to [12] for their form when full evanescent coupling is included.

3.2. Propagator matrices. We introduce the rescaled propagating mode amplitudes $\hat{a}_j^\varepsilon$, $\hat{b}_j^\varepsilon$, $j = 1, \ldots, N(\omega)$, given by

$$\hat{a}_j^\varepsilon(\omega, z) = \hat{a}_j\left(\omega, \frac{z}{\varepsilon^2}\right), \qquad \hat{b}_j^\varepsilon(\omega, z) = \hat{b}_j\left(\omega, \frac{z}{\varepsilon^2}\right). \tag{3.8}$$

The two-point linear boundary value problem (3.4), (3.5), (3.7) for $(\hat{a}^\varepsilon, \hat{b}^\varepsilon)$ can be solved using propagator matrices. We first put the problem in vector-matrix form:

$$\frac{dX_\omega^\varepsilon}{dz} = \frac{1}{\varepsilon}\mathbf{H}_\omega\left(\frac{z}{\varepsilon^2}\right) X_\omega^\varepsilon. \tag{3.9}$$

Here the $2N(\omega)$-vector X_ω^ε, obtained by concatenating the $N(\omega)$-vectors $\hat{a}^\varepsilon$ and $\hat{b}^\varepsilon$ and the $2N(\omega) \times 2N(\omega)$ matrix $\mathbf{H}_\omega$, is defined by

$$X_\omega^\varepsilon(z) = \begin{bmatrix} \hat{a}^\varepsilon(\omega, z) \\ \hat{b}^\varepsilon(\omega, z) \end{bmatrix}, \qquad \mathbf{H}_\omega(z) = \begin{bmatrix} \mathbf{H}_\omega^{(a)}(z) & \mathbf{H}_\omega^{(b)}(z) \\ \overline{\mathbf{H}_\omega^{(b)}}(z) & \overline{\mathbf{H}_\omega^{(a)}}(z) \end{bmatrix}, \tag{3.10}$$

where the entries of the $N(\omega) \times N(\omega)$ matrices $\mathbf{H}_\omega^{(a)}(z)$ and $\mathbf{H}_\omega^{(b)}(z)$ are given by

$$H_{\omega,jl}^{(a)}(z) = \frac{ik^2}{2}\frac{C_{jl}(z)}{\sqrt{\beta_j\beta_l}} e^{i(\beta_l - \beta_j)z}, \qquad H_{\omega,jl}^{(b)}(z) = \frac{ik^2}{2}\frac{C_{jl}(z)}{\sqrt{\beta_j\beta_l}} e^{-i(\beta_l + \beta_j)z}. \tag{3.11}$$

The propagator matrices $\mathbf{P}_\omega^\varepsilon(z)$ are the $2N(\omega) \times 2N(\omega)$ random matrices solution of the initial value problem

$$\frac{d\mathbf{P}_\omega^\varepsilon}{dz} = \frac{1}{\varepsilon}\mathbf{H}_\omega\left(\frac{z}{\varepsilon^2}\right)\mathbf{P}_\omega^\varepsilon, \tag{3.12}$$

with the initial condition $\mathbf{P}_\omega^\varepsilon(z = 0) = \mathbf{I}$. The solution of (3.4), (3.5), (3.7) satisfies

$$\begin{bmatrix} \hat{a}^\varepsilon(\omega, L) \\ 0 \end{bmatrix} = \mathbf{P}_\omega^\varepsilon(L) \begin{bmatrix} \hat{a}_0(\omega) \\ \hat{b}^\varepsilon(\omega, 0) \end{bmatrix}, \tag{3.13}$$

so that $\hat{a}^\varepsilon(\omega, L)$ can be expressed in terms of the entries of the propagator matrix $\mathbf{P}_\omega^\varepsilon(L)$. The symmetry relation (3.10) satisfied by the matrix $\mathbf{H}_\omega$ imposes the condition that the propagator has the form

$$\mathbf{P}_\omega^\varepsilon(z) = \begin{bmatrix} \mathbf{P}_\omega^{\varepsilon,a}(z) & \mathbf{P}_\omega^{\varepsilon,b}(z) \\ \overline{\mathbf{P}_\omega^{\varepsilon,b}}(z) & \overline{\mathbf{P}_\omega^{\varepsilon,a}}(z) \end{bmatrix}, \tag{3.14}$$

where $\mathbf{P}_\omega^{\varepsilon,a}(z)$ and $\mathbf{P}_\omega^{\varepsilon,b}(z)$ are $N(\omega)\times N(\omega)$ matrices. Note that the matrix $\mathbf{P}_\omega^{\varepsilon,a}$ describes the coupling between different right-going modes, while $\mathbf{P}_\omega^{\varepsilon,b}$ describes the coupling between right-going and left-going modes.

3.3. The forward scattering approximation. The limit as $\varepsilon\to 0$ of $\mathbf{P}_\omega^\varepsilon$ can be obtained and identified as a multidimensional diffusion process, meaning that the entries of the limit matrix satisfy a system of linear stochastic differential equations. This follows from the application of the diffusion-approximation theorem proved in [19]. The stochastic differential equations for the limit entries of $\mathbf{P}_\omega^{\varepsilon,b}(z)$ are coupled to the limit entries of $\mathbf{P}_\omega^{\varepsilon,a}(z)$ through the coefficients

$$\int_0^\infty \mathbb{E}[C_{jl}(0)C_{jl}(z)]\cos((\beta_j(\omega)+\beta_l(\omega))z)\,dz,\qquad j,l=1,\dots,N(\omega)\,.$$

This is because the phase factors present in the matrix $\mathbf{H}_\omega^{(b)}(z)$ are $\pm(\beta_j+\beta_l)z$. On the other hand, the stochastic differential equations for the limit entries of $\mathbf{P}_\omega^{\varepsilon,a}(z)$ are coupled to each other through the coefficients

$$\int_0^\infty \mathbb{E}[C_{jl}(0)C_{jl}(z)]\cos((\beta_j(\omega)-\beta_l(\omega))z)\,dz,\qquad j,l=1,\dots,N(\omega)\,.$$

This is because the phase factors present in the matrix $\mathbf{H}_\omega^{(a)}(z)$ are $\pm(\beta_j-\beta_l)z$. If we assume that the power spectral density of the process ν (i.e., the Fourier transform of its z-autocorrelation function) possesses a cut-off frequency, then it is natural to consider the case where

$$\int_0^\infty \mathbb{E}[C_{jl}(0)C_{jl}(z)]\cos((\beta_j(\omega)+\beta_l(\omega))z)\,dz=0,\qquad j,l=1,\dots,N(\omega),\tag{3.15}$$

while (at least) some of the intracoupling coefficients (those with $|j-l|=1$) are not zero. As a result of this assumption, the asymptotic coupling between $\mathbf{P}_\omega^{\varepsilon,a}(z)$ and $\mathbf{P}_\omega^{\varepsilon,b}(z)$ becomes zero. If we also take into account the initial condition $\mathbf{P}_\omega^{\varepsilon,b}(z=0)=\mathbf{0}$, then the limit of $\mathbf{P}_\omega^{\varepsilon,b}(z)$ is $\mathbf{0}$.

In the forward scattering approximation, we neglect the left-going (backward) propagating modes. As we have just seen, it is valid in the limit $\varepsilon\to 0$ when the condition (3.15) holds. In this case we can consider the simplified coupled mode equation given by

$$\frac{d\hat a^\varepsilon}{dz}=\frac{1}{\varepsilon}\mathbf{H}_\omega^{(a)}\left(\frac{z}{\varepsilon^2}\right)\hat a^\varepsilon\,,\tag{3.16}$$

where $\mathbf{H}_\omega^{(a)}$ is the $N(\omega)\times N(\omega)$ complex matrix given by (3.11). The system (3.16) is provided with the initial condition $\hat a_j^\varepsilon(\omega,z=0)=\hat a_{j,0}(\omega)$. Note that the matrix $\mathbf{H}_\omega^{(a)}$ is skew Hermitian, which implies the conservation relation $\sum_{j=1}^N|\hat a_j^\varepsilon(L)|^2=\sum_{j=1}^N|\hat a_{j,0}|^2$. We finally introduce the *transfer*, or propagator matrix $\mathbf{T}^\varepsilon(\omega,z)$, which is the fundamental solution of (3.16). It is the $N(\omega)\times N(\omega)$ matrix solution of

$$\frac{d}{dz}\mathbf{T}^\varepsilon(\omega,z)=\frac{1}{\varepsilon}\mathbf{H}_\omega^{(a)}\left(\frac{z}{\varepsilon^2}\right)\mathbf{T}^\varepsilon(\omega,z)\,,\tag{3.17}$$

starting from $\mathbf{T}^\varepsilon(\omega,0)=\mathbf{I}$. The (j,l)-entry of the transfer matrix is the transmission coefficient $T_{jl}^\varepsilon(\omega,L)$, i.e., the output amplitude of the mode j when the input wave is a pure l mode with amplitude one. The transfer matrix $\mathbf{T}^\varepsilon(\omega,L)$ is unitary because $\mathbf{H}_\omega^{(a)}$ is skew Hermitian.

4. The time harmonic problem. In this section we consider the system of random differential equations (3.16) for a single frequency ω. Most of the results presented in this section can be found in [14] and are known collectively as the "coupled mode theory." We reproduce this theory because the original two-frequency analysis presented in the next section will give a new point of view on it.

4.1. The coupled mode diffusion process. We now apply the diffusion approximation theorem [19] to the system (3.16). The limit distribution of $\hat{a}^\varepsilon$ as $\varepsilon \to 0$ is a diffusion on $\mathbb{C}^{N(\omega)}$. We will assume that the longitudinal wavenumbers β_j, along with their sums and differences, are distinct. In this case the infinitesimal generator of the limit $\hat{a}$ has a simple form, provided we write it in terms of $\hat{a}$ and $\bar{\hat{a}}$ rather than in terms of the real and imaginary parts of $\hat{a}$. We get the following result.

PROPOSITION 4.1. *The mode amplitudes $(\hat{a}_j^\varepsilon(\omega,z))_{j=1,\dots,N}$ converge in distribution as $\varepsilon \to 0$ to the diffusion process $(\hat{a}_j(\omega,z))_{j=1,\dots,N}$, whose infinitesimal generator is*

$$\mathcal{L} = \frac{1}{4}\sum_{j\neq l}\Gamma_{jl}^{(c)}(\omega)\left(A_{jl}\overline{A_{jl}} + \overline{A_{jl}}A_{jl}\right) + \frac{1}{2}\sum_{j,l}\Gamma_{jl}^{(1)}(\omega)A_{jj}\overline{A_{ll}}$$

$$+\frac{i}{4}\sum_{j\neq l}\Gamma_{jl}^{(s)}(\omega)(A_{ll}-A_{jj})\,, \tag{4.1}$$

$$A_{jl} = \hat{a}_j\frac{\partial}{\partial \hat{a}_l} - \bar{\hat{a}}_l\frac{\partial}{\partial \bar{\hat{a}}_j} = -\overline{A_{lj}}\,. \tag{4.2}$$

Here we have defined the complex derivatives in the standard way: if $z = x + iy$, then $\partial_z = (1/2)(\partial_x - i\partial_y)$ and $\partial_{\bar{z}} = (1/2)(\partial_x + i\partial_y)$. The coefficients $\Gamma^{(c)}$, $\Gamma^{(s)}$, and $\Gamma^{(1)}$ are given by

$$\Gamma_{jl}^{(c)}(\omega) = \frac{\omega^4\gamma_{jl}^{(c)}(\omega)}{4\bar{c}^4\beta_j(\omega)\beta_l(\omega)} \quad \text{if } j\neq l\,, \qquad \Gamma_{jj}^{(c)}(\omega) = -\sum_{n\neq j}\Gamma_{jn}^{(c)}(\omega)\,, \tag{4.3}$$

$$\gamma_{jl}^{(c)}(\omega) = 2\int_0^\infty \cos\left((\beta_j(\omega)-\beta_l(\omega))z\right)\mathbb{E}[C_{jl}(0)C_{jl}(z)]dz\,, \tag{4.4}$$

$$\Gamma_{jl}^{(s)}(\omega) = \frac{\omega^4\gamma_{jl}^{(s)}(\omega)}{4\bar{c}^4\beta_j(\omega)\beta_l(\omega)} \quad \text{if } j\neq l\,, \qquad \Gamma_{jj}^{(s)}(\omega) = -\sum_{n\neq j}\Gamma_{jn}^{(s)}(\omega)\,, \tag{4.5}$$

$$\gamma_{jl}^{(s)}(\omega) = 2\int_0^\infty \sin\left((\beta_j(\omega)-\beta_l(\omega))z\right)\mathbb{E}[C_{jl}(0)C_{jl}(z)]dz\,, \tag{4.6}$$

$$\Gamma_{jl}^{(1)}(\omega) = \frac{\omega^4\gamma_{jl}^{(1)}}{4\bar{c}^4\beta_j(\omega)\beta_l(\omega)} \quad \text{for all } j,l\,, \tag{4.7}$$

$$\gamma_{jl}^{(1)} = 2\int_0^\infty \mathbb{E}[C_{jj}(0)C_{ll}(z)]dz\,. \tag{4.8}$$

Let us discuss some qualitative properties of the diffusion process $\hat{a}$.

(1) The coefficients of the second derivatives of the generator $\mathcal{L}$ are homogeneous of degree two, while the coefficients of the first derivatives are homogeneous of degree one. As a consequence we can write closed differential equations for moments of any order, as we shall see in the next sections.

(2) The coefficients $\gamma_{jl}^{(c)}$, and thus $\Gamma_{jl}^{(c)}$, are proportional to the power spectral densities of the stationary process $C_{jl}(z)$ for $j \neq l$. They are, therefore, nonnegative. In this paper we assume that the off-diagonal entries of the matrix $\Gamma^{(c)}$ are positive.

(3) We have $A_{jn}\big(\sum_{l=1}^N |\hat{a}_l|^2\big) = 0$ for any j, n, so that the infinitesimal generator satisfies $\mathcal{L}\big(\sum_{l=1}^N |\hat{a}_l|^2\big) = 0$. This implies that the diffusion process is supported on a sphere of $\mathbb{C}^N$, whose radius R_0 is determined by the initial condition $R_0^2 = \sum_{l=1}^N |\hat{a}_{l,0}(\omega)|^2$. The operator $\mathcal{L}$ is not self-adjoint on the sphere because of the term $\Gamma^{(s)}$ in (4.1). This means that the process is not reversible. However, the uniform measure on the sphere is invariant, and the generator is strongly elliptic. From the theory of irreducible Markov processes with compact state space, we know that the process is ergodic, which means in particular that for large z, the limit process $\hat{a}(z)$ converges to the uniform distribution over the sphere of radius R_0. This fact can be used to compute the limit distribution of the mode powers $(|\hat{a}_j|^2)_{j=1,\dots,N}$ for large z, which is the uniform distribution over $\mathcal{H}_N$,

$$\mathcal{H}_N = \left\{ (P_j)_{j=1,\dots,N},\, P_j \geq 0,\, \sum_{j=1}^N P_j = R_0^2 \right\}. \tag{4.9}$$

In the next section we carry out a more detailed analysis that is valid for any z.

(4) As noted at the end of section 3.1, coupling to the evanescent modes does affect $\mathcal{L}$ in (4.1). All the coefficients (4.3)–(4.8) remain the same except for $\Gamma_{jl}^{(s)}(\omega)$ in (4.5) which has an additional term [12]. This modification affects (6.1) and (6.7) in the following sections.

4.2. Coupled power equations. The generator of the limit process $\hat{a}$ possesses an important symmetry, which follows from noting that, when applying the generator to a function of $(|\hat{a}_1|^2, \dots, |\hat{a}_N|^2)$, we obtain another function of $(|\hat{a}_1|^2, \dots, |\hat{a}_N|^2)$. This implies that the limit process $(|\hat{a}_1(z)|^2)_{j=1,\dots,N}$ is itself a Markov process.

PROPOSITION 4.2. *The mode powers $(|\hat{a}_j^\varepsilon(\omega, z)|^2)_{j=1,\dots,N}$ converge in distribution as $\varepsilon \to$ to the diffusion process $(P_j(\omega, z))_{j=1,\dots,N}$, whose infinitesimal generator is*

$$\mathcal{L}_P = \sum_{j \neq l} \Gamma_{jl}^{(c)}(\omega) \left[P_l P_j \left(\frac{\partial}{\partial P_j} - \frac{\partial}{\partial P_l} \right) \frac{\partial}{\partial P_j} + (P_l - P_j) \frac{\partial}{\partial P_j} \right]. \tag{4.10}$$

As pointed out above, the diffusion process $(P_j(\omega, z))_{j=1,\dots,N}$ is supported in $\mathcal{H}_N$. As a first application of this result, we compute the mean mode powers:

$$P_j^{(1)}(\omega, z) = \mathbb{E}[P_j(\omega, z)] = \lim_{\varepsilon \to 0} \mathbb{E}[|\hat{a}_j^\varepsilon(\omega, z)|^2].$$

Using the generator $\mathcal{L}_P$ we get the following proposition.

PROPOSITION 4.3. *The mean mode powers $\mathbb{E}[|\hat{a}_j^\varepsilon(\omega, z)|^2]$ converge to $P_j^{(1)}(\omega, z)$, which is the solution of the linear system*

$$\frac{dP_j^{(1)}}{dz} = \sum_{n \neq j} \Gamma_{jn}^{(c)}(\omega) \left(P_n^{(1)} - P_j^{(1)} \right), \tag{4.11}$$

starting from $P_j^{(1)}(\omega, z = 0) = |\hat{a}_{j,0}(\omega)|^2$, $j = 1, \dots, N$.

The solution of this system can be written in terms of the exponential of the matrix $\Gamma^{(c)}(\omega)$. We note that the vector $P^{(1)}(\omega, z)$ has a probabilistic interpretation, which we consider in some detail in section 6.3. We give here some basic properties. First, the matrix $\Gamma^{(c)}(\omega)$ is symmetric and real, its off-diagonal terms are positive, and its diagonal terms are negative. The sums over the rows and columns are all

zero. As a consequence of the Perron–Frobenius theorem, $\Gamma^{(c)}(\omega)$ has zero as a simple eigenvalue, and all other eigenvalues are negative. The eigenvector associated with the zero eigenvalue is the uniform vector $(1,\dots,1)^T$. This shows that

$$\sup_{j=1,\dots,N}\left|P_j^{(1)}(\omega,z)-\frac{1}{N}R_0^2\right|\le Ce^{-z/L_{\mathrm{equip}}(\omega)},$$

where $R_0^2=\sum_{j=1}^N|\hat a_{j,0}(\omega)|^2$ and $L_{\mathrm{equip}}(\omega)$ is the absolute value of the reciprocal of the second eigenvalue of $\Gamma^{(c)}(\omega)$. In other words, the mean mode powers converge exponentially fast to the uniform distribution, which means that we have asymptotic *equipartition* of mode energy. The length $L_{\mathrm{equip}}(\omega)$ is the equipartition distance for the mean mode powers.

4.3. Fluctuations theory. Proposition 4.2 also allows us to study the fluctuations of the mode powers by looking at the fourth-order moments of the mode amplitudes:

$$P_{jl}^{(2)}(\omega,z)=\lim_{\varepsilon\to0}\mathbb{E}\left[|\hat a_j^\varepsilon(\omega,z)|^2|\hat a_l^\varepsilon(\omega,z)|^2\right]=\mathbb{E}[P_j(\omega,z)P_l(\omega,z)]\,.$$

Using the generator $\mathcal{L}_P$ we get a system of ordinary differential equations for limit fourth-order moments $(P_{jl}^{(2)})_{j,l=1,\dots,N}$ of the form

$$\begin{aligned}\frac{dP_{jj}^{(2)}}{dz}&=\sum_{n\ne j}\Gamma_{jn}^{(c)}\left(4P_{jn}^{(2)}-2P_{jj}^{(2)}\right),\\ \frac{dP_{jl}^{(2)}}{dz}&=-2\Gamma_{jl}^{(c)}P_{jl}^{(2)}+\sum_n\Gamma_{ln}^{(c)}\left(P_{jn}^{(2)}-P_{jl}^{(2)}\right)+\sum_n\Gamma_{jn}^{(c)}\left(P_{ln}^{(2)}-P_{jl}^{(2)}\right),\qquad j\ne l\,.\end{aligned}$$

The initial conditions are $P_{jl}^{(2)}(z=0)=|\hat a_{j,0}|^2|\hat a_{l,0}|^2$. This is a system of linear ordinary differential equations with constant coefficients that can be solved by computing the exponent of the evolution matrix.

It is straightforward to check that the function $P_{jl}^{(2)}\equiv1+\delta_{jl}$ is a stationary solution of the fourth-order moment system. Using the positivity of $\Gamma_{jl}^{(c)}$, $j\ne l$, we conclude that this stationary solution is asymptotically stable, which means that the solution $P_{jl}^{(2)}(z)$ starting from $P_{jl}^{(2)}(z=0)=|\hat a_{j,0}|^2|\hat a_{l,0}|^2$ converges as $z\to\infty$ to

$$P_{jl}^{(2)}(z)\overset{z\to\infty}{\longrightarrow}\begin{cases}\dfrac{1}{N(N+1)}R_0^4 & \text{if } j\ne l\,,\\ \dfrac{2}{N(N+1)}R_0^4 & \text{if } j=l\,,\end{cases}$$

where $R_0^2=\sum_{j=1}^N|\hat a_{j,0}|^2$. This implies that the correlation of $P_j(z)$ and $P_l(z)$ converges to $-1/(N-1)$ if $j\ne l$ and to $(N-1)/(N+1)$ if $j=l$ as $z\to\infty$. We see from the $j\ne l$ result that if, in addition, the number of modes N becomes large, then the mode powers become uncorrelated. The $j=l$ result shows that, whatever the number of modes N, the mode powers P_j are not statistically stable quantities.

5. Pulse propagation in waveguides. Bandwidth plays a basic role in the propagation of pulses in a waveguide because of dispersion. We assume that a point

source located inside the waveguide at $(z = 0, \mathbf{x} = \mathbf{x}_0)$ emits a pulse with carrier frequency ω_0 and bandwidth of order ε^2:

$$\mathbf{F}^\varepsilon(t,\mathbf{x},z) = f^\varepsilon(t)\delta(\mathbf{x}-\mathbf{x}_0)\delta(z)\mathbf{e}_z\,, \qquad f^\varepsilon(t) = f(\varepsilon^2 t)e^{i\omega_0 t}\,. \tag{5.1}$$

We have assumed in this model a narrowband source whose duration is of order ε^{-2}, that is, of the same order as the travel time through the waveguide whose length is L/ε^2. This is not the typical situation encountered in ultrasonic and underwater sound experiments in connection with time reversal [22, 15], where broadband pulses are used and, in the notation of this paper, $f^\varepsilon(t) = f(\varepsilon^p t)e^{i\omega_0 t}$ with $0 \le p < 2$. The analysis of these broadband regimes is carried out in [10]. The main results are similar in the two regimes, except for those regarding statistical stability in time reversal. We discuss this issue in sections 8.3 and 8.4, where we show that statistical stability in the narrowband case (5.1) can be achieved by mode diversity rather than by frequency diversity. The latter is typical for time-reversal refocusing of broadband pulses [10]. This is the reason that we consider the narrowband case in this paper.

The point source (5.1) generates left-going propagating modes that we do not need to consider, as they propagate in a homogeneous half-space, and right-going modes that we do analyze. As shown in section 2.3, the interface conditions at $z = 0$, which are initial conditions in the forward scattering approximation, have the form

$$\hat a_j^\varepsilon(\omega,0) = \frac{1}{2}\sqrt{\beta_j(\omega)}\hat f^\varepsilon(\omega)\phi_j(\mathbf{x}_0)\,, \qquad \hat f^\varepsilon(\omega) = \frac{1}{\varepsilon^2}\hat f\left(\frac{\omega-\omega_0}{\varepsilon^2}\right)$$

for $j \le N(\omega)$. The transmitted field observed in the plane $z = L/\varepsilon^2$ and at time t/ε^2 has the form

$$p_{tr}^\varepsilon(t,\mathbf{x},L) := p\left(\frac{t}{\varepsilon^2},\mathbf{x},\frac{L}{\varepsilon^2}\right),$$

$$p_{tr}^\varepsilon(t,\mathbf{x},L) = \frac{1}{4\pi\varepsilon^2}\int\sum_{j,l=1}^{N(\omega)}\frac{\sqrt{\beta_l(\omega)}}{\sqrt{\beta_j(\omega)}}\phi_j(\mathbf{x})\phi_l(\mathbf{x}_0)\hat f\left(\frac{\omega-\omega_0}{\varepsilon^2}\right)T_{jl}^\varepsilon(\omega)e^{i\frac{\beta_j(\omega)L-\omega t}{\varepsilon^2}}d\omega\,.$$

We change variables $\omega = \omega_0 + \varepsilon^2 h$ and expand $\beta_j(\omega_0+\varepsilon^2 h)$ with respect to ε:

$$p_{tr}^\varepsilon(t,\mathbf{x},L) = \frac{1}{4\pi}\int\sum_{j,l=1}^{N}\frac{\sqrt{\beta_l}}{\sqrt{\beta_j}}\phi_j(\mathbf{x})\phi_l(\mathbf{x}_0) \times\hat f(h)T_{jl}^\varepsilon(\omega_0+\varepsilon^2 h)e^{i\frac{\beta_j(\omega_0)L-\omega_0 t}{\varepsilon^2}}e^{i[\beta_j'(\omega_0)L-t]h}dh\,. \tag{5.2}$$

Here $\beta_j'(\omega)$ is the ω-derivative of $\beta_j(\omega)$. We do not show the dependence of N on ω_0 after we approximate $N(\omega_0+\varepsilon^2 h)$ by $N(\omega_0)$.

In a homogeneous waveguide we have that $T_{jl}^\varepsilon = \delta_{jl}$ and

$$p_{tr}^\varepsilon\,(t,\mathbf{x},L) = \frac{1}{2}\sum_{j=1}^{N}\phi_j(\mathbf{x})\phi_j(\mathbf{x}_0)e^{i\frac{\beta_j(\omega_0)L-\omega_0 t}{\varepsilon^2}}f\left(t-\beta_j'(\omega_0)L\right)\,. \tag{5.3}$$

The transmitted field is therefore a superposition of modes, each of which is centered around its travel time $\beta_j'(\omega_0)L$. The modal dispersion makes the overall spreading of the transmitted field linearly increasing with L.

In a random waveguide, the integral representation (5.2) shows that the moments of the transmitted field depend on the joint statistics of the entries of the transfer matrix at different frequencies. The next section will give the needed results.

6. Statistics of the transfer matrix.

6.1. Single frequency statistics of the transfer matrix. Using Proposition 4.1 with the special initial conditions $\hat{a}_j(0) = \delta_{jl}$, we obtain the first moments of the transmission coefficients.

PROPOSITION 6.1. *The expectations of the transmission coefficients* $\mathbb{E}[T^\varepsilon_{jl}(\omega, L)]$ *converge to zero as* $\varepsilon \to 0$ *if* $j \neq l$ *and to* $\bar{T}_j(\omega, L)$ *if* $j = l$*, where*

$$\bar{T}_j(\omega, L) = \exp\left(\frac{\Gamma^{(c)}_{jj}(\omega)L}{2} - \frac{\Gamma^{(1)}_{jj}(\omega)L}{2} + \frac{i\Gamma^{(s)}_{jj}(\omega)L}{2}\right). \tag{6.1}$$

The real part of the exponential factor is $[\Gamma^{(c)}_{jj}(\omega) - \Gamma^{(1)}_{jj}(\omega)]L/2$. The coefficient $\Gamma^{(c)}_{jj}(\omega)$ is negative. The coefficient $\Gamma^{(1)}_{jj}(\omega)$ is nonnegative because it is proportional to the power spectral density of C_{jj} at zero-frequency. As a result, the damping coefficient has a negative real part, and therefore the mean transmission coefficients decay exponentially with propagation distance.

Using Proposition 4.3, we immediately get the following result.

PROPOSITION 6.2. *The mean square moduli of the transmission coefficients have limits as* $\varepsilon \to 0$*,* $\lim_{\varepsilon\to 0} \mathbb{E}[|T^\varepsilon_{jl}(\omega, L)|^2] = \mathcal{T}^{(l)}_j(\omega, L)$*, which are the solutions of the system of linear equations*

$$\frac{d\mathcal{T}^{(l)}_j}{dz} = \sum_{n\neq j} \Gamma^{(c)}_{jn}(\omega)\left(\mathcal{T}^{(l)}_n - \mathcal{T}^{(l)}_j\right), \qquad \mathcal{T}^{(l)}_j(\omega, z = 0) = \delta_{jl}\,. \tag{6.2}$$

The coefficients $\Gamma^{(c)}_{jl}$ *are given by* (4.3)*.*

From the analysis of section 4.2, we know that the mean square moduli of the entries of the transfer matrix converge exponentially fast to the constant $1/N$.

6.2. Transport equations for the autocorrelation of the transfer matrix. In many physically interesting contexts, such as in calculating the mean transmitted intensity or the mean refocused field amplitude, we need two-frequency statistical information. We now introduce a proposition that describes the two-frequency statistical properties that we will need in the applications discussed in this paper.

PROPOSITION 6.3. *The autocorrelation function of the transmission coefficients at two nearby frequencies admits a limit as* $\varepsilon \to 0$*:*

$$\mathbb{E}[T^\varepsilon_{jj}(\omega, L)\overline{T^\varepsilon_{ll}(\omega - \varepsilon^2 h, L)}] \stackrel{\varepsilon\to 0}{\longrightarrow} e^{Q_{jl}(\omega)L} \text{ if } j \neq l\,, \tag{6.3}$$

$$\mathbb{E}[T^\varepsilon_{jl}(\omega, L)\overline{T^\varepsilon_{jl}(\omega - \varepsilon^2 h, L)}] \stackrel{\varepsilon\to 0}{\longrightarrow} e^{-i\beta'_j(\omega)hL} \int \mathcal{W}^{(l)}_j(\omega, \tau, L)e^{ih\tau}d\tau\,, \tag{6.4}$$

$$\mathbb{E}[T^\varepsilon_{jm}(\omega, L)\overline{T^\varepsilon_{ln}(\omega - \varepsilon^2 h, L)}] \stackrel{\varepsilon\to 0}{\longrightarrow} 0 \text{ in the other cases,} \tag{6.5}$$

where $(\mathcal{W}^{(l)}_j(\omega, \tau, z))_{j=1,\ldots,N(\omega)}$ *is the solution of the system of transport equations*

$$\frac{\partial \mathcal{W}^{(l)}_j}{\partial z} + \beta'_j(\omega)\frac{\partial \mathcal{W}^{(l)}_j}{\partial \tau} = \sum_{n\neq j} \Gamma^{(c)}_{jn}(\omega)\left(\mathcal{W}^{(l)}_n - \mathcal{W}^{(l)}_j\right), \tag{6.6}$$

starting from $\mathcal{W}_j^{(l)}(\omega,\tau,z=0)=\delta(\tau)\delta_{jl}$. *The coefficients* $\Gamma_{jl}^{(c)}$ *are given by* (4.3). *The damping factors* Q_{jl} *are*

$$Q_{jl}(\omega)=\frac{\Gamma_{jj}^{(c)}(\omega)+\Gamma_{ll}^{(c)}(\omega)}{2}-\frac{\Gamma_{jj}^{(1)}(\omega)+\Gamma_{ll}^{(1)}(\omega)-2\Gamma_{jl}^{(1)}(\omega)}{2} +i\frac{\Gamma_{jj}^{(s)}(\omega)-\Gamma_{ll}^{(s)}(\omega)}{2}\,. \tag{6.7}$$

We note that the real parts of the damping factors Q_{jl} are negative and that the solutions of the transport equations are measures. If $j\neq l$, then $\mathcal{W}_j^{(l)}$ has a continuous density, but $\mathcal{W}_l^{(l)}$ has a Dirac mass at $\tau=\beta_l'(\omega)z$ with weight $\exp(\Gamma_{ll}^{(c)}z)$, where $\Gamma_{ll}^{(c)}$ is given by (4.3), and a continuous density denoted by $\mathcal{W}_{l,c}^{(l)}(\omega,\tau,z)$:

$$\mathcal{W}_l^{(l)}(\omega,\tau,z)d\tau=e^{\Gamma_{ll}^{(c)}(\omega)z}\delta(\tau-\beta_l'(\omega)z)d\tau+\mathcal{W}_{l,c}^{(l)}(\omega,\tau,z)d\tau\,.$$

Note also that by integrating the system of transport equations with respect to τ, we recover the result of Proposition 6.2.

The system of transport equations describes the coupling between the N right-going modes. It is the main theoretical result of this paper. It describes the evolution of the coupled powers of the modes in frequency and time, with transport velocities equal to the group velocities of the modes $1/\beta_j'(\omega)$. Therefore, the transport equations (6.6) could have been written as the natural space-time generalization of the coupled power equations (6.2). The mathematical content of Proposition 6.3 gives the precise connection between the quantities that satisfy this simple and intuitive space-time extension of (6.2) and the moments of the random transfer matrix. The two-frequency nature of the statistical quantities that satisfy the transport equations is clear in Proposition 6.3.

Proof. For fixed indices m and n we consider the product of two transfer matrices,

$$U_{jl}^{\varepsilon}(\omega,h,z)=T_{jm}^{\varepsilon}(\omega,z)\overline{T_{ln}^{\varepsilon}(\omega-\varepsilon^2h,z)}\,,$$

and note that it is the solution of

$$\frac{dU_{jl}^{\varepsilon}}{dz}=\sum_{j_1=1}^{N}\frac{1}{\varepsilon}H_{\omega,jj_1}^{(a)}\left(\frac{z}{\varepsilon^2}\right)U_{j_1l}^{\varepsilon}+\sum_{l_1=1}^{N}\frac{1}{\varepsilon}\overline{H_{\omega-\varepsilon^2h,ll_1}^{(a)}}\left(\frac{z}{\varepsilon^2}\right)U_{jl_1}^{\varepsilon}\,,$$

with the initial conditions $U_{jl}^{\varepsilon}(\omega,h,z=0)=\delta_{mj}\delta_{nl}$. We expand $\beta_\cdot(\omega-\varepsilon^2h)$ with respect to ε, and we introduce the Fourier transform

$$V_{jl}^{\varepsilon}(\omega,\tau,z)=\frac{1}{2\pi}\int e^{-ih(\tau-\beta_l'(\omega)z)}U_{jl}^{\varepsilon}(\omega,h,z)dh\,,$$

which is the solution of

$$\frac{\partial V_{jl}^{\varepsilon}}{\partial z}+\beta_l'(\omega)\frac{\partial V_{jl}^{\varepsilon}}{\partial\tau}=\sum_{j_1=1}^{N}\frac{1}{\varepsilon}H_{\omega,jj_1}^{(a)}\left(\frac{z}{\varepsilon^2}\right)V_{j_1l}^{\varepsilon}+\sum_{l_1=1}^{N}\frac{1}{\varepsilon}\overline{H_{\omega,ll_1}^{(a)}}\left(\frac{z}{\varepsilon^2}\right)V_{jl_1}^{\varepsilon}\,,$$

with the initial conditions $V_{jl}^{\varepsilon}(\omega,\tau,z=0)=\delta_{mj}\delta_{nl}\delta(\tau)$. We can now apply the diffusion approximation theorem [19] and get the result stated in the proposition. ∎

6.3. Probabilistic interpretation of the transport equations. The transport equations (6.6) have a probabilistic representation that can be used for Monte Carlo simulations as well as for getting a diffusion approximation result. It is primarily this diffusion approximation that we want to derive in this section. We will use it in the applications that follow. We introduce the jump Markov process J_z, whose state space is $\{1,\dots,N(\omega)\}$ and whose infinitesimal generator is

$$\mathcal{L}\phi(j) = \sum_{l\neq j} \Gamma^{(c)}_{jl}(\omega)\left(\phi(l)-\phi(j)\right) . \tag{6.8}$$

We also define the process $\mathcal{B}_z$ by

$$\mathcal{B}_z = \int_0^z \beta'_{J_s} ds\,, \quad z\geq 0\,, \tag{6.9}$$

which is well defined because J_z is piecewise constant. In a manner similar to that in [1], we get the probabilistic representation of the solutions to system (6.2) and the solutions to the transport equations (6.6) in terms of the jump Markov process J_z:

$$\mathcal{T}^{(n)}_j(\omega,L) = \mathbb{P}\left(J_L = j \mid J_0 = n\right) , \tag{6.10}$$

$$\int_{\tau_0}^{\tau_1} \mathcal{W}^{(n)}_j(\omega,\tau,L)d\tau = \mathbb{P}\left(J_L = j\ ,\ \mathcal{B}_L\in[\tau_0,\tau_1] \mid J_0 = n\right) . \tag{6.11}$$

The process J_z is an irreducible, reversible, and ergodic Markov process. Its distribution converges as $z\to\infty$ to the uniform distribution over $\{1,\dots,N\}$. The convergence is exponential with a rate that is equal to the second eigenvalue of the matrix $\Gamma^{(c)} = (\Gamma^{(c)}_{jl})_{j,l=1,\dots,N}$. The first eigenvalue of this matrix is zero and the associated eigenvector is the uniform distribution over $\{1,\dots,N\}$. The second eigenvalue can be written in the form $-1/L_{\text{equip}}$, which defines the equipartition distance L_{equip}.

We next determine the asymptotic distribution of the process $\mathcal{B}_z$. From the ergodic theorem we have that with probability one,

$$\frac{\mathcal{B}_z}{z} \overset{z\to\infty}{\longrightarrow} \overline{\beta'(\omega)}\,, \text{ where } \overline{\beta'(\omega)} = \frac{1}{N(\omega)}\sum_{j=1}^{N(\omega)} \beta'_j(\omega)\,. \tag{6.12}$$

We can interpret the z large limit to mean that z is considerably larger than L_{equip}.

For a planar waveguide we have that $\beta_j = \sqrt{\omega^2/c^2 - \pi^2 j^2/d^2}$ and $N(\omega) = [(\omega d)/(\pi\bar{c})]$. In the continuum limit $N(\omega)\gg 1$, we obtain the expression $\overline{\beta'(\omega)} = \pi/(2\bar{c})$, which is independent of ω. This ω independence property is likely to hold for a broad class of waveguides.

By applying a central limit theorem for functionals of ergodic Markov processes, we find that, in distribution,

$$\frac{\mathcal{B}_z - \overline{\beta'(\omega)}z}{\sqrt{z}} \overset{z\to\infty}{\longrightarrow} \mathcal{N}(0,\sigma^2_{\beta'(\omega)})\,. \tag{6.13}$$

Here $\mathcal{N}(0,\sigma^2_{\beta'(\omega)})$ is a zero-mean Gaussian random variable with variance

$$\sigma^2_{\beta'(\omega)} = 2\int_0^\infty \mathbb{E}_e\left[(\beta'_{J_0}(\omega)-\overline{\beta'(\omega)})(\beta'_{J_s}(\omega)-\overline{\beta'(\omega)})\right] ds\,, \tag{6.14}$$

where $\mathbb{E}_e$ stands for expectation with respect to the stationary process J_z. These limit theorems imply that when $L \gg L_{\rm equip}$, we have

$$\mathcal{T}_j^{(n)}(\omega, L) \overset{L\gg L_{\rm equip}}{\simeq} \frac{1}{N(\omega)}\,, \tag{6.15}$$

$$\mathcal{W}_j^{(n)}(\omega,\tau,L) \overset{L\gg L_{\rm equip}}{\simeq} \frac{1}{N(\omega)}\frac{1}{\sqrt{2\pi\sigma^2_{\beta'(\omega)}L}}\exp\left(-\frac{(\tau-\overline{\beta'(\omega)}L)^2}{2\sigma^2_{\beta'(\omega)}L}\right)\,. \tag{6.16}$$

The asymptotic result (6.15) shows that $\mathcal{T}_j^{(n)}$ becomes independent of n, the initial mode index, and uniform over $j \in \{1,\dots,N(\omega)\}$. This is the regime of energy equipartition among all propagating modes. The asymptotic result (6.16) is equivalent to the diffusion approximation for the system of transport equations (6.6).

7. Pulse propagation in random waveguides. We consider the transmitted field (5.2) obtained in the setup described in section 5, and we now address the case of a random waveguide. By Proposition 6.1, the mean transmitted field in the asymptotic $\varepsilon \to 0$ is given by

$$\mathbb{E}[p^\varepsilon_{tr}(t,\mathbf{x},L)] = \frac{1}{2}\sum_{j=1}^N \phi_j(\mathbf{x})\phi_j(\mathbf{x}_0)\bar{T}_j(\omega_0,L)e^{i\frac{\beta_j(\omega_0)L-\omega_0 t}{\varepsilon^2}} f\left(t-\beta_j'(\omega_0)L\right)\,.$$

As in the homogeneous case, the mean field is a superposition of modes in the random case, but the mean transmission coefficients are exponentially damped and vanish for L large, $L > L_{\rm equip}(\omega_0)$. Therefore, the mean field vanishes for large L. We now turn our attention to the mean intensity, which accounts for the conversion of the coherent field into incoherent wave fluctuations. We express the transmitted intensity as the expectation of a double integral

$$\begin{aligned}\mathbb{E}\left[\left|p^\varepsilon_{tr}(t,\mathbf{x},L)\right|^2\right] = \frac{1}{16\pi^2}\sum_{j,l=1}^N\sum_{m,n=1}^N \frac{\sqrt{\beta_l\beta_n}}{\sqrt{\beta_j\beta_m}}\phi_j(\mathbf{x})\phi_l(\mathbf{x}_0)\phi_m(\mathbf{x})\phi_n(\mathbf{x}_0)\\ \times e^{i\frac{[\beta_j(\omega_0)-\beta_m(\omega_0)]L}{\varepsilon^2}}\int\int \hat{f}\,(h)\,\overline{\hat{f}\,(h')}\mathbb{E}[T^\varepsilon_{jl}(\omega_0+\varepsilon^2 h)\overline{T^\varepsilon_{mn}(\omega_0+\varepsilon^2 h')}]\\ \times e^{i[\beta_j'(\omega_0)L-t]h-[\beta_m'(\omega_0)L-t]h'}dhdh'\,.\end{aligned}$$

Using Proposition 6.3 we see that there are two contributions to this integral:

$$\mathbb{E}\left[\left|p^\varepsilon_{tr}(t,\mathbf{x},L)\right|^2\right] = I_1^\varepsilon(t,\mathbf{x},L) + I_2^\varepsilon(t,\mathbf{x},L)\,. \tag{7.1}$$

The limit of the first contribution is

$$\begin{aligned}I_1^\varepsilon(t,\mathbf{x},L) \overset{\varepsilon\to 0}{\simeq} \frac{1}{4}\sum_{j\neq m=1}^N \phi_j(\mathbf{x})\phi_j(\mathbf{x}_0)\phi_m(\mathbf{x})\phi_m(\mathbf{x}_0)e^{i\frac{[\beta_j(\omega_0)-\beta_m(\omega_0)]L}{\varepsilon^2}}\\ \times e^{Q_{jm}(\omega_0)L}f(t-\beta_j'(\omega_0)L)f(t-\beta_m'(\omega_0)L)\,.\end{aligned} \tag{7.2}$$

We see that it decays exponentially with the propagation distance because of the damping factors $\exp(Q_{jm}(\omega_0)L)$. We can therefore neglect this contribution for $L \gg L_{\rm equip}(\omega_0)$. The limit of the second contribution is

$$I_2^\varepsilon(t,\mathbf{x},L) \overset{\varepsilon\to 0}{\simeq} \frac{1}{4}\sum_{j,l=1}^N \frac{\beta_l}{\beta_j}\phi_j^2(\mathbf{x})\phi_l^2(\mathbf{x}_0)\int \mathcal{W}_j^{(l)}(\omega_0,\tau,L)f(t-\tau)^2 d\tau\,. \tag{7.3}$$

In the asymptotic equipartition regime $L \gg L_{\rm equip}(\omega_0)$ we use the diffusion approximation (6.16). We conclude that

$$\lim_{\varepsilon\to 0} \mathbb{E}\left[\left|p^\varepsilon_{tr}\left(t,\mathbf{x},L\right)\right|^2\right] \overset{L\gg L_{\rm equip}}{\simeq} H_{\omega_0,\mathbf{x}_0}(\mathbf{x}) \times [K_{\omega_0,L} * (f^2)](t)\,, \tag{7.4}$$

where the spatial profile $H_{\omega_0,\mathbf{x}_0}$ and the time convolution kernel $K_{\omega_0,L}$ are given by

$$H_{\omega_0,\mathbf{x}_0}(\mathbf{x}) = \frac{1}{4N(\omega_0)} \sum_{j=1}^{N(\omega_0)} \frac{\phi_j^2(\mathbf{x})}{\beta_j(\omega_0)} \times \sum_{l=1}^{N(\omega_0)} \phi_l^2(\mathbf{x}_0)\beta_l(\omega_0)\,, \tag{7.5}$$

$$K_{\omega_0,L}(t) = \frac{1}{\sqrt{2\pi\sigma^2_{\beta'(\omega_0)}L}} \exp\left(-\frac{(t-\overline{\beta'(\omega_0)}L)^2}{2\sigma^2_{\beta'(\omega_0)}L}\right). \tag{7.6}$$

To sum up, the main results of this section are that

- the mean field decays exponentially with propagation distance,
- the mean transmitted intensity converges to the transverse spatial profile $H_{\omega_0,\mathbf{x}_0}$, and
- the mean transmitted intensity is concentrated around the time $\overline{\beta'(\omega_0)}L$ with a spread that is of the order of $\sigma_{\beta'(\omega_0)}\sqrt{L} \sim \sqrt{L L_{\rm equip}(\omega_0)}/\bar{c}$ for a pulse with carrier frequency ω_0.

Note that $\sigma_{\beta'(\omega_0)}\sqrt{L} \ll L/\bar{c}$, which means that pulse spreading increases as $\sqrt{L}$ in a random waveguide while it increases linearly in a homogeneous one. This is because the modes are strongly coupled together and propagate with the same "average" group velocity $1/\overline{\beta'(\omega_0)}$ in the random waveguide. The "average" group velocity is the harmonic average of the group velocities of the modes $1/\beta'_j(\omega_0)$. These results are intuitively clear, but they were not discussed in detail in the early literature [14, 5]. We note that in (7.4) it is the mean of the pulse intensity that has the asymptotic form that we have derived. The pulse intensity fluctuations can also be computed and are not small.

8. Time reversal in a waveguide.

8.1. Time reversal setup. We now consider time reversal in a waveguide. A point source located in the plane $z = 0$ at the lateral position $\mathbf{x}_0$ emits a pulse $f^\varepsilon(t)$ of the form (5.1). A time-reversal mirror is located in the plane $z = L/\varepsilon^2$ and occupies the subdomain $\mathcal{D}_M \subset \mathcal{D}$. The transmitted wave observed in the plane $z = L/\varepsilon^2$ at time t/ε^2 is (5.2). The time-reversal mirror records the field from time t_0/ε^2 up to time t_1/ε^2, time reverses it, and sends it back into the waveguide. The new source at the time-reversal mirror that generates the back propagating waves is

$$\begin{aligned}\mathbf{F}^\varepsilon_{\rm TR}(t,\mathbf{x},z) &= f^\varepsilon_{\rm TR}(t,\mathbf{x})\delta\left(z-\frac{L}{\varepsilon^2}\right)\mathbf{e}_z\,,\\ f^\varepsilon_{\rm TR}(t,\mathbf{x}) &= p^\varepsilon_{tr}\left(t_1-\varepsilon^2 t,\mathbf{x},L\right) G_1(t_1-\varepsilon^2 t) G_2(\mathbf{x})\,,\end{aligned} \tag{8.1}$$

where p^ε_{tr} is given by (5.2), G_1 is the time-window function of the form $G_1(t) = \mathbf{1}_{[t_0,t_1]}(t)$, and G_2 is the spatial-window function $G_2(\mathbf{x}) = \mathbf{1}_{\mathcal{D}_M}(\mathbf{x})$. We have seen that the power delay spread of the transmitted signal is not very long in the forward scattering approximation. This is because there is no backscattering to produce long codas (i.e., long incoherent wave fluctuations). Moreover, we focus our attention more on spatial effects in this paper, so it is reasonable to assume that we record the

field for all time at the time-reversal mirror. This means that we have $G_1(t) = 1$ and $\hat{G}_1(h) = 2\pi\delta(h)$. Therefore, the left-going propagating modes generated by this source have amplitudes

$$\hat{b}_m(\omega) = -\frac{\sqrt{\beta_m(\omega)}}{2}\int_{\mathcal{D}} \hat{f}^{\varepsilon}_{\rm TR}(\omega,\mathbf{x})\phi_m(\mathbf{x})d\mathbf{x}e^{i\beta_m(\omega)\frac{L}{\varepsilon^2}}$$

$$= \frac{1}{4\varepsilon^2}\sum_{j,l=1}^{N}\frac{\sqrt{\beta_l\beta_m}}{\sqrt{\beta_j}}M_{mj}\phi_l(\mathbf{x}_0)\overline{\hat{f}\left(\frac{\omega-\omega_0}{\varepsilon^2}\right)}\,\overline{T^{\varepsilon}_{jl}(\omega)}e^{i[\beta_m(\omega)-\beta_j(\omega)]\frac{L}{\varepsilon^2}+i\omega\frac{t_1}{\varepsilon^2}}\,, \tag{8.2}$$

where the coupling coefficients M_{jl} are given by

$$M_{jl} = \int_{\mathcal{D}} \phi_j(\mathbf{x})G_2(\mathbf{x})\phi_l(\mathbf{x})d\mathbf{x}\,. \tag{8.3}$$

We have explicit formulas for the coupling coefficients M_{jl} in two cases as follows:

- If the mirror spans the complete cross section $\mathcal{D}$ of the waveguide, then we have $G_2(\mathbf{x}) = 1$ and $M_{jl} = \delta_{jl}$.
- If the mirror is point-like at $\mathbf{x} = \mathbf{x}_1$, meaning $G_2(\mathbf{x}) = |\mathcal{D}|\delta(\mathbf{x}-\mathbf{x}_1)$, with the factor $|\mathcal{D}|$ added for dimensional consistency, then $M_{jl} = |\mathcal{D}|\phi_j(\mathbf{x}_1)\phi_l(\mathbf{x}_1)$.

The refocused field observed in the plane $z = 0$ in the Fourier domain is given by

$$\hat{p}_{\rm TR}(\omega,\mathbf{x},0) = \sum_{m,n=1}^{N}\frac{(\mathbf{T}^{\varepsilon})^T_{nm}(\omega)\hat{b}_m(\omega)}{\sqrt{\beta_n}}\phi_n(\mathbf{x})\,. \tag{8.4}$$

Here $(\mathbf{T}^{\varepsilon})^T(\omega)$ is the transfer matrix for the left-going modes propagating from L/ε^2 to 0, and it is the transpose of $\mathbf{T}^{\varepsilon}(\omega)$. This follows from the unitarity of the transfer matrix $\mathbf{T}^{\varepsilon}(\omega)$. In the time domain, the refocused field observed at time $t_{\rm obs}/\varepsilon^2$ is

$$p_{\rm TR}\left(\frac{t_{\rm obs}}{\varepsilon^2},\mathbf{x},0\right) = \frac{1}{4\pi}\sum_{j,l,m,n=1}^{N}\frac{\sqrt{\beta_l\beta_m}}{\sqrt{\beta_j\beta_n}}M_{mj}\phi_n(\mathbf{x})\phi_l(\mathbf{x}_0)e^{i[\beta_m-\beta_j](\omega_0)\frac{L}{\varepsilon^2}+i\omega_0\frac{t_1-t_{\rm obs}}{\varepsilon^2}}$$

$$\times\int\overline{\hat{f}\,(h)}\,\overline{T^{\varepsilon}_{jl}(\omega_0+\varepsilon^2 h)}T^{\varepsilon}_{mn}(\omega_0+\varepsilon^2 h)e^{i\{[\beta'_m-\beta'_j](\omega_0)L+(t_1-t_{\rm obs})\}h}dh\,. \tag{8.5}$$

8.2. Refocusing in a homogeneous waveguide. In this case, $T^{\varepsilon}_{jl} = \delta_{jl}$ and the refocused field is

$$p_{\rm TR}\left(\frac{t_{\rm obs}}{\varepsilon^2},\mathbf{x},0\right) = \frac{1}{2}e^{i\omega_0\frac{t_1-t_{\rm obs}}{\varepsilon^2}}\sum_{j,m=1}^{N}e^{i[\beta_m-\beta_j](\omega_0)\frac{L}{\varepsilon^2}}$$

$$\times M_{mj}\phi_m(\mathbf{x})\phi_j(\mathbf{x}_0)f\left([\beta'_m-\beta'_j](\omega_0)L+t_1-t_{\rm obs}\right)\,. \tag{8.6}$$

The refocused field is a weighted sum of modes, whose weights depend on the mirror through the coefficients M_{mj}. The oscillatory terms in (8.6) produce transverse sidelobes in the refocused field, as seen in Figure 8.1.

8.3. The mean refocused field in a random waveguide. In the analysis of time reversal considered up to now [3, 9, 2, 21], statistical stability is shown by a frequency decoherence argument. More precisely, it is shown that the refocused field is the superposition of many approximately uncorrelated frequency components, which ensures self-averaging in the time domain. This argument cannot be used in

the narrowband case (5.1) that we consider here. This is because the decoherence frequency, which is of order ε^2, is comparable to the bandwidth, which is also of order ε^2. In broadband cases with a bandwidth of order ε^p, $p \in [0,2)$, statistical stability can be obtained by the usual frequency decoherence argument [10].

First we compute the mean refocused field and then consider the statistical stability. By taking the expectation of (8.5), we find that the mean refocused field involves the second-order moments of the transfer matrix. From Proposition 6.3 we have the limit values of these second-order moments, so we can write

$$\mathbb{E}\left[p_{\mathrm{TR}}\left(\frac{t_{\mathrm{obs}}}{\varepsilon^2},\mathbf{x},0\right)\right] = p_1^\varepsilon + p_2^\varepsilon\,, \tag{8.7}$$

$$p_1^\varepsilon \overset{\varepsilon\to 0}{\simeq} \frac{1}{2}\sum_{j\neq m=1}^{N} M_{mj}\phi_m(\mathbf{x})\phi_j(\mathbf{x}_0)e^{i[\beta_m-\beta_j](\omega_0)\frac{L}{\varepsilon^2}+i\omega_0\frac{t_1-t_{\mathrm{obs}}}{\varepsilon^2}} \times e^{Q_{jm}(\omega_0)L} f\left([\beta'_m-\beta'_j](\omega_0)L+t_1-t_{\mathrm{obs}}\right)\,,$$

$$p_2^\varepsilon \overset{\varepsilon\to 0}{\simeq} \frac{1}{2}e^{i\omega_0\frac{t_1-t_{\mathrm{obs}}}{\varepsilon^2}} f\left(t_1-t_{\mathrm{obs}}\right)\sum_{j,l=1}^{N} M_{jj}\phi_l(\mathbf{x})\phi_l(\mathbf{x}_0)\mathcal{T}_j^{(l)}(\omega_0,L)\,.$$

The term p_1^ε decays exponentially with propagation distance because of the damping factors coming from Q_{jm}. We can therefore neglect this term in the asymptotic equipartition regime. The term p_2^ε does contribute and it refocuses around the time $t_{\mathrm{obs}} = t_1$ with the original pulse shape time reversed. The spatial focusing profile is a weighted sum of modes, with weights that depend on the time-reversal mirror through the coefficients M_{jl} and on the mean square transmission coefficients $\mathcal{T}_j^{(l)}$.

In the asymptotic equipartition regime $L \gg L_{\mathrm{equip}}$, the coefficients $\mathcal{T}_j^{(l)}(\omega_0,L)$ converge to $1/N$ for all j and l, which for the mean refocused field gives

$$\lim_{\varepsilon\to 0}\mathbb{E}\left[p_{\mathrm{TR}}\left(\frac{t_{\mathrm{obs}}}{\varepsilon^2},\mathbf{x},0\right)\right] \overset{L\gg L_{\mathrm{equip}}}{\simeq} e^{i\omega_0\frac{t_1-t_{\mathrm{obs}}}{\varepsilon^2}} f\left(t_1-t_{\mathrm{obs}}\right) \times\frac{1}{N(\omega_0)}\sum_{j}^{N(\omega_0)} M_{jj}\times\frac{1}{2}\sum_{l=1}^{N(\omega_0)}\phi_l(\mathbf{x})\phi_l(\mathbf{x}_0)\,. \tag{8.8}$$

We note the difference between this expression and (8.6) in a homogeneous waveguide. The oscillatory terms in (8.6) which generate the side-lobes are suppressed in (8.8). The analytical understanding of side-lobe suppression in time reversal in random waveguides is a new result in this paper.

The spatial refocusing profile can then be computed explicitly because it does not depend on the mirror shape or size. In the case of a planar waveguide, we have $\phi_j(x) = \sqrt{2/d}\sin(\pi jx/d)$, and in the continuum limit $N \gg 1$, we have

$$\frac{1}{2}\sum_{l=1}^{N}\phi_l(x)\phi_l(x_0) \overset{N\gg 1}{\simeq} \frac{1}{\lambda_0}\mathrm{sinc}\left(2\pi\frac{x-x_0}{\lambda_0}\right)\,.$$

The mean refocused field is therefore concentrated around the original source location x_0 with a resolution of half of a wavelength, which is the diffraction limit.

8.4. Statistical stability of the refocused field. As we noted already, we cannot claim that the refocused field is statistically stable by using the same argument

as in the broadband case, because here we have a narrowband pulse. However, we can achieve statistical stability through the summation over the modes. We will show this in the quasi-monochromatic case, where the pulse envelope $f(t) = 1$ and $\hat{f}(h) = 2\pi\delta(h)$. In this case, it is clear that statistical stability cannot arise from time averaging, and the refocused field is

$$p_{\rm TR}\left(\frac{t_{\rm obs}}{\varepsilon^2},\mathbf{x},0\right) = \frac{1}{2}\sum_{j,l,m,n=1}^{N}\frac{\sqrt{\beta_l\beta_m}}{\sqrt{\beta_j\beta_n}}M_{mj}\phi_n(\mathbf{x})\phi_l(\mathbf{x}_0) \times e^{i[\beta_m-\beta_j](\omega_0)\frac{L}{\varepsilon^2}+i\omega_0\frac{t_1-t_{\rm obs}}{\varepsilon^2}}\overline{T^\varepsilon_{jl}(\omega_0)}T^\varepsilon_{mn}(\omega_0)\,. \tag{8.9}$$

From (8.8), the mean refocused field at $\mathbf{x} = \mathbf{x}_0$ is in the asymptotic equipartition regime

$$\lim_{\varepsilon\to 0}\mathbb{E}\left[p_{\rm TR}\left(\frac{t_{\rm obs}}{\varepsilon^2},\mathbf{x}_0,0\right)\right] \overset{L\gg L_{\rm equip}}{\simeq} e^{i\omega_0\frac{t_1-t_{\rm obs}}{\varepsilon^2}}\frac{R_0^2}{2N}\sum_{j=1}^{N}M_{jj}\,, \tag{8.10}$$

where $R_0^2 = \sum_{l=1}^{N}\phi_l^2(\mathbf{x}_0)$. We now compute the second moment of the refocused field observed at $\mathbf{x} = \mathbf{x}_0$. By taking the expectation of the square of (8.9), we find that this moment involves fourth-order moments of the transfer matrix. From the results of section 4.3, we have

$$\lim_{\varepsilon\to 0}\mathbb{E}[\overline{T^\varepsilon_{jl}}T^\varepsilon_{mn}\overline{T^\varepsilon_{j'l'}}T^\varepsilon_{m'n'}] \overset{L\gg L_{\rm equip}}{\simeq} \begin{cases} \frac{2}{N(N+1)} & \text{if } (j,l)=(m,n)=(j',l')=(m',n')\,,\\ \frac{1}{N(N+1)} & \text{if } (j,l)=(m,n)\neq(j',l')=(m',n')\,,\\ \frac{1}{N(N+1)} & \text{if } (j,l)=(m',n')\neq(j',l')=(m,n)\,,\\ 0 & \text{otherwise}\,. \end{cases}$$

Using these fourth-order moment results in the expression for the second moment of the refocused field we see that, in the limit $\varepsilon \to 0$ and in the asymptotic equipartition regime $L \gg L_{\rm equip}$,

$$\lim_{\varepsilon\to 0}\mathbb{E}\left[\left|p_{\rm TR}\left(\frac{t_{\rm obs}}{\varepsilon^2},\mathbf{x}_0,0\right)\right|^2\right] \overset{L\gg L_{\rm equip}}{\simeq} \frac{R_0^4}{4N(N+1)}\left[\left(\sum_j M_{jj}\right)^2+\sum_{j,j'}M_{jj'}^2\right].$$

Let us introduce the relative standard deviation S of the refocused field amplitude,

$$S^2 := \lim_{\varepsilon\to 0}\frac{\mathbb{E}\left[\left|p_{\rm TR}\left(\frac{t_{\rm obs}}{\varepsilon^2},\mathbf{x}_0,0\right)\right|^2\right]-\left|\mathbb{E}\left[p_{\rm TR}\left(\frac{t_{\rm obs}}{\varepsilon^2},\mathbf{x}_0,0\right)\right]\right|^2}{\left|\mathbb{E}\left[p_{\rm TR}\left(\frac{t_{\rm obs}}{\varepsilon^2},\mathbf{x}_0,0\right)\right]\right|^2}\,. \tag{8.11}$$

We have statistical stability when S is small. In the asymptotic equipartition regime, S^2 is given by

$$S^2 \overset{L\gg L_{\rm equip}}{\simeq} -\frac{1}{N+1}+\frac{N}{N+1}\frac{1}{Q_{\rm mirror}}\,,\qquad Q_{\rm mirror} = \frac{\sum_{j,l=1}^{N}M_{jj}M_{ll}}{\sum_{j,l=1}^{N}M_{jl}^2}\,. \tag{8.12}$$

The quality factor $Q_{\rm mirror}$ depends only on the time-reversal mirror. We will have statistical stability when the number of modes N is large and when the quality factor $Q_{\rm mirror}$ is large. This analytical criterion for statistical stability in narrowband time reversal is a new result in this paper.

We can consider two extreme cases:

- If the time-reversal mirror spans the waveguide cross section, then $M_{jl} = \delta_{jl}$ and the quality factor is equal to N, which is optimal since the relative standard deviation is then zero for any N. This result is not surprising since the time-reversal mirror records the transmitted signal fully, in both time and space, which implies optimal refocusing.
- If the time-reversal mirror is point-like at $\mathbf{x}_1$, then $M_{jl} = \phi_j(\mathbf{x}_1)\phi_l(\mathbf{x}_1)$ and the quality factor is 1, which is bad, because the relative standard deviation S is asymptotically equal to $\sqrt{N-1}/\sqrt{N+1}$. The fluctuations of the refocused field are, therefore, of the same order as the mean field, which means that there is no statistical stability.

In the next section, we address a particular case which allows explicit calculations.

8.5. Numerical illustration. In this section we illustrate the time reversal results in the quasi-monochromatic case for a particular random waveguide. We consider a random planar waveguide with diameter d. The random process ν is stationary and mixing in the z-direction. Its autocorrelation function is

$$\mathbb{E}[\nu(x,z)\nu(x',z')] = \sigma^2 \exp\left(-\frac{|z-z'|}{l_c}\right) R(x,x')\,,$$

where the support of R is contained in $[0,d]^2$, σ is the standard deviation of the medium fluctuations, and l_c is the axial correlation length. This decomposition of the autocorrelation function makes it easy to compute the effective coefficients $\Gamma^{(c)}_{jl}$, $\Gamma^{(s)}_{jl}$, and $\Gamma^{(1)}_{jl}$. We obtain

$$\gamma^{(c)}_{jl} = \frac{2\sigma^2 l_c G_{l,j}}{1+(\beta_j-\beta_l)^2 l_c^2}\,, \qquad \gamma^{(s)}_{jl} = \frac{2\sigma^2(\beta_j-\beta_l) l_c^2 G_{l,j}}{1+(\beta_j-\beta_l)^2 l_c^2}\,, \qquad \gamma^{(1)}_{jl} = 2 l_c G_{l,j}\,,$$
$$G_{l,j} = S_{j-l,j-l} + S_{j+l,j+l} - S_{j-l,j+l} - S_{j+l,j-l}\,,$$

where

$$S_{j,l} = \frac{1}{d^2}\int_0^d \int_0^d \cos\left(\frac{j\pi x}{d}\right)\cos\left(\frac{l\pi x'}{d}\right) R(x,x')dxdx'\,.$$

For simplicity, we shall make two hypotheses. First, we introduce a band-limiting idealization; i.e., we assume that the support of S lies with a finite square so that $S(j,l) \neq 0$ only if $|j| \leq 1$ and $|l| \leq 1$. Second, we assume that l_c is smaller than $\beta_j - \beta_{j-1}$ for all j. The expressions of the effective coefficients can then be simplified. The matrix $\Gamma^{(s)}_{jl}$ is essentially zero, while the matrices $\Gamma^{(c)}_{jl}$ and $\Gamma^{(s)}_{jl}$ are tridiagonal,

$$\Gamma^{(c)}_{jl} \simeq \Gamma^{(1)}_{jl} \simeq \frac{\omega^4\sigma^2 l_c d S_{1,1}}{4\bar{c}^4\beta_j(\omega)\beta_l(\omega)} \text{ if } |j-l| = 1\,, \qquad \Gamma^{(1)}_{jj} \simeq \frac{\omega^4\sigma^2 l_c d S_{0,0}}{4\bar{c}^4\beta_j^2(\omega)}\,,$$

and $\Gamma^{(c)}_{jj}$ is chosen so that the lines of the matrix are zero.

Homogeneous waveguide. In the homogeneous case the spatial profile of the refocused field is (8.6). Let us consider a time-reversal mirror of size a located in $x \in [d/2-a/2, d/2+a/2]$: $G_2(x) = \mathbf{1}_{[d/2-a/2,d/2+a/2]}(x)$. We then have

$$M_{jl} = \frac{a}{d}\left[\cos\left(\frac{(j-l)\pi}{2}\right)\operatorname{sinc}\left(\frac{(j-l)\pi a}{2d}\right) - \cos\left(\frac{(j+l)\pi}{2}\right)\operatorname{sinc}\left(\frac{(j+l)\pi a}{2d}\right)\right].$$

Using these formulas, we plot in Figure 8.1 the spatial profile of the refocused field for different sizes a of the time-reversal mirror. The peak at the original source location is there in all cases, but for small time-reversal mirrors, large side-lobes appear.

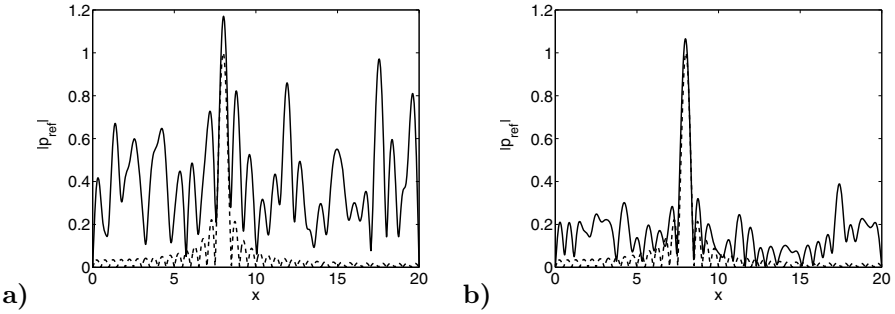


FIG. 8.1. *Transverse profile of the refocused field in a homogeneous waveguide with diameter d and length L. Here $d = 20$, $L = 200$, and $\lambda_0 = 1$, so there are 40 modes. The original source location is $x_0 = 8$. The dashed curve is the sinc profile, which is the focusing profile of a full-size time-reversal mirror. The solid curves are refocusing profiles for time-reversal mirrors of size $a = 2.5$* (a) *and $a = 10$* (b).

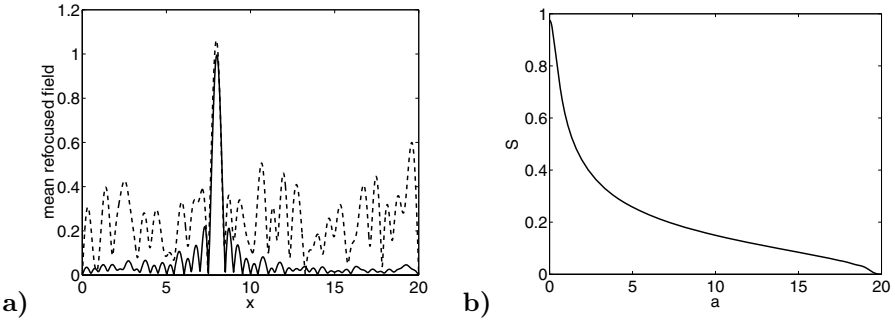


FIG. 8.2. (a) *Transverse profile of the mean refocused field in a random waveguide with diameter d and length L. Here $d = 20$, $L = 200$, $\lambda_0 = 1$, the time-reversal mirror has size $a = 5$, and the original source location is $x_0 = 8$. The axial correlation length of the random medium is $l_c = 0.25$ and the random fluctuations have standard deviation σ. The dashed curve is the spatial profile obtained in homogeneous medium $\sigma = 0$. The solid lines stand for the mean profiles obtained in random media $\sigma = 0.015$. The random case is very close to the equipartition regime, for which the focusing profile is a sinc.* (b) *The relative standard deviation S, from* (8.12)*, of the refocused field in the equipartition regime as a function of the mirror size a. Here $d = 20$ and $\lambda_0 = 1$.*

Random waveguide. The mean spatial profile of the refocused field is (8.7). In the absence of randomness, this formula reduces to (8.6). The mean spatial profile is plotted in Figure 8.2(a), which illustrates the transition from the poor spatial refocusing in a homogeneous medium to the diffraction-limited refocusing obtained in the equipartition regime. In the equipartition regime $L \gg L_{\rm equip}$, the mean spatial profile is given by (8.8) and becomes independent of the mirror size, up to an amplitude factor. However, the statistical stability of the refocused field depends on the size of the time-reversal mirror, as shown in Figure 8.2(b).

9. Summary and conclusions. The main result of this paper is the derivation from first principles of the system of transport equations (6.6) for the coupled mode powers, in the asymptotic limit of section 3. It is easy to write such equations by simply adding a time-dispersive term in (4.11). We identify here the field quantity that has the mean which satisfies this space-time transport equation.

We apply the transport equation (6.6) to pulse spreading in order to get (7.4),

which shows how randomness reduces time dispersion at the expense of introducing random fluctuations. We also apply it to time reversal in a random waveguide, in a narrowband regime, and show in (8.8) how side-lobes are suppressed in refocusing. We also show in (8.12) how statistical stability depends on the quality factor Q_{mirror}. This quality factor does not depend on the random medium in the energy equipartition regime, but it does depend on the size of the time-reversal mirror relative to the waveguide cross section.

REFERENCES

[1] M. Asch, W. Kohler, G. Papanicolaou, M. Postel, and B. White, *Frequency content of randomly scattered signals*, SIAM Rev., 33 (1991), pp. 519–625.

[2] P. Blomgren, G. Papanicolaou, and H. Zhao, *Super-resolution in time-reversal acoustics*, J. Acoust. Soc. Amer., 111 (2002), pp. 230–248.

[3] J.-F. Clouet and J.-P. Fouque, *A time reversal-method for an acoustical pulse propagating in randomly layered media*, Wave Motion, 25 (1997), pp. 361–368.

[4] A. Derode, P. Roux, and M. Fink, *Robust acoustic time reversal with high-order multiple scattering*, Phys. Rev. Lett., 75 (1995), pp. 4206–4209.

[5] L. B. Dozier and F. D. Tappert, *Statistics of normal mode amplitudes in a random ocean. I. Theory*, J. Acoust. Soc. Amer., 63 (1978), pp. 353–365.

[6] L. B. Dozier and F. D. Tappert, *Statistics of normal mode amplitudes in a random ocean. II. Computations*, J. Acoust. Soc. Amer., 64 (1978), pp. 533–547.

[7] M. Fink, *Time reversed acoustics*, Physics Today, 20 (1997), pp. 34–40.

[8] M. Fink, D. Cassereau, A. Derode, C. Prada, P. Roux, M. Tanter, J.-L. Thomas, and F. Wu, *Time-reversed acoustics*, Reports on Progress in Physics, 63 (2000), pp. 1933–1995.

[9] J.-P. Fouque, J. Garnier, A. Nachbin, and K. Sølna, *Time reversal refocusing for point source in randomly layered media*, Wave Motion, 42 (2005), pp. 238–260.

[10] J.-P. Fouque, J. Garnier, G. Papanicolaou, and K. Sølna, *Wave Propagation and Time Reversal in Randomly Layered Media*, Springer, New York, 2007.

[11] J. Garnier, *Energy distribution of the quantum harmonic oscillator under random time-dependent perturbations*, Phys. Rev. E, 60 (1999), pp. 3676–3687.

[12] J. Garnier, *The role of evanescent modes in randomly perturbed single-mode waveguides*, Discrete Contin. Dyn. Syst. Ser. B, 8 (2007), pp. 455–472.

[13] W. Kohler, *Power reflection at the input of a randomly perturbed rectangular waveguide*, SIAM J. Appl. Math., 32 (1977), pp. 521–533.

[14] W. Kohler and G. Papanicolaou, *Wave propagation in a randomly inhomogeneous ocean*, in Wave Propagation and Underwater Acoustics, Lecture Notes in Phys. 70, J. B. Keller and J. S. Papadakis, eds., Springer, Berlin, 1977, pp. 153–223.

[15] W. A. Kuperman, W. S. Hodgkiss, H. C. Song, T. Akal, C. Ferla, and D. R. Jackson, *Phase conjugation in the ocean: Experimental demonstration of an acoustic time-reversal mirror*, J. Acoust. Soc. Amer., 103 (1998), pp. 25–40.

[16] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes*, MIT Press, Cambridge, MA, 1984.

[17] D. Marcuse, *Theory of Dielectric Optical Waveguides*, Academic Press, New York, 1974.

[18] A. Nachbin and G. Papanicolaou, *Water waves in shallow channels of rapidly varying depth*, J. Fluid Mech., 241 (1992), pp. 311–332.

[19] G. Papanicolaou and W. Kohler, *Asymptotic theory of mixing stochastic ordinary differential equations*, Comm. Pure Appl. Math., 27 (1974), pp. 641–668.

[20] G. Papanicolaou and W. Kohler, *Asymptotic analysis of deterministic and stochastic equations with rapidly varying components*, Comm. Math. Phys., 45 (1975), pp. 217–232.

[21] G. Papanicolaou, L. Ryzhik, and K. Sølna, *Statistical stability in time reversal*, SIAM J. Appl. Math., 64 (2004), pp. 1133–1155.

[22] P. Roux and M. Fink, *Time reversal in a waveguide: Study of the temporal and spatial focusing*, J. Acoust. Soc. Amer., 107 (2000), pp. 2418–2429.

[23] H. C. Song, W. A. Kuperman, and W. S. Hodgkiss, *Iterative time reversal in the ocean*, J. Acoust. Soc. Amer., 105 (1999), pp. 3176–3184.

SIAM J. APPL. MATH.
Vol. 67, No. 6, pp. 1740–1775

THE DYNAMICS AND INTERACTION OF QUANTIZED VORTICES IN THE GINZBURG–LANDAU–SCHRÖDINGER EQUATION*

YANZHI ZHANG†, WEIZHU BAO‡, AND QIANG DU§

Abstract. The dynamic laws of quantized vortex interactions in the Ginzburg–Landau–Schrödinger equation (GLSE) are analytically and numerically studied. A review of the reduced dynamic laws governing the motion of vortex centers in the GLSE is provided. The reduced dynamic laws are solved analytically for some special initial data. By directly simulating the GLSE with an efficient and accurate numerical method proposed recently in [Y. Zhang, W. Bao, and Q. Du, *Numerical simulation of vortex dynamics in Ginzburg–Landau–Schrödinger equation*, European J. Appl. Math., to appear], we can qualitatively and quantitatively compare quantized vortex interaction patterns of the GLSE with those from the reduced dynamic laws. Some conclusive findings are obtained, and discussions on numerical and theoretical results are made to provide further understanding of vortex interactions in the GLSE. Finally, the vortex motion under an inhomogeneous potential in the GLSE is also studied.

Key words. Ginzburg–Landau equation, nonlinear Schrödinger equation, complex Ginzburg–Landau equation, Ginzburg–Landau–Schrödinger equation, vortex state, reduced dynamic laws, vortex interaction

AMS subject classifications. 35Q55, 65T99, 65Z05, 65N12, 65N35, 81-08

DOI. 10.1137/060671528

1. Introduction.

One of the most well-studied equations in nonlinear science is the Ginzburg–Landau–Schrödinger equation (GLSE) of the form [36]

$$(\alpha - i\beta)\partial_t \psi(\mathbf{x}, t) = \nabla^2 \psi + \frac{1}{\varepsilon^2}\left(V(\mathbf{x}) - |\psi|^2\right)\psi, \qquad \mathbf{x} \in \mathbb{R}^2, \quad t > 0, \tag{1.1}$$

$$\psi(\mathbf{x}, 0) = \psi_0(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^2. \tag{1.2}$$

Here, t is time, $\mathbf{x} = (x, y)^T \in \mathbb{R}^2$ is the Cartesian coordinate vector, (r, θ) is the polar coordinate system, $\psi = \psi(\mathbf{x}, t)$ is a complex-valued wave function (or order parameter), $V(\mathbf{x})$ is a real-valued external potential satisfying $\lim_{|\mathbf{x}|\to\infty} V(\mathbf{x}) = 1$, $\varepsilon > 0$ is a constant, and α and β are two nonnegative constants satisfying $\alpha + \beta > 0$. A vortex-like solution satisfies a nonzero far-field condition as follows: For a given integer $m \in \mathbb{Z}$,

$$|\psi(\mathbf{x}, t)| \to 1 \text{ (e.g., } \psi \to e^{im\theta}), \quad t \geq 0, \quad \text{when } r = |\mathbf{x}| = \sqrt{x^2 + y^2} \to \infty. \tag{1.3}$$

The GLSE (1.1) describes a large variety of nonlinear phenomena, including nonlinear waves, phase transitions, superconductivity, superfluidity, liquid crystals, and

*Received by the editors October 5, 2006; accepted for publication (in revised form) May 22, 2007; published electronically October 5, 2007. We acknowledge support from National University of Singapore grant R-146-000-083-112, and from NSF-DMS 0409297, NSF-CCF 0430349, and NSF-DMR 0205232.

http://www.siam.org/journals/siap/67-6/67152.html

†Department of Mathematics, National University of Singapore, Singapore 117543 (zhyanzhi@gmail.com).

‡Department of Mathematics and Center for Computational Science and Engineering, National University of Singapore, Singapore 117543 (bao@math.nus.edu.sg, http://www.math.nus.edu.sg/~bao).

§Department of Mathematics, Pennsylvania State University, University Park, PA 16802 (qdu@math.psu.edu, http://www.math.psu.edu/qdu).

strings in the field theory. For example, when $\alpha = 1$ and $\beta = 0$, it collapses into the nonlinear heat equation (NLHE) or the Ginzburg–Landau equation (GLE) [27, 28]. The GLE with a complex order parameter is well known for modeling superconductivity [10, 11, 14, 12, 19], while that with a real order parameter corresponds to the Allen–Cahn equation in phase transition [13]. When $\alpha = 0$ and $\beta = 1$, the GLSE reduces to the nonlinear Schrödinger equation (NLSE) [27, 31, 22] for modeling, for example, superfluidity or Bose–Eistein condensation (BEC). While $\alpha > 0$ and $\beta > 0$, it is the complex Ginzburg–Landau equation (CGLE), or NLSE with a damping term [3], which also arises in the study of the hydrodynamic instability [1].

It is known that there are stationary *vortex solutions* with a single winding number or index $m \in \mathbb{Z}$ of the GLSE (1.1) with $\varepsilon = 1$ and $V(\mathbf{x}) \equiv 1$ [27, 14, 36], which take the form

$$\phi_m(\mathbf{x}) = f_m(r)\, e^{im\theta}, \qquad \mathbf{x} = (r\cos\theta, r\sin\theta)^T \in \mathbb{R}^2, \tag{1.4}$$

where the modulus $f_m(r)$ is a real-valued function satisfying

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{df_m(r)}{dr}\right) - \frac{m^2}{r^2} f_m(r) + \left(1 - f_m^2(r)\right) f_m(r) = 0, \qquad 0 < r < \infty, \tag{1.5}$$

$$f_m(0) = 0, \qquad f_m(r) = 1 \quad \text{when} \quad r \to \infty. \tag{1.6}$$

The modulus as well as the core sizes of such vortex states have been calculated in the literature [27, 36] by numerically solving the boundary value problem (1.5)–(1.6). Numerical and analytical results suggest that the vortex states with winding number $m = \pm 1$ are dynamically stable, and, respectively, $|m| > 1$ dynamically unstable [27, 34, 25, 26, 22, 2, 36] (note that the stability and interaction laws of a quantized vortex in the Gross–Pitaevskii equation for BEC [3, 4, 5] may be very different from that studied here due to the different far-field boundary conditions).

In this paper, we study the GLSE (1.1) with initial conditions containing several, say N, vortices. A precise definition of vortex solutions can be found in [22, 19]. We are mainly concerned with the following initial condition:

$$\psi_0(\mathbf{x}) = \prod_{j=1}^{N} \phi_{m_j}\left(\mathbf{x} - \mathbf{x}_j^0\right) = \prod_{j=1}^{N} \phi_{m_j}\left(x - x_j^0,\, y - y_j^0\right), \qquad \mathbf{x} \in \mathbb{R}^2, \tag{1.7}$$

where N is the total number of vortices and ϕ_{m_j} is the vortex state as defined in (1.4) with winding number $m_j = \pm 1$ (see [36] for their numerical solutions). We may then consider the interaction of N vortices with their initial centers shifted from the origin $(0,0)$ to $\mathbf{x}_j^0 = \left(x_j^0, y_j^0\right)^T$ $(1 \le j \le N)$. Taking $m = \sum_{j=1}^{N} m_j$ in (1.3), we refer to vortices with the same winding numbers as *like vortices* and those with different winding numbers as *opposite vortices*.

When $\varepsilon = 1$ and $V(\mathbf{x}) \equiv 1$ in (1.1), it is known that for N well-separated vortices of winding numbers $m_j = \pm 1$ and locations $\mathbf{x}_j$ $(1 \le j \le N)$, the leading asymptotic expansion of the energy is

$$E \sim \sum_{j=1}^{N} E_j - \pi \sum_{j \ne l} m_j\, m_l \ln\left|\mathbf{x}_l - \mathbf{x}_j\right|, \tag{1.8}$$

where E_j is the self-energy of the vortex at $\mathbf{x}_j$, and the second term corresponds to the well-known Kirchoff–Onsager Hamiltonian. From (1.8), we can obtain the

vortex dynamic laws of the induced motion in the leading order, i.e., the adiabatic approximation [27]. For the GLE, i.e., $\alpha = 1$ and $\beta = 0$ in (1.1), the vortex dynamics satisfies [27, 14, 15, 19]

$$\kappa \mathbf{v}_j(t) := \kappa \frac{d\mathbf{x}_j(t)}{dt} = 2m_j \sum_{l=1,\, l\neq j}^{N} m_l \frac{\mathbf{x}_j(t) - \mathbf{x}_l(t)}{|\mathbf{x}_j(t) - \mathbf{x}_l(t)|^2}, \qquad t \geq 0, \tag{1.9}$$

$$\mathbf{x}_j(0) = \mathbf{x}_j^0, \qquad 1 \leq j \leq N, \tag{1.10}$$

where κ is a constant determined from the initial setup (1.7). On the other hand, for the NLSE, i.e., $\alpha = 0$ and $\beta = 1$ in (1.1), it satisfies [27, 14, 8, 19]

$$\mathbf{v}_j(t) := \frac{d\mathbf{x}_j(t)}{dt} = 2 \sum_{l=1,\, l\neq j}^{N} m_l \frac{\mathbf{J}\,(\mathbf{x}_j(t) - \mathbf{x}_l(t))}{|\mathbf{x}_j(t) - \mathbf{x}_l(t)|^2}, \qquad t \geq 0, \tag{1.11}$$

$$\mathbf{x}_j(0) = \mathbf{x}_j^0, \qquad 1 \leq j \leq N, \tag{1.12}$$

where $\mathbf{J}$ is a symplectic matrix defined as

$$\mathbf{J} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{1.13}$$

For asymptotic study of the vortex motions in the GLE and the NLSE, we refer to [7, 8, 9, 6, 20, 18, 29, 31, 32, 33, 35] and references therein.

The aim of this paper is to provide a more detailed and accurate account of the vortex dynamics governed by the GLSE, in particular, to address some open questions concerning the range of validity of the reduced dynamic laws. Our approach is to first solve analytically the ordinary differential equations (ODEs) (1.9) and (1.11) for any N under a few types of initial data, and then compare these solutions with those from direct simulation results of the GLSE (1.1) by using the efficient and accurate numerical method proposed recently in [36]. The key features of the numerical method include (i) the application of a time-splitting technique for decoupling the nonlinearity in the GLSE; (ii) the adoption of polar coordinates to effectively match and resolve the nonzero far-field conditions (1.3) in phase space; and (iii) the utilization of Fourier pseudospectral discretization in the transverse direction and a second order (or fourth order) finite difference or (finite element) discretization in the radial direction [36].

There are naturally many interesting issues concerning the vortex dynamics in various limiting cases, such as the interaction of well-separated vortices with smaller and smaller vortex cores ($\varepsilon \to 0$), and when the distances between the vortices become comparable with the core sizes (initially, both ε and the distances are of $O(1)$). Our approach and numerical methods are applicable to both of these situations, but due to page limitation, our main focus here is on the latter and we leave the discussion on the former to future studies. The main findings in this paper provide justification of the asymptotic vortex dynamic laws in some situations while unveiling limitations in other cases; they also reveal interesting phenomena on the sound wave propagation and the radiation effect associated with the vortex interaction.

The results of the paper are organized as follows. In section 2, based on the nonlinear ODEs of the reduced dynamic laws, we prove the conservation of the mass center and signed mass center of the N vortex centers, respectively, and solve analytically the reduced dynamic laws with a few types of initial data. In section 3, the dynamics and interaction of quantized vortices in the GLE are directly simulated by

solving (1.1) and compared with those from the reduced dynamic laws. Similar results for the NLSE are reported in section 4. The vortex motions in the CGLE, and in the GLSE under an inhomogeneous external potential, are reported in section 5. Finally, some conclusions are drawn in section 6.

2. The reduced dynamic laws. In this section, we first prove the conservation of the mass center and signed mass center of the N vortex centers in the reduced dynamic laws (1.9) and (1.11) for the GLE and the NLSE, respectively. These conservation properties can be used to solve the dynamic laws in special cases and to compare with the direct numerical simulation results of the GLE and the NLSE. We then solve the nonlinear ODEs analytically for several special types of initial data; such analytical solutions can again be compared with the numerical solutions of the GLE and the NLSE.

2.1. Conservation laws. Define, respectively, the mass center $\overline{\mathbf{x}}$ and the signed mass center $\widetilde{\mathbf{x}}$ of the N vortices as

$$\overline{\mathbf{x}}(t) := \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j(t) \quad \text{and} \quad \widetilde{\mathbf{x}}(t) := \frac{1}{N}\sum_{j=1}^{N} m_j\,\mathbf{x}_j(t). \tag{2.1}$$

Let

$$Z_j = \sum_{l=1,\, l\neq j}^{N} m_l \frac{\mathbf{x}_j(t)-\mathbf{x}_l(t)}{|\mathbf{x}_j(t)-\mathbf{x}_l(t)|^2}.$$

It is easy to see that

$$\sum_{j=1}^{N} m_j Z_j = \sum_{j=1}^{N-1}\sum_{j<1\le N}^{N} m_j\, m_l \left[\frac{\mathbf{x}_j(t)-\mathbf{x}_l(t)}{|\mathbf{x}_j(t)-\mathbf{x}_l(t)|^2} + \frac{\mathbf{x}_l(t)-\mathbf{x}_j(t)}{|\mathbf{x}_l(t)-\mathbf{x}_j(t)|^2}\right] = 0. \tag{2.2}$$

Then we have the following.

LEMMA 2.1. *The mass center of the N vortices in the reduced dynamic laws* (1.9) *for the GLE is conserved, i.e.,*

$$\overline{\mathbf{x}}(t) := \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j(t) \equiv \overline{\mathbf{x}}(0) := \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j(0) = \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j^0, \qquad t \ge 0. \tag{2.3}$$

Proof. Summing (1.9) for $1 \le j \le N$ and noting (2.1) and (2.2), we get for $t \ge 0$,

$$\frac{d\overline{\mathbf{x}}(t)}{dt} = \frac{1}{N}\sum_{j=1}^{N}\frac{d\mathbf{x}_j(t)}{dt} = \frac{2}{\kappa N}\sum_{j=1}^{N} m_j Z_j = 0.$$

Thus the conservation law (2.3) is a combination of the above and (1.7). ∎

Similarly, we have the following.

LEMMA 2.2. *The signed mass center of the N vortices in the reduced dynamic laws* (1.11) *for the NLSE is conserved, i.e.,*

$$\widetilde{\mathbf{x}}(t) := \frac{1}{N}\sum_{j=1}^{N} m_j\,\mathbf{x}_j(t) \equiv \widetilde{\mathbf{x}}(0) := \frac{1}{N}\sum_{j=1}^{N} m_j\,\mathbf{x}_j(0) = \frac{1}{N}\sum_{j=1}^{N} m_j\,\mathbf{x}_j^0, \quad t \ge 0. \tag{2.4}$$

Proof. Multiplying (1.11) by $m_j N \mathbf{J}^{-1}$, summing (1.11) for $1 \le j \le N$, and noting (2.1) and (2.2), we have for $t \ge 0$,

$$N\mathbf{J}^{-1}\frac{d\widetilde{\mathbf{x}}(t)}{dt} = \sum_{j=1}^{N} m_j \mathbf{J}^{-1}\frac{d\mathbf{x}_j(t)}{dt} = 2\sum_{j=1}^{N} m_j Z_j = 0.$$

Thus the conservation law (2.4) is a combination of the above and (1.7). ∎

2.2. Initial conditions used for the study of vortex dynamics. Due to the special structures of the nonlinear ODEs (1.9) and (1.11), we can solve them analytically when the N vortices are initially located symmetrically on a circle or at its center. By the conservation of the mass center and signed mass center in (1.9) and (1.11), we assume without loss of generality that the circle is centered at the origin with radius $r_0 = a > 0$.

For simplicity, we denote θ_0 as a given constant, denote $m_0 = +1$ or -1, and consider the following five patterns for the initial conditions in (1.7).

Pattern I. N ($N \ge 2$) like vortices uniformly located on a circle, i.e.,

$$\mathbf{x}_j^0 = a\left(\cos\left(\frac{2j\pi}{N}+\theta_0\right), \sin\left(\frac{2j\pi}{N}+\theta_0\right)\right)^T, \quad \text{and } m_j = m_0 \quad \text{for } 1 \le j \le N. \tag{2.5}$$

Pattern II. N ($N \ge 3$) like vortices located on a circle and its center, i.e.,

$$\mathbf{x}_N^0 = (0,0)^T, \qquad m_{_N} = m_0, \tag{2.6}$$

and for $1 \le j \le N-1$,

$$\mathbf{x}_j^0 = a\left(\cos\left(\frac{2j\pi}{N-1}+\theta_0\right), \sin\left(\frac{2j\pi}{N-1}+\theta_0\right)\right)^T \quad \text{with} \quad m_j = m_0. \tag{2.7}$$

Pattern III. The same as Pattern II, except $m_{_N} = -m_0$ for the center vortex.

Pattern IV. Two opposite vortices, i.e., for $j = 1, 2$,

$$\mathbf{x}_j^0 = a\left(\cos\left(j\pi+\theta_0\right), \sin\left(j\pi+\theta_0\right)\right)^T \quad \text{with} \quad m_1 = -m_2 = m_0. \tag{2.8}$$

Pattern V. Three vortices ($N = 3$) with nonsymmetric initial setups.

Here we consider the following three different cases (with $m_1 = m_3 = +1$):

Case 1. $\mathbf{x}_1^0 = (-a, -b/2)^T$, $\mathbf{x}_2^0 = (0, b)^T$, $\mathbf{x}_3^0 = (a, -b/2)^T$, $m_2 = +1$.

Case 2. $\mathbf{x}_1^0 = (-\sqrt{3}a/2, -a/2)^T$, $\mathbf{x}_2^0 = (0, a)^T$, $\mathbf{x}_3^0 = (\sqrt{3}a/2, -a/2)^T$, $m_2 = -1$.

Case 3. $\mathbf{x}_1^0 = (-a, -b/2)^T$, $\mathbf{x}_2^0 = (0, b)^T$, $\mathbf{x}_3^0 = (a, -b/2)^T$, $m_2 = -1$.

Notice that for all five types of patterns, we have $\overline{\mathbf{x}}(t) = \overline{\mathbf{x}}(0) = (0,0)^T$ for $t \ge 0$. Moreover, for the first three patterns and the first case of Pattern V, we have $\widetilde{\mathbf{x}}(t) = \widetilde{\mathbf{x}}(0) = 0$.

2.3. Analytical solutions of the reduced dynamics for the GLE. Noting (2.3), we can solve the nonlinear ODEs (1.9) analytically when the initial conditions in (1.10) are given by Patterns I–IV.

LEMMA 2.3. *If the initial data in* (1.10) *satisfy* (2.5)*, i.e., Pattern* I*, then the solutions of* (1.9)–(1.10) *can be given, for* $1 \le j \le N$ *with* $N \ge 2$*, by*

$$\mathbf{x}_j(t) = \sqrt{a^2 + \frac{2\,(N-1)}{\kappa}\,t}\,\left(\cos\left(\frac{2j\pi}{N}+\theta_0\right), \sin\left(\frac{2j\pi}{N}+\theta_0\right)\right)^T, \quad t \ge 0. \tag{2.9}$$

Proof. For any $N \geq 2$, based on the structures of the ODEs (1.9) and the initial data (2.5), we take the ansatz for the solution as

$$\mathbf{x}_j(t) = c_{_N}(t)\mathbf{x}_j^0, \qquad t \geq 0, \quad 1 \leq j \leq N, \tag{2.10}$$

where $c_{_N}(t)$ is a function of time t and $c_{_N}(0) = 1$. Substituting (2.10) into (1.9), applying a dot-product on both sides by $\mathbf{x}_j^0$, and noting (2.5), we get

$$\begin{aligned} c_{_N}'(t) &= \frac{2}{\kappa a^2 c_{_N}(t)} \sum_{l=1,\, l\neq j}^{N} m_j m_l \frac{(\mathbf{x}_j^0 - \mathbf{x}_l^0)\cdot \mathbf{x}_j^0}{|\mathbf{x}_j^0 - \mathbf{x}_l^0|^2} \\ &= \frac{2}{\kappa a^2 c_{_N}(t)} \sum_{l=1,\, l\neq j}^{N} \frac{a^2 - \mathbf{x}_l^0 \cdot \mathbf{x}_j^0}{2a^2 - 2\mathbf{x}_l^0 \cdot \mathbf{x}_j^0} = \frac{N-1}{\kappa a^2 c_{_N}(t)}, \quad t \geq 0. \end{aligned}$$

Solving the above ODE and noting that $c_{_N}(0) = 1$, we obtain

$$c_{_N}(t) = \sqrt{1 + \frac{2(N-1)}{a^2\kappa}\, t}, \qquad t \geq 0. \tag{2.11}$$

Thus the solution (2.9) is a combination of (2.10), (2.11), and (2.5). ∎

From the results in Lemma 2.3 we can see that, when the N vortices are uniformly located on a circle initially, i.e., as in Pattern I, by the reduced dynamic law each vortex moves outside along the line passing through its initial location and the origin, and these N vortices are located on a circle at any time t with its radius increasing by time $c_{_N}(t)$ as in (2.11).

LEMMA 2.4. *If the initial data in* (1.10) *satisfy* (2.6)–(2.7)*, i.e., Pattern* II*, then the solutions of* (1.9)–(1.10) *are*

$$\mathbf{x}_N(t) \equiv (0,0)^T, \qquad t \geq 0, \tag{2.12}$$

and for $1 \leq j \leq N-1$ *with* $N \geq 3$*,*

$$\mathbf{x}_j(t) = \sqrt{a^2 + \frac{2N}{\kappa}\, t}\, \left(\cos\left(\frac{2j\pi}{N-1} + \theta_0\right), \sin\left(\frac{2j\pi}{N-1} + \theta_0\right)\right)^T, \quad t \geq 0. \tag{2.13}$$

Proof. Due to the symmetry of the ODEs (1.9), the initial data (2.5), and the conservation of mass center (2.3), we can immediately obtain the solution (2.12). As in the proof of Lemma 2.3, we assume

$$\mathbf{x}_j(t) = d_N(t)\mathbf{x}_j^0, \qquad t \geq 0, \quad 1 \leq j \leq N-1,$$

where $d_N(t)$ is a function of time t and $d_N(0) = 1$. Substituting the above into (1.9), applying a dot-product on both sides by $\mathbf{x}_j^0$, and noting (2.7) and (2.12), we get

$$\begin{aligned} d_N'(t) &= \frac{2}{\kappa a^2 d_N(t)} \left[m_j m_{_N} \frac{(\mathbf{x}_j^0 - \mathbf{x}_N^0)\cdot \mathbf{x}_j^0}{|\mathbf{x}_j^0 - \mathbf{x}_N^0|^2} + \sum_{l=1,\, l\neq j}^{N-1} m_j m_l \frac{(\mathbf{x}_j^0 - \mathbf{x}_l^0)\cdot \mathbf{x}_j^0}{|\mathbf{x}_j^0 - \mathbf{x}_l^0|^2} \right] \\ &= \frac{2}{\kappa a^2 d_N(t)} \left[m_0^2 + \sum_{l=1,\, l\neq j}^{N-1} m_0^2 \frac{a^2 - \mathbf{x}_l^0 \cdot \mathbf{x}_j^0}{2a^2 - 2\mathbf{x}_l^0 \cdot \mathbf{x}_j^0} \right] = \frac{N}{\kappa a^2 d_N(t)}, \quad t \geq 0. \end{aligned}$$

Solving the above ODE and noting that $d_N(0) = 1$, we obtain

$$d_N(t) = \sqrt{1 + \frac{2N}{a^2\kappa}\, t}, \qquad t \geq 0. \tag{2.14}$$

Thus the solution (2.13) is a combination of the above and (2.7). ∎

From the results in Lemma 2.4 we can see that, for the dynamics of (1.9)–(1.10) in Pattern II, by the reduced dynamic law the vortex initially at the center of the circle does not move for any time $t \geq 0$, each of the other $N-1$ vortices moves outside along the line passing through its initial location and the origin, and these $N-1$ vortices are located on a circle at any time t with its radius increasing by time $d_N(t)$ as in (2.14).

LEMMA 2.5. *If the initial data in* (1.10) *are as in Pattern* III*, then the solutions of* (1.9)–(1.10) *are*

$$\mathbf{x}_N(t) \equiv (0,0)^T, \qquad t \geq 0, \tag{2.15}$$

and for $1 \leq j \leq N-1$ *with* $N \geq 3$*,*

$$\mathbf{x}_j(t) = \sqrt{a^2 + \frac{2(N-4)}{\kappa}\, t}\, \left(\cos\left(\frac{2j\pi}{N-1} + \theta_0\right), \sin\left(\frac{2j\pi}{N-1} + \theta_0\right)\right)^T. \tag{2.16}$$

The proof follows from the analogous results in Lemma 2.4. From the results in Lemma 2.5 we can see that, for the dynamics of (1.9)–(1.10) in Pattern III, by the reduced dynamic law (i) the vortex initially at the origin does not move during the interaction, each of the other $N-1$ vortices moves along the line passing through its initial location and the origin, and these $N-1$ vortices are located on a circle at any time t; (ii) when $N = 3$, the two vortices with the same index move towards each other and collide with the vortex having the opposite index at the origin and at time $t = t_c = \kappa a^2/2$; (iii) when $N = 4$, all four vortices do not move but remain at their initial locations for any $t \geq 0$; and (iv) when $N \geq 5$, the $N-1$ vortices with the same index move outside and never collide with the vortex with the opposite index no matter how small the initial radius of the circle is.

LEMMA 2.6. *If the initial data in* (1.10) *satisfy* (2.8)*, i.e., Pattern* IV*, then the solutions of* (1.9)–(1.10) *can be given by*

$$\mathbf{x}_j(t) = \sqrt{a^2 - \frac{2}{\kappa}\, t}\, \left(\cos\left(j\pi + \theta_0\right), \sin\left(j\pi + \theta_0\right)\right)^T, \quad 0 \leq t \leq t_c, \quad j = 1, 2, \tag{2.17}$$

with $t_c = \kappa a^2/2$.

The proof is similar to that of Lemma 2.3. From the results in Lemma 2.6 we can see that, for the dynamics of (1.9)–(1.10) in Pattern IV, when $0 \leq t < t_c = a^2\kappa/2$ the two vortices move towards each other along a line passing through their initial locations and collide at the origin at time $t = t_c = O(a^2)$ according to the reduced dynamic law.

2.4. Analytical solutions of the reduced dynamics for the NLSE. Similarly, noting (2.4) we can also solve the nonlinear ODEs (1.11) analytically when the initial conditions in (1.12) are given by Patterns I–IV.

LEMMA 2.7. *If the initial data in* (1.12) *satisfy* (2.5)*, i.e., Pattern* I*, then the solutions of* (1.11)–(1.12) *can be given, for* $1 \leq j \leq N$ *with* $N \geq 2$*, by*

$$\mathbf{x}_j(t) = a\left(\cos\left(\frac{2j\pi}{N} + \theta_0 + \frac{m_0(N-1)}{a^2}\, t\right), \sin\left(\frac{2j\pi}{N} + \theta_0 + \frac{m_0(N-1)}{a^2}\, t\right)\right)^T. \tag{2.18}$$

Proof. For any $N \geq 2$, based on the structures of the ODEs (1.11) and the initial data (2.5), we take the ansatz for the solution with $1 \leq j \leq N$ as

$$\mathbf{x}_j(t) = a\left(\cos\left(\frac{2j\pi}{N} + \theta_0 + \alpha_N(t)\right), \sin\left(\frac{2j\pi}{N} + \theta_0 + \alpha_N(t)\right)\right)^T, \quad t \geq 0, \tag{2.19}$$

where $\alpha_N(t)$ is a function of time and $\alpha_N(0) = 0$.

Now, let $\mathbf{x}_j^{\perp}(t) = a\left(-\sin\left(\frac{2j\pi}{N} + \theta_0 + \alpha_N(t)\right), \cos\left(\frac{2j\pi}{N} + \theta_0 + \alpha_N(t)\right)\right)^T$. By (1.13), we have the elementary identity

$$\begin{aligned} &\sum_{l=1,\, l\neq j}^{N} m_0 \frac{\mathbf{x}_j^{\perp} \cdot (\mathbf{J}\mathbf{x}_j) - \mathbf{x}_j^{\perp} \cdot (\mathbf{J}\mathbf{x}_l)}{|\mathbf{x}_j|^2 + |\mathbf{x}_l|^2 - 2\mathbf{x}_j \cdot \mathbf{x}_l} \\ &= \sum_{l=1,\, l\neq j}^{N} m_0 \frac{1 - \cos\left(\frac{2(j-l)\pi}{N}\right)}{2 - 2\cos\left(\frac{2(j-l)\pi}{N}\right)} = \frac{m_0(N-1)}{2}. \end{aligned} \tag{2.20}$$

Inserting (2.19) into (1.11) and applying a dot-product on both sides with $\mathbf{x}_j^{\perp}(t)$, we get

$$\alpha_N'(t) = \frac{2}{a^2} \sum_{l=1,\, l\neq j}^{N} m_l \frac{\mathbf{x}_j^{\perp} \cdot [\mathbf{J}(\mathbf{x}_j - \mathbf{x}_l)]}{|\mathbf{x}_j - \mathbf{x}_l|^2} = \frac{m_0(N-1)}{a^2}, \qquad t \geq 0.$$

Solving the above ODE and noting that $\alpha_N(0) = 0$, we obtain $\alpha_N(t) = m_0(N-1)t/a^2$ for $t \geq 0$. Thus a combination of the above leads to the solution (2.18). ∎

From the results in Lemma 2.7 we can see that, when the N like vortices are uniformly located on a circle initially, i.e., for the dynamics of (1.11)–(1.12) in Pattern I, they rotate along the circle (counterclockwise if $m_0 = +1$ and clockwise if $m_0 = -1$) with angular frequency $\omega = (N-1)/a^2$ (cf. Figure 13(a),(d)).

LEMMA 2.8. *If the initial data in* (1.12) *satisfy* (2.6)–(2.7)*, i.e., Pattern* II*, then the solutions of* (1.11)–(1.12) *are*

$$\mathbf{x}_N(t) \equiv (0,0)^T, \qquad t \geq 0, \tag{2.21}$$

and for $1 \leq j \leq N-1$ *with* $N \geq 3$*,*

$$\mathbf{x}_j(t) = a\left(\cos\left(\frac{2j\pi}{N-1} + \theta_0 + \frac{m_0 N}{a^2}\, t\right), \sin\left(\frac{2j\pi}{N-1} + \theta_0 + \frac{m_0 N}{a^2}\, t\right)\right)^T. \tag{2.22}$$

Proof. Due to the symmetry of the ODEs (1.11), the initial data (2.6)–(2.7), and the conservation of signed mass center (2.4), we can immediately get the solution (2.21). As in the proof of Lemma 2.7, we assume for $1 \leq j \leq N-1$ that

$$\mathbf{x}_j(t) = a\left(\cos\left(\frac{2j\pi}{N-1} + \theta_0 + \beta_N(t)\right), \sin\left(\frac{2j\pi}{N-1} + \theta_0 + \beta_N(t)\right)\right)^T, \tag{2.23}$$

where $\beta_N(t)$ is a function of time and $\beta_N(0) = 0$. Inserting (2.23) into (1.11), applying a dot-product on both sides with

$$\mathbf{x}_j^{\perp}(t) = a\left(-\sin\left(\frac{2j\pi}{N-1} + \theta_0 + \beta_N(t)\right), \cos\left(\frac{2j\pi}{N-1} + \theta_0 + \beta_N(t)\right)\right)^T,$$

and noting (1.13), (2.21), and (2.20) (with N replaced by $N-1$), we get

$$\begin{aligned}\beta_N'(t) &= \frac{2}{a^2}\left[m_{_N}\frac{\mathbf{x}_j^{\perp}\cdot[\mathbf{J}(\mathbf{x}_j-\mathbf{x}_N)]}{|\mathbf{x}_j-\mathbf{x}_N|^2}+\sum_{l=1,\,l\neq j}^{N-1} m_l\frac{\mathbf{x}_j^{\perp}\cdot[\mathbf{J}(\mathbf{x}_j-\mathbf{x}_l)]}{|\mathbf{x}_j-\mathbf{x}_l|^2}\right]\\ &= \frac{2}{a^2}\left[m_0\frac{\mathbf{x}_j^{\perp}\cdot(\mathbf{J}\mathbf{x}_j)}{|\mathbf{x}_j|^2}+\sum_{l=1,\,l\neq j}^{N-1} m_0\frac{\mathbf{x}_j^{\perp}\cdot(\mathbf{J}\mathbf{x}_j)-\mathbf{x}_j^{\perp}\cdot(\mathbf{J}\mathbf{x}_l)}{|\mathbf{x}_j|^2+|\mathbf{x}_l|^2-2\mathbf{x}_j\cdot\mathbf{x}_l}\right]=\frac{m_0N}{a^2}\end{aligned}$$

for $t\geq 0$. Solving the above ODE and noting that $\beta_N(0)=0$, we obtain $\beta_N(t)=m_0Nt/a^2$ for $t\geq 0$. Thus a combination of the above leads to the solution (2.22). ∎

From the results in Lemma 2.8 we can see that, for the dynamics of (1.11)–(1.12) in Pattern II, the vortex initially at the center of the circle does not move for any time $t\geq 0$, and the other $N-1$ vortices rotate along the circle (counterclockwise if $m_0=+1$ and clockwise if $m_0=-1$) with angular frequency $\omega=N/a^2$ (cf. Figure 14(a),(d)).

LEMMA 2.9. *If the initial data in* (1.12) *are as in Pattern* III*, then the solutions of* (1.11)–(1.12) *are*

$$\mathbf{x}_N(t)\equiv(0,0)^T,\qquad t\geq 0, \tag{2.24}$$

and for $1\leq j\leq N-1$ *with* $N\geq 3$*,*

$$\mathbf{x}_j(t)=a\left(\cos\left(\frac{2j\pi}{N-1}+\theta_0+m_0\omega_N t\right),\,\sin\left(\frac{2j\pi}{N-1}+\theta_0+m_0\omega_N t\right)\right)^T, \tag{2.25}$$

where $\omega_N=(N-4)/a^2$.

The proof is similar to that of Lemma 2.8. From the results in Lemma 2.9, we can see for the dynamics of (1.11)–(1.12) in Pattern III that (i) the vortex initially at the origin does not move during the interaction; (ii) when $N=3$, the two vortices initially located on a circle rotate along the same circle (clockwise if $m_0=+1$ and counterclockwise if $m_0=-1$) with frequency $\omega(a)=1/a^2$ (cf. Figure 15(a)); (iii) the case of $N=4$ is rather special, and the reduced dynamics implies that all four vortices do not move and stay at their initial locations for any $t\geq 0$ (cf. Figure 15(d)); and (iv) when $N\geq 5$, the $N-1$ vortices initially located on a circle rotate along the same circle (counterclockwise if $m_0=+1$ and clockwise if $m_0=-1$) with angular frequency $\omega_N=(N-4)/a^2$ (cf. Figure 15(g)).

LEMMA 2.10. *If the initial data in* (1.12) *satisfy* (2.8)*, i.e., Pattern* IV*, then the solutions of* (1.11)–(1.12) *can be given by*

$$\mathbf{x}_j(t)=\mathbf{x}_j^0+\frac{m_0}{a}\,t\,(-\sin\theta_0,\,\cos\theta_0)^T,\qquad t\geq 0,\quad j=1,2. \tag{2.26}$$

Proof. From the conservation of the signed mass center (2.4), we have

$$\widetilde{\mathbf{x}}(t)=\frac{\mathbf{x}_1(t)-\mathbf{x}_2(t)}{2}\equiv\frac{\mathbf{x}_1(0)-\mathbf{x}_2(0)}{2}=a\,(\cos\theta_0,\,\sin\theta_0)^T,\quad t\geq 0. \tag{2.27}$$

On the other hand, from the ODEs (1.11), we obtain

$$\frac{d\mathbf{x}_1(t)}{dt}=-2m_0\frac{\mathbf{J}\,(\mathbf{x}_1(t)-\mathbf{x}_2(t))}{|\mathbf{x}_1(t)-\mathbf{x}_2(t)|^2},\qquad \frac{d\mathbf{x}_2(t)}{dt}=2m_0\frac{\mathbf{J}\,(\mathbf{x}_2(t)-\mathbf{x}_1(t))}{|\mathbf{x}_2(t)-\mathbf{x}_1(t)|^2}.$$

Summing up the above equations and combining with (2.27), we get

$$\frac{d\mathbf{x}_1(t)}{dt} = -\frac{m_0}{a}\mathbf{J}\,(\cos\theta_0,\,\sin\theta_0)^T, \quad t \geq 0, \quad \text{with} \quad \mathbf{x}_1(0) = \mathbf{x}_1^0. \tag{2.28}$$

Solving (2.28) and noting (2.27), we obtain (2.26) immediately. □

From the results in Lemma 2.10 we can see that, for the dynamics of (1.11)–(1.12) in Pattern IV, the two opposite vortices move along two parallel lines which are perpendicular to the line passing through their initial locations with constant velocity (cf. Figures 16(b) and 20(a))

$$\mathbf{v}(t) = \frac{d\mathbf{x}_1(t)}{dt} = \frac{d\mathbf{x}_2(t)}{dt} \equiv \frac{m_0}{a}\,(-\sin\theta_0,\,\cos\theta_0)^T, \qquad t \geq 0. \tag{2.29}$$

3. Numerical results for vortex dynamics in the GLE. In this section, we report the numerical results of the vortex dynamics and interaction by directly simulating the GLE; i.e., we take $\alpha = 1$, $\beta = 0$, $\varepsilon = 1$, and $V(\mathbf{x}) \equiv 1$ in (1.1), with the efficient and accurate time-splitting method introduced in [36]. For the choice of mesh size and time step, as well as the size of the bounded computational domain, we refer to [36]. For comparison, we also exhibit the motion of the vortex centers solved from the reduced dynamics (1.9) in each case. In the figures, the symbols used include + (center of a vortex with index $m = +1$), − (center of a vortex with index $m = -1$), and o (collision position of two or more opposite vortices).

3.1. Interactions of N ($N \geq 2$) like vortices, Patterns I and II. Figure 1 displays the surface plots of $-|\psi|$ at different times when the initial data in (1.7) are chosen as (2.5) with $N = 2$, $m_0 = +1$, and $a = 2$, and Figure 2 shows the time evolution of the vortex centers for different number of vortices $N \geq 2$, i.e., Pattern I. In addition, Figure 3 shows the time evolution of the vortex centers when the initial data in (1.7) is chosen as (2.6)–(2.7) with $m_0 = +1$ and $a = 3$ for different number of vortices $N \geq 3$, i.e., Pattern II.

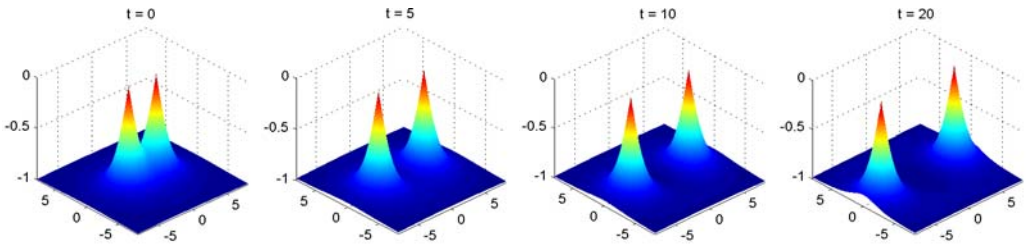


FIG. 1. *Surface plots of* $-|\psi|$ *at different times for the GLE when the initial condition is chosen as Pattern* I *with* $N = 2$, $m_0 = +1$, *and* $a = 2$ *in* (2.5).

From Figures 1–3, and additional numerical experiments not shown here, we can draw the following conclusions for the interaction of N like vortices in the GLE when the initial data are chosen as either Pattern I or II:

(i) The mass center of the vortex centers is conserved for any time $t \geq 0$ (cf. Figures 2 and 3), which confirms the conservation law in (2.3).

(ii) Vortices with the same index undergo a repulsive interaction and they never collide (cf. Figures 1, 2, and 3). Their speeds depend on their distances to the origin, i.e., the larger the distance, the slower the motion (cf. Figures 2 and 3). In addition,

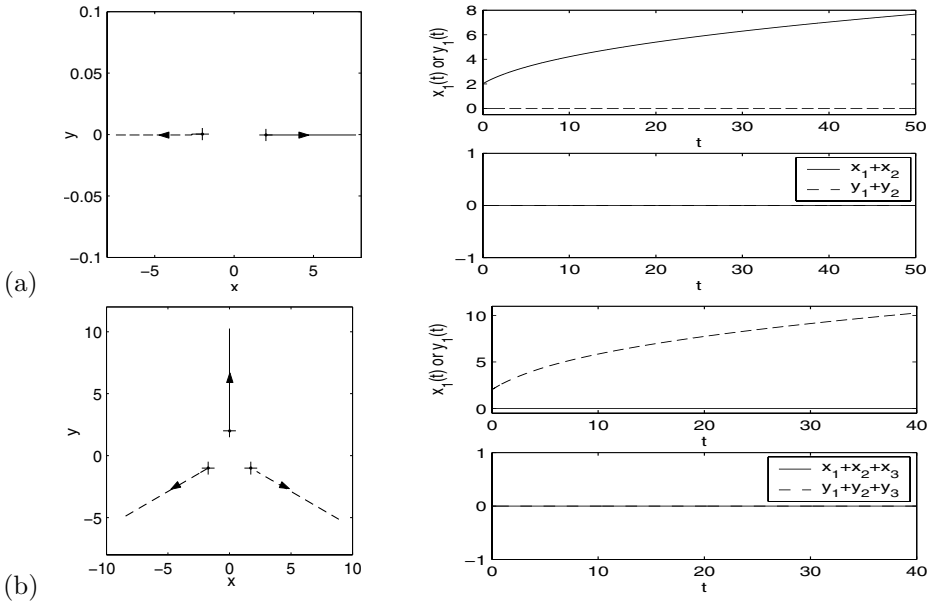


FIG. 2. *Time evolution of vortex centers by directly simulating the GLE when the initial data are chosen as Pattern* I *with* $a = 2$ *and* $m_0 = +1$ *for different* N. (a) $N = 2$; (b) $N = 3$.

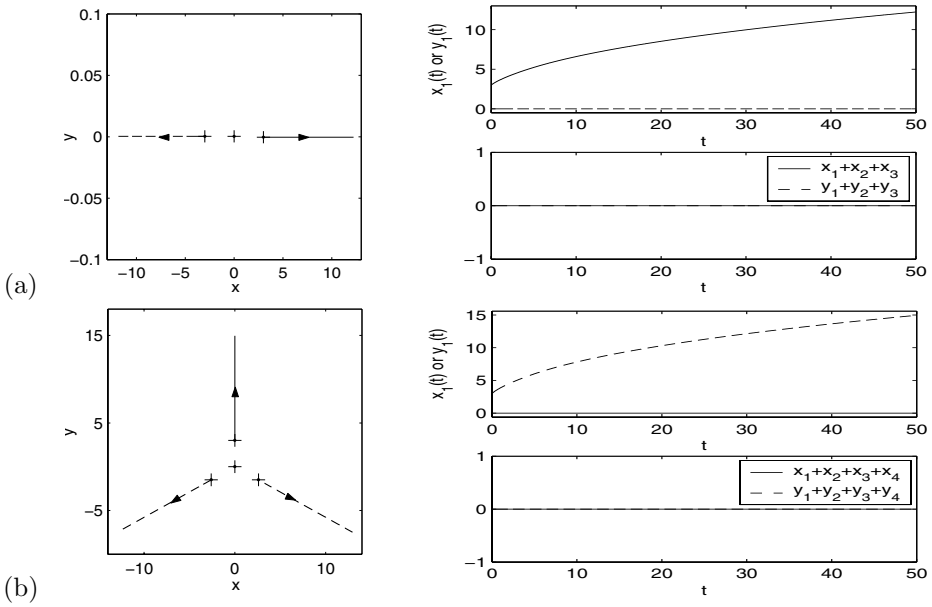


FIG. 3. *Time evolution of vortex centers by directly simulating the GLE when the initial data are chosen as Pattern* II *with* $a = 3$ *and* $m_0 = +1$ *for different* N. (a) $N = 3$; (b) $N = 4$.

in Pattern II the vortex initially at the origin does not move during the dynamics (cf. Figure 3), which confirms the analytical solution (2.12).

(iii) Due to the symmetry of the initial data, the vortices of those initially located on a circle move along lines passing through their initial locations and the origin, and

at any time $t \geq 0$, these vortex centers are always on a circle (cf. Figures 2 and 3), which confirms the analytical solutions (2.9) and (2.13).

(iv) In Patterns I and II, the solutions of the reduced dynamic laws agree qualitatively with our numerical results of the GLE, and quantitatively if a proper κ in (1.9) is chosen, which depends on the initial setup in (1.7). For example, in Pattern I with $N = 2$, we numerically find that the two solutions are the same when we choose $\kappa \approx 2.1279$, 2.1690, 2.2589, and 2.3116 for $a = 4$, 5, 10, and 20, respectively, which suggests that $\frac{1}{\kappa} \approx 0.424 + \frac{0.1897}{a}$ when $a \geq 4$.

3.2. Interactions of N ($N \geq 3$) opposite vortices, Pattern III. Figure 4 shows time evolution of the vortex centers when the initial data in (1.7) are chosen as Pattern III with $m_0 = +1$ and $a = 3$ for different $N \geq 3$.

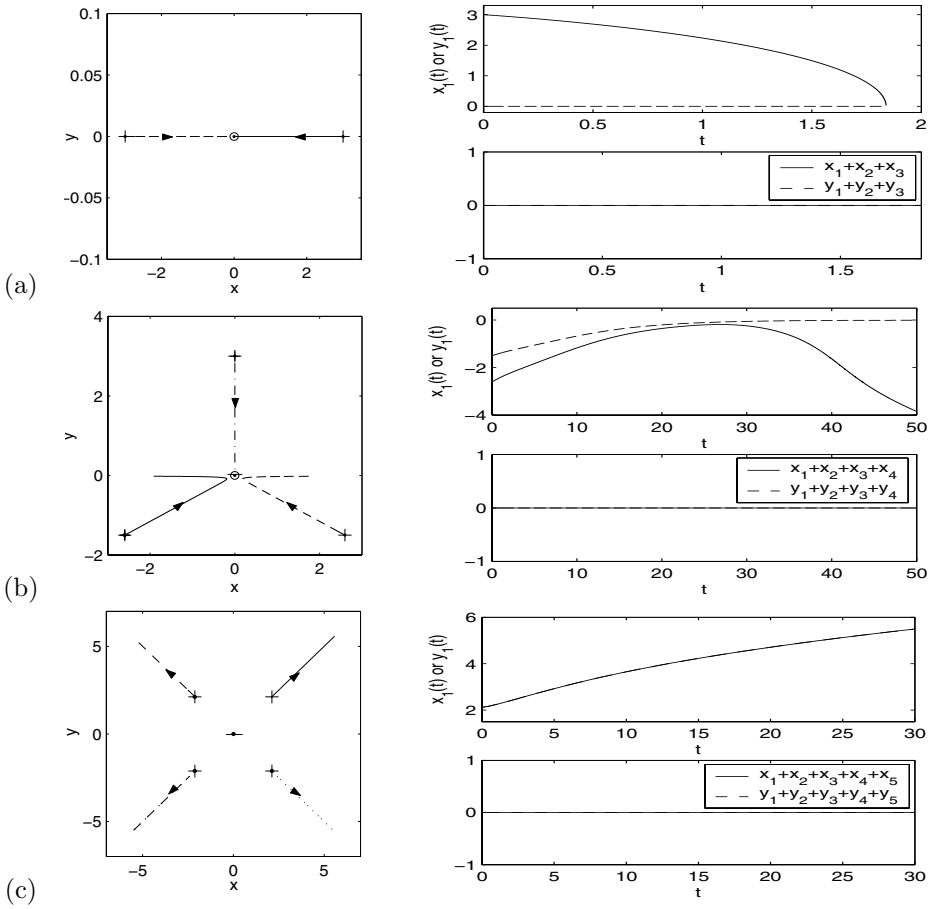


FIG. 4. *Time evolution of vortex centers by directly simulating the GLE when the initial data are chosen as Pattern* III *with* $a = 3$ *and* $m_0 = +1$ *for different* N. (a) $N = 3$; (b) $N = 4$; (c) $N = 5$.

From Figure 4, and additional numerical experiments not shown here, we can draw the following conclusions for the interaction of N opposite vortices in the GLE when the initial data are chosen as Pattern III:

(i) The mass center of the vortex centers is conserved for any time $t \geq 0$ (cf. Figure 4), which confirms the conservation law in (2.3).

(ii) The vortex initially at the origin does not move for any time $t \geq 0$ (cf. Figure 4), which confirms the analytical solution (2.15). The vortices of those initially located on a circle move to the origin when $N = 3$ or 4 and, respectively, move away when $N \geq 5$, along lines passing through their initial location and the origin, and at any time $t \geq 0$, these vortex centers are always on a circle (cf. Figure 4), which confirms the analytical solutions (2.16). Their speeds depend on their distances to the origin, i.e., the larger the distance, the slower the motion.

(iii) When $N = 3$ or 4, collisions between the vortex centers are observed at a critical time t_c (cf. Figure 4(a),(b)). The collision time is quadratically proportional to the initial distance a. Before collision, the interaction is attractive. When $N = 3$, they collide at the origin, and after the collision, there is one vortex with index $m = m_0$ left, and it stays at the origin forever (cf. Figure 4(a)). On the other hand, when $N = 4$, one of the three vortices initially located on the circle collides with the one initially at the origin. After the collision, two like vortices remain and they undergo a repulsive interaction (cf. Figure 4(b)).

(iv) When $N \geq 5$, the vortices undergo repulsive interactions and never collide (cf. Figure 4(c)).

(v) In Pattern III, when $N = 3$ or $N \geq 5$, the solutions of the reduced dynamic laws qualitatively agree with our numerical results of the GLE. On the contrary, they are completely different for $N = 4$. One may argue that a possible cause is the fact that this case corresponds to a degenerate case of the reduced dynamics (1.9)–(1.10) for which the vortices remain stationary, and thus the next order effect becomes important in the underlying vortex motion of the original GLE. In fact, the collision time needed for $N = 4$ ($t_c \approx 28$; cf. Figure 4(b)) is much longer than that for $N = 3$ ($t_c \approx 1.8$; cf. Figure 4(a)) with the same initial radius of the circle at $a = 3$.

3.3. Interactions of two opposite vortices, Pattern IV. Figure 5 displays the surface plots of $-|\psi|$ at different times and Figure 6 shows time evolution of the vortex centers when the initial data in (1.7) are chosen as (2.8) with $m_0 = +1$ and $a = 1.5$, i.e., Pattern IV.

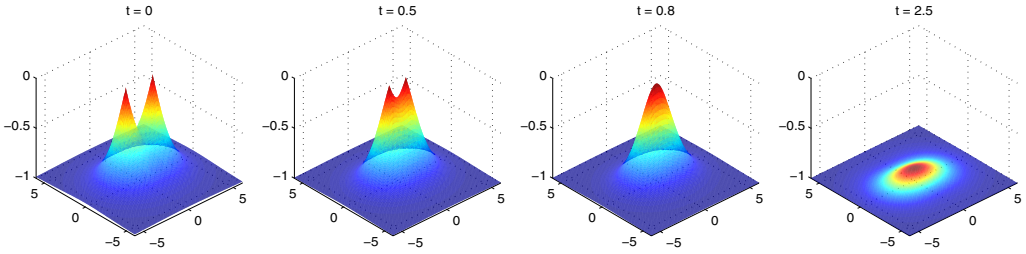


FIG. 5. *Surface plots of* $-|\psi|$ *at different times for the GLE when the initial condition is chosen as Pattern* IV (2.8) *with* $m_0 = +1$ *and* $a = 1.5$.

From Figures 5–6, we can draw the following conclusions for the interaction of two opposite vortices in the GLE when the initial condition is chosen as Pattern IV:

(i) The mass center of the two vortex centers is conserved for any time $t \geq 0$ (cf. Figure 6), which again confirms the conservation law in (2.3).

(ii) Two vortices with opposite winding numbers undergo an attractive interaction (cf. Figure 5), and their centers move along a straight line passing through their locations at $t = 0$ (cf. Figure 6). The speed of the motion for the two vortex

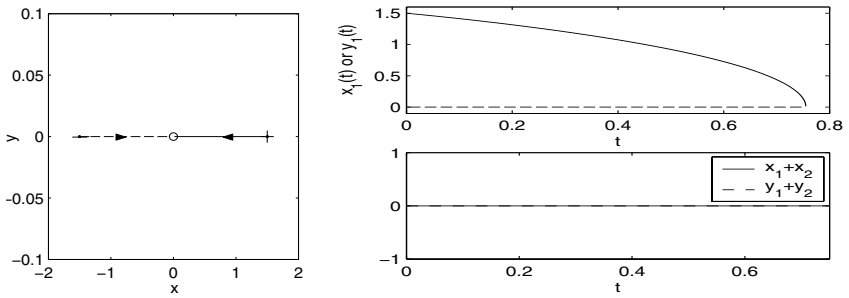


FIG. 6. *Time evolution of vortex centers by directly simulating the GLE when the initial data are chosen as Pattern* IV *with* $a = 1.5$ *and* $m_0 = +1$.

centers depends on their distance. The smaller the distance, the faster the motion (cf. Figure 6).

(iii) There exists a critical time $t_c > 0$, and the two opposite vortices collide with each other at the origin (cf. Figure 5). From our numerical results, we find numerically that the collision time depends on the distance of the two vortex centers at $t = 0$ as

$$t_c \approx \frac{1}{14.8710}\, d_0^{2.0715} \quad \text{with} \quad d_0 = 2a, \qquad a > 0. \tag{3.1}$$

This immediately implies that $t_c = O\left(a^2\right)$, which confirms the analytical result of the collision time in Lemma 2.6.

(iv) Again, in Pattern IV the solutions of the reduced dynamic laws agree qualitatively with our numerical results of the GLE, and quantitatively if a proper κ in (1.9) is chosen, which depends on the initial distance between the two vortex centers.

3.4. Interactions of vortices with nonsymmetric setups. Figures 7–9 show the time evolution of the vortex centers when the initial data in (1.7) are chosen as the three cases in Pattern V, respectively.

Based on Figures 7–9 and our additional numerical experiments, we can draw the following conclusions for three vortices with nonsymmetric initial setups:

(i) When they have the same index, they never collide (cf. Figure 7). On the contrary, when they have opposite indices, they collide at a finite time (cf. Figures 8 and 9), and after collision, only one vortex is left.

(ii) The mass centers of the vortex centers are not conserved (cf. Figures 7–9) during the dynamics within the time frame we computed the solutions, which suggests that there is a considerable discrepancy between the reduced dynamics law (1.9)–(1.10) and the original dynamics in some regimes. One may argue that a possible cause is the fact that the reduced dynamic law is the adiabatic approximation in the leading order when the N vortices are well separated, and thus the next order effect becomes important in the underlying vortex motion of the original GLE when the N vortices are not well separated. In fact, in our numerical results, the larger the distance between the vortex centers, the better the conservation of the mass center (cf. Figures 7–9). This suggests that the reduced dynamics law (1.9)–(1.10) is still a reasonable approximation to the vortex motion of the original GLE in a nonsymmetric initial setup when the N vortices are well separated.

4. Numerical results for vortex dynamics in the NLSE. Similarly, in this section we report the numerical results of the vortex dynamics and interaction by

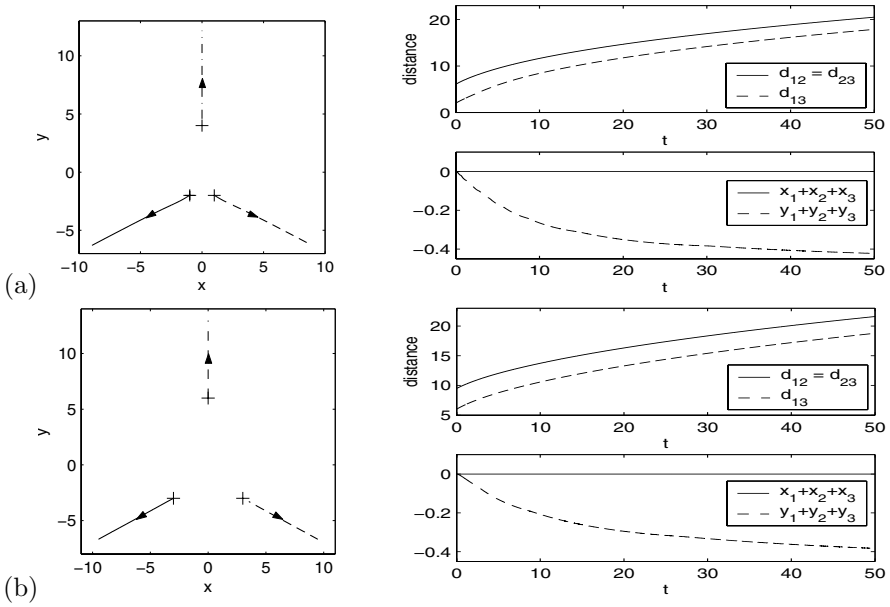


FIG. 7. *Time evolution of vortex centers by directly simulating the GLE when the initial data are chosen as Case* 1 *in Pattern* V *with different* a *and* b. (a) $a = 1$, $b = 4$; (b) $a = 3$, $b = 6$.

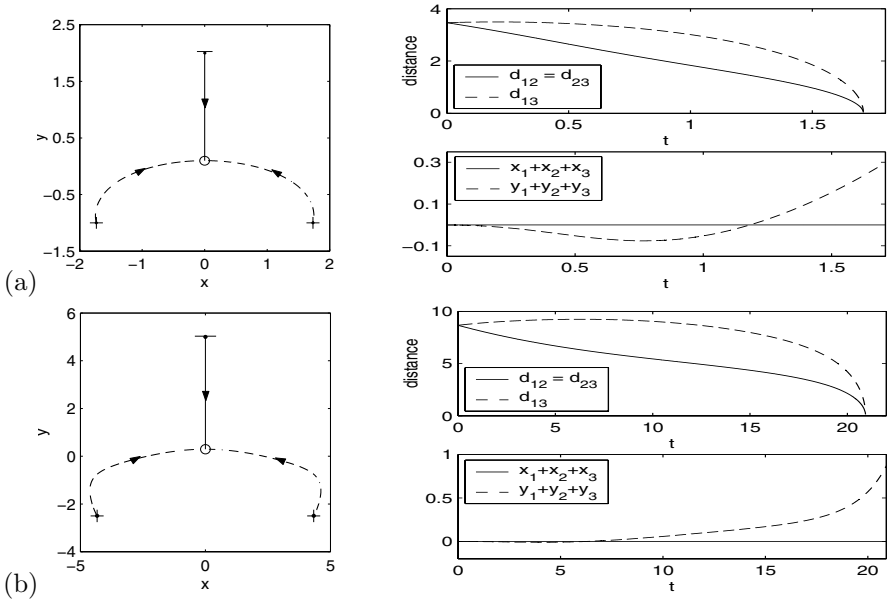


FIG. 8. *Time evolution of vortex centers by directly simulating the GLE when the initial data are chosen as Case* 2 *in Pattern* V *with different* a. (a) $a = 2$; (b) $a = 5$.

directly simulating the nonlinear Schrödinger equation; i.e., we take $\alpha = 0$, $\beta = 1$, $\varepsilon = 1$, and $V(\mathbf{x}) \equiv 1$ in (1.1), with the numerical method introduced in [36]. All the computational setups are the same as in the previous section.

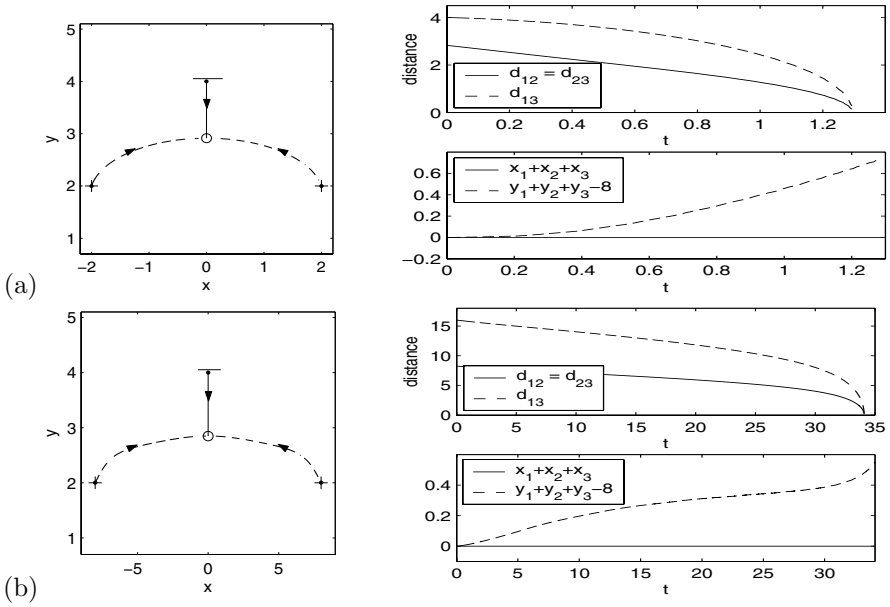


FIG. 9. *Time evolution of vortex centers by directly simulating the GLE when the initial data are chosen as Case* 3 *in Pattern* V *with different* a *and* b. (a) $a = 2$, $b = 4$; (b) $a = 8$, $b = 4$.

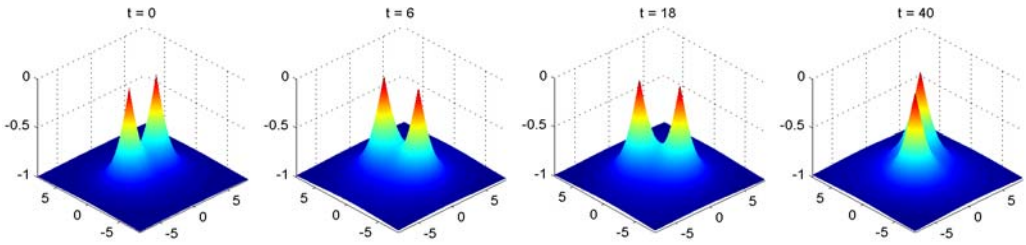


FIG. 10. *Surface plots of* $-|\psi|$ *at different times for the NLSE when the initial condition is chosen as Pattern* I (2.5) *with* $N = 2$, $m_0 = +1$, *and* $a = 2$.

4.1. Interactions of N ($N \geq 2$) like vortices, Patterns I and II. Figures 10–12 give the surface plots of $-|\psi|$ at different times, the slice plots of $|\psi(x, 0, t)|$ at different times, and some dynamical laws when the initial data in (1.7) are chosen as (2.5) with $N = 2$ and $m_0 = +1$, i.e., interaction of two like vortices. In addition, Figure 13 shows time evolution of the vortex centers when the initial condition is chosen as Pattern I (2.5) for different $N \geq 2$, and Figure 14 depicts similar results for Pattern II (2.6)–(2.7).

From Figures 10–14, and additional numerical experiments not shown here, we can draw the following conclusions for the interaction of N like vortices in the NLSE when the initial condition is chosen as either Pattern I or II:

(i) The signed mass center of the vortex centers is conserved for any time $t \geq 0$ (cf. Figures 13(b),(c),(e),(f); 14(b),(c),(e),(f)), which confirms the conservation law in (2.4).

(ii) Vortices with the same index behave like point vortices in an ideal fluid and never collide (cf. Figures 10, 13, and 14). In fact, there exists a critical time $t_0 > 0$,

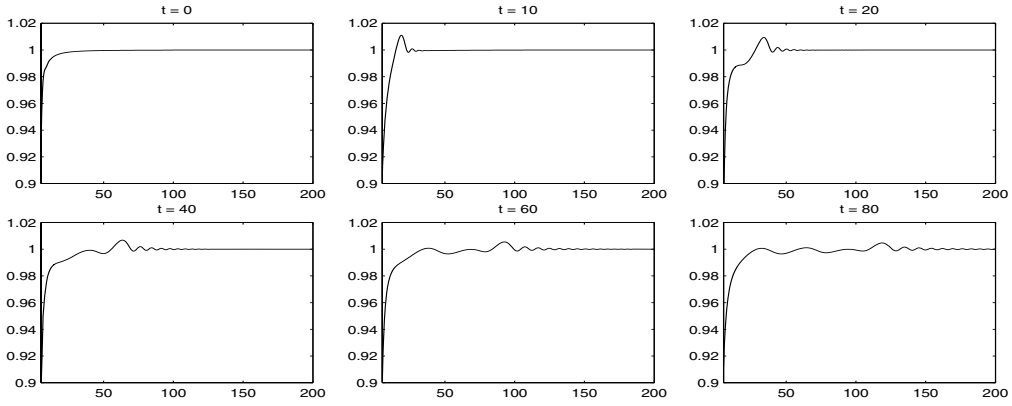


FIG. 11. *Plots of* $|\psi(x,0,t)|$ $(x \geq 4)$ *at different times for the NLSE when the initial condition is chosen as Pattern* I (2.5) *with* $N = 2$, $m_0 = +1$, *and* $a = 2$, *showing the sound wave propagation.*

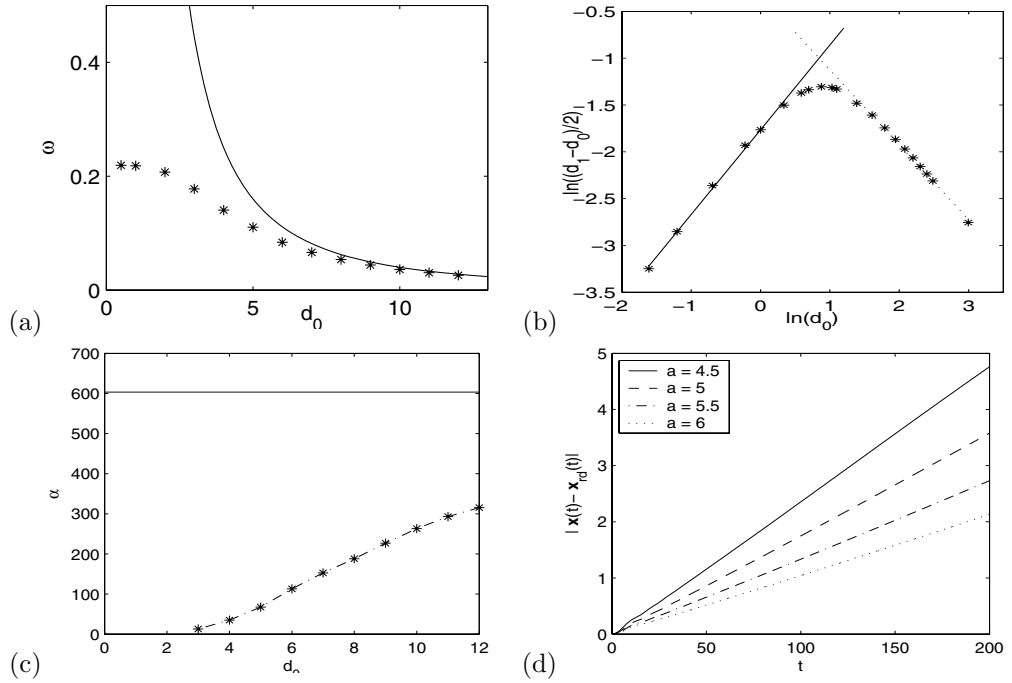


FIG. 12. *Dynamical laws of interaction between two like vortices, i.e., Pattern* I *with* $N = 2$, *in the NLSE.* (a) *Frequency* ω *of the rotation (solid line is from* (2.18) *and asterisks are our numerical results);* (b) *diameter* $d_1 = |\mathbf{x}_1(t_0) - \mathbf{x}_2(t_0)|$ *when the two vortices start to rotate on a circle;* (c) $\alpha(d_0)$ *in* (4.1) (*the asterisks are our numerical results and the solid line is from the theoretical predication* $\alpha = 2^6 \cdot 3\pi$ [30]); (d) *errors of the vortex centers between the solution* (2.18) *of the reduced dynamic laws* (*denoted* $\mathbf{x}_{\rm rd}(t)$) *and our directly simulating results of the NLSE* (*denoted* $\mathbf{x}(t)$) *for different initial distance* $d_0 = 2a$.

depending on the initial distance to the origin, i.e., a, such that before time t_0, i.e., when $0 \leq t \leq t_0$, the vortices initially located on a circle move from their initial locations to another circle, and the change in distance between each vortex to the origin is rapid (cf. Figures 13(b), 14(b)); after time t_0, i.e., for $t \geq t_0$, the vortices

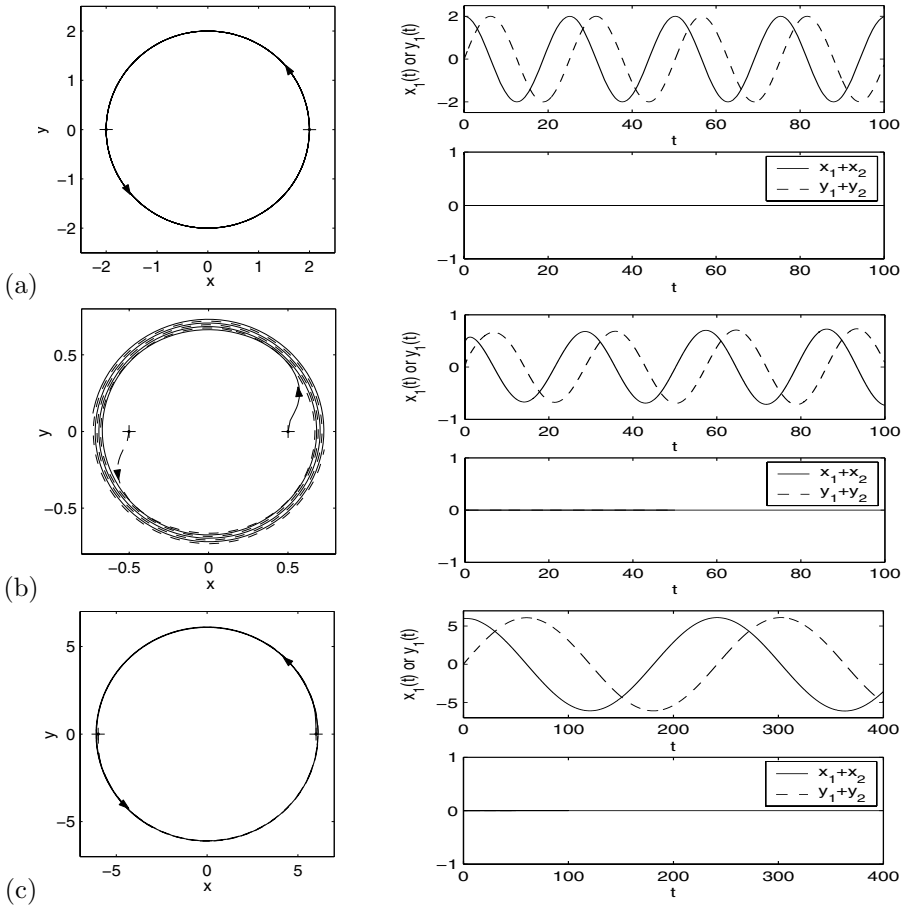


FIG. 13. *Time evolution of vortex centers for the NLSE when the initial condition is chosen as Pattern* I *with different* N. (a) *and* (d) *are from the reduced dynamic laws* (2.18)*;* (b) *and* (e) *and* (c) *and* (f) *show direct simulation results of the NLSE with* $a = O(r_1^0)$ *and* $a \gg r_1^0$*, respectively. Case* 1: $N = 2$ *with* (b) $a = 0.5$ *and* (c) $a = 6$.

rotate uniformly along a circle (counterclockwise when winding number $m_0 = +1$ and clockwise when $m_0 = -1$) with angular frequency ω depending on a and the radius of the circle increasing very slowly. The sound wave propagation is clearly observed during the interaction (cf. Figure 11). In addition, in Pattern II, the vortex initially at the origin does not move during the dynamics (cf. Figure 14(b),(c),(e),(f)), which confirms the analytical solution (2.21).

(iii) For Pattern I with $N = 2$, we also present the comparison quantitatively (cf. Figure 12). In this case, denote $d_0 = |\mathbf{x}_1^0 - \mathbf{x}_2^0| = 2a$ and $d_1 = |\mathbf{x}_1(t_0) - \mathbf{x}_2(t_0)|$ for the initial distance and the diameter of the circle at time $t = t_0$ of the two vortices, respectively. The angular frequency predicted by the reduced dynamics is confirmed by our numerical simulations (cf. Figure 12(a)) when $d_0 = 2a$ is large, and it is invalid when d_0 is small; i.e., the reduced dynamics is invalid when the vortex pair initially has overlapping support. Furthermore, even when the two vortices are well separated, the reduced dynamics fails to take into account the effect of the excessive energy and the radiation, which play important roles in the NLSE vortex dynamics. For example,

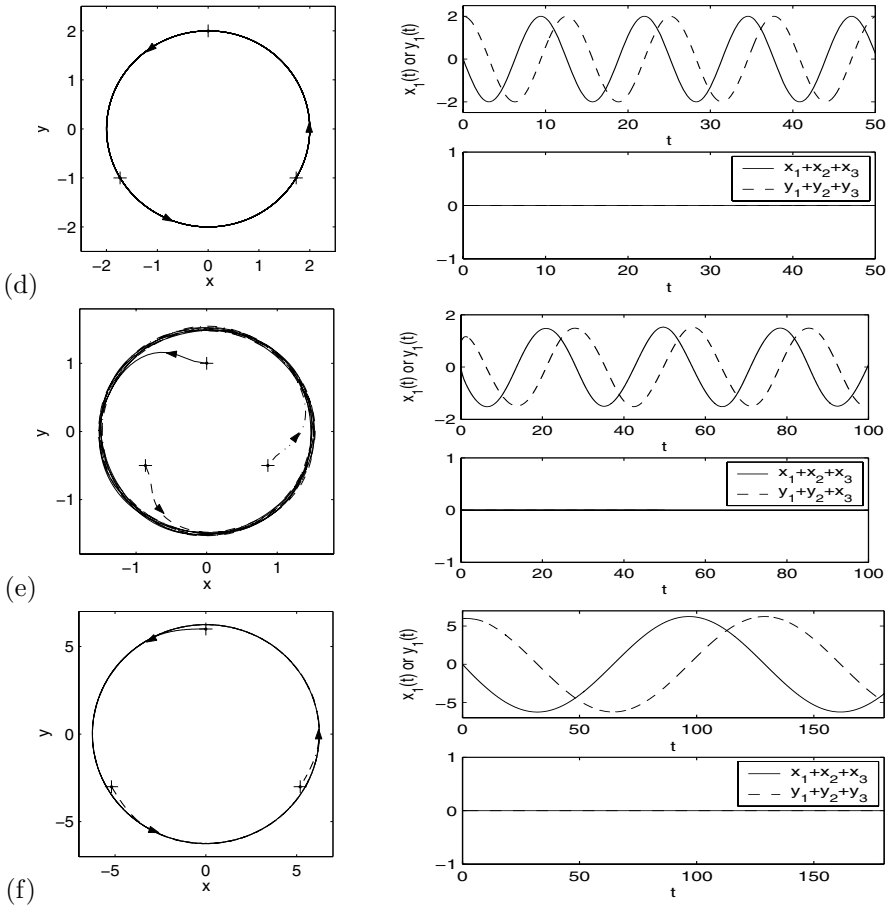


FIG. 13 (*cont.*). *Case* 2: $N = 3$ *with* (e) $a = 1$ *and* (f) $a = 6$.

by analyzing the next-order approximation for the interaction of two well-separated vortices in the NLSE, it was derived in [30] that the diameter of the circle increases on the order of $O(t^{1/6})$, i.e., asymptotically,

$$d(t) = |\mathbf{x}_1(t) - \mathbf{x}_2(t)| = \left(|\mathbf{x}_1^0 - \mathbf{x}_2^0|^6 + 2^6 \cdot 3\pi t\right)^{1/6} = \left(d_0^6 + 2^6 \cdot 3\pi t\right)^{1/6}. \tag{4.1}$$

This departs from the constant distance prediction made from the reduced dynamic laws (1.11). Numerically, we fit the distance between the two vortex centers $d(t) = |\mathbf{x}_1(t) - \mathbf{x}_2(t)|$ for $t \geq t_0$ by

$$d(t) = |\mathbf{x}_1(t) - \mathbf{x}_2(t)| = \left(d(t_0)^6 + \alpha(d_0)(t - t_0)\right)^{1/6}, \qquad t \geq t_0, \tag{4.2}$$

with $\alpha(d_0)$ being a constant depending on d_0. The results show that (4.1) is a very good prediction (cf. Figure 12(c)). Of course, much more detailed information on the vortex dynamics in the NLSE can be found through our numerical simulations. For example, our simulations suggest that when the initial distance between the two vortex centers increases, the time t_0 increases, the diameter $d_1 = d(t_0)$ of the circle at $t = t_0$ increases (cf. Figure 12(b)), and $\alpha(d_0)$ in (4.2) increases (cf. Figure 12(c)). From Fig-

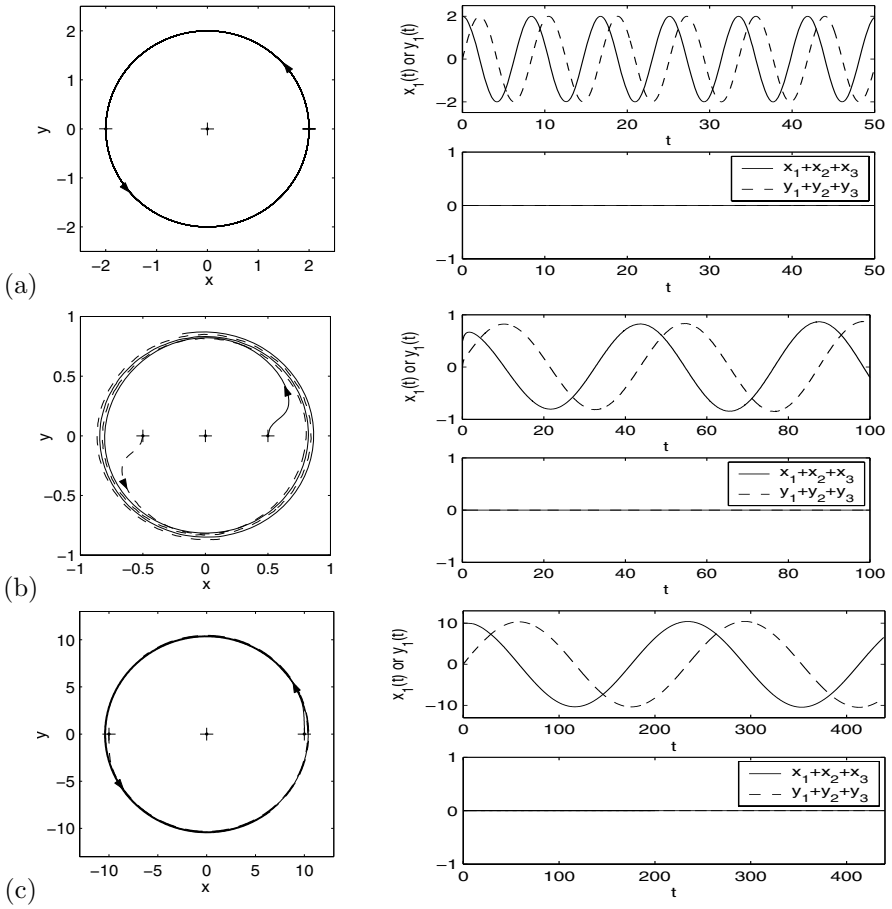


FIG. 14. *Time evolution of vortex centers for the NLSE when the initial condition is chosen as Pattern* II *with different* N. (a) *and* (d) *are from the reduced dynamic laws* (2.21)–(2.22)*;* (b) *and* (e) *and* (c) *and* (f) *show direct simulation results of the NLSE with* $a = O(r_1^0)$ *and* $a \gg r_1^0$*, respectively. Case* 1: $N = 3$ *with* (b) $a = 0.5$ *and* (c) $a = 10$.

ure 12(b), we have the numerical dynamical laws for the diameter d_1 for different d_0:

$$d_1 := d(t_0) \approx \begin{cases} d_0 + d_0^{0.9053}/2.9189, & d_0 < r_1^0, \\ d_0 + 1.4453/d_0^{0.7996}, & d_0 > 2r_1^0, \end{cases}$$

where $r_1^0 \approx 1.75$ [36] is the core size for the vortex state ϕ_m in (1.4) of the GLSE with winding number $m = \pm 1$.

(iv) In Patterns I and II, the solutions of the reduced dynamic laws agree qualitatively with our numerical results of the NLSE, and quantitatively when time t is small and they are well separated, i.e., $a \gg r_1^0 = 1.75$ [36]. In general, for a fixed initial distance, i.e., a, the error increases when time increases; for a given time, the error decreases when the initial distance increases (cf. Figure 12(d)). This again suggests that the reduced dynamics for governing time evolution of the vortex centers in the NLSE is valid only when time t is small and the initial distance is very large. Corrections must be added, e.g., such as (4.1), when either the time t is large or the initial distance is not very large.

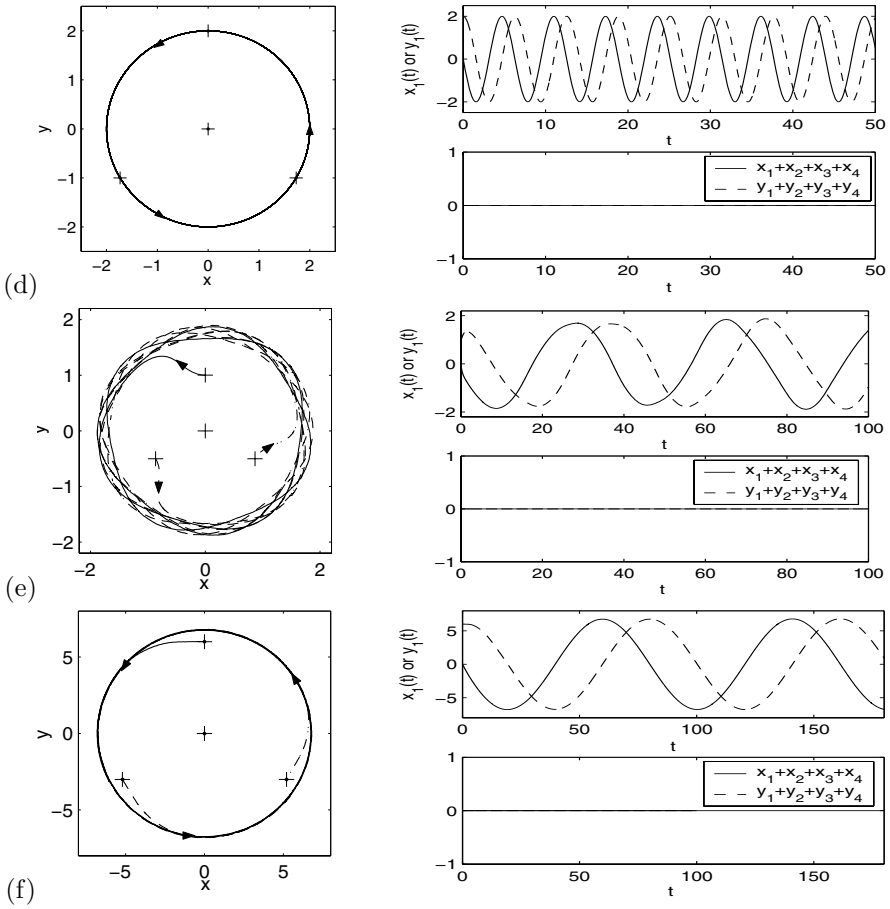


FIG. 14 (*cont.*). *Case* 2: $N = 4$ *with* (e) $a = 1$ *and* (f) $a = 6$.

(v) The results in Figure 12, as well as in Figure 21, also confirm Kirchoff's laws rigorously derived in [23, 24] for the interaction of two vortices in the NLSE, i.e., $\alpha = 0$, $\beta = 1$, $V(\mathbf{x}) \equiv 1$ in (1.1), when $\varepsilon \to 0$ with the initial distance between the two vortex centers fixed. In fact, the vortex interactions of (1.1) with $V(\mathbf{x}) \equiv 1$, $\varepsilon = 1$, and increased initial distances between the vortex centers are equivalent to those of (1.1) with $V(\mathbf{x}) \equiv 1$, $\varepsilon \to 0$, and fixed initial distances between the vortex centers by applying a rescaling.

4.2. Interactions of N ($N \geq 3$) opposite vortices, Pattern III. Figure 15 shows the time evolution of the vortex centers for different $N \geq 3$ when the initial data in (1.7) are chosen as in Pattern III with $m_0 = +1$ for different N and a.

From Figure 15, and additional numerical experiments not shown here, we can draw the following conclusions for the interaction of N opposite vortices in the NLSE when the initial data are chosen as Pattern III:

(i) The signed mass center of the vortex centers is conserved for any time $t \geq 0$ (cf. Figure 15(b),(c),(e),(f),(h),(i)), which confirms the conservation law in (2.4).

(ii) The vortex initially at the origin does not move for any time $t \geq 0$ (cf. Figure 15(b),(c),(e),(f),(h),(i)), which confirms the analytical solution (2.24). After a critical

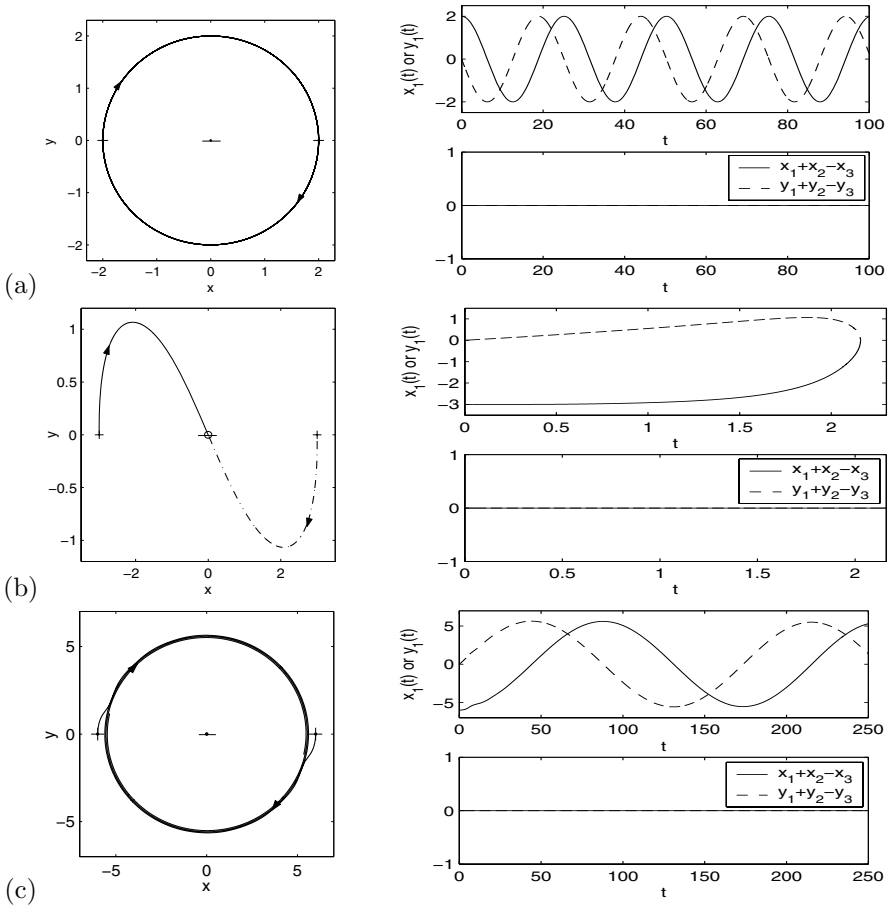


FIG. 15. *Time evolution of vortex centers for the NLSE when the initial condition is chosen as Pattern* III *with different* N *and* a. (a), (d), *and* (g) *are from the reduced dynamic laws* (2.25); (b), (e), *and* (h) *and* (c), (f), *and* (i) *show direct simulation results of the NLSE with* $a = O(r_1^0)$ *and* $a \gg r_1^0$, *respectively. Case* 1: $N = 3$ *with* (b) $a = 3$ *and* (c) $a = 6$.

time t_c depending on the initial radius a, the vortices initially located on a circle rotate clockwise when $N = 3$ and $a > a_{\rm cr} \approx 2r_1^0$ or $N = 4$, and, respectively, counterclockwise when $N \geq 5$, along a circle, and at any time $t \geq 0$, these vortex centers are always on a circle (cf. Figure 15(b),(c),(e),(f),(h),(i)), which confirms the analytical solutions (2.25). Their angular frequencies depend on their distances to the origin, i.e., the larger the distance, the slower the motion.

(iii) For the case of $N = 3$, when the initial radius $a < 2r_1^0$, the three vortices undergo attractive interactions, and the two vortices initially on a circle move symmetrically towards the center before a critical time t_c. When $t = t_c$, they collide at the origin (cf. Figure 15(b)), and after it, only one vortex with a winding number m_0 is left and stays at the point $(0, 0)$ for any time $t > t_c$. On the other hand, when $a > 2r_1^0$, the two vortices rotate (clockwise for $m_0 = +1$, and, respectively, counterclockwise for $m_0 = -1$) on a circle whose radius increases very slowly with time t (cf. Figure 15(c)).

(iv) When $N = 4$, the three vortices initially on a circle move to another circle

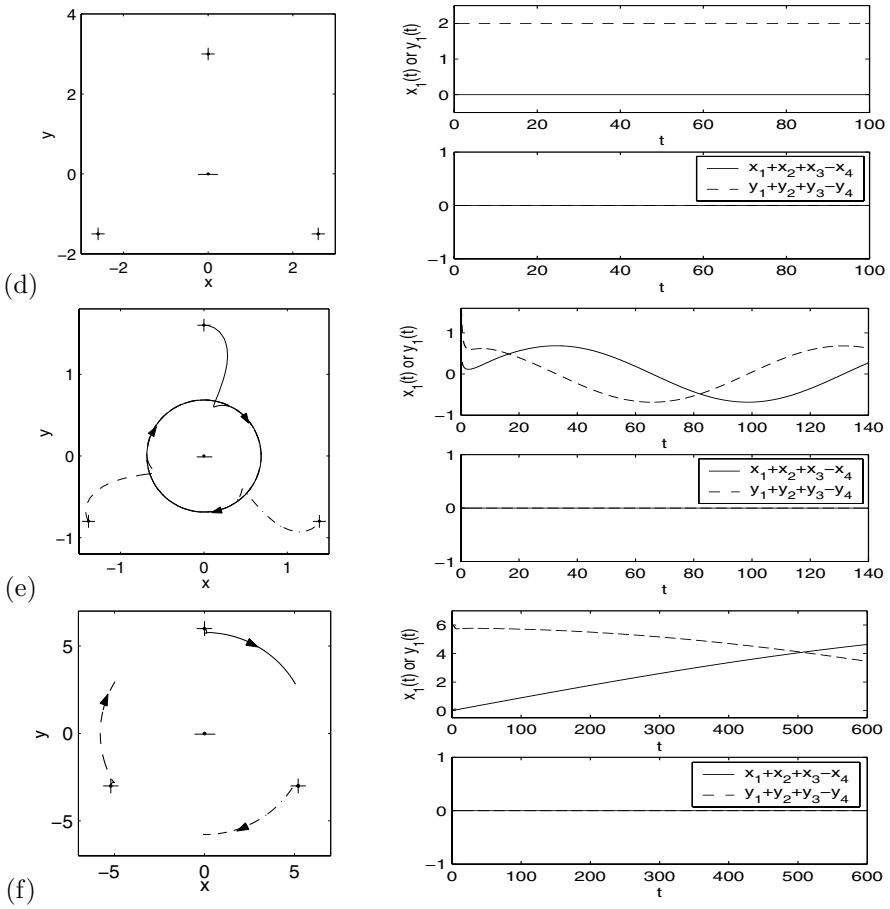


FIG. 15 (*cont.*). *Case* 2: $N = 4$ *with* (e) $a = 1.6$ *and* (f) $a = 6$.

with radius $a_1 < a$; then they rotate (clockwise for $m_0 = +1$, and, respectively, counterclockwise for $m_0 = -1$) on a circle whose radius increases very slowly with time t (cf. Figure 15(e),(f)). The four vortices never collide no matter how small a is.

(v) When $N \geq 5$, the $N-1$ vortices initially on a circle move to another circle with radius $a_1 > 0$; then they rotate (counterclockwise for $m_0 = +1$, and, respectively, clockwise for $m_0 = -1$) on a circle whose radius increases very slowly with time t (cf. Figure 15(h),(i)).

(vi) In Pattern III, when $N = 3$ and $a > 2r_1^0$ or $N \geq 5$, the solutions of the reduced dynamic laws agree qualitatively with our numerical results of the NLSE. On the contrary, when $N = 4$ or $N = 3$ with $a < 2r_1^0$, they are completely different! This may be attributed to the lack of well separation between the vortex cores and/or the next-order effect in the underlying vortex motion of the original NLSE. In fact, the angular frequency for $N = 4$ is much larger than that for $N = 3$ with the same initial radius of the circle (cf. Figure 15(c),(f)).

4.3. Interactions of two opposite vortices, Pattern IV. Figure 16 displays the surface plots of $-|\psi|$ at different times when the initial data in (1.7) are chosen as (2.8), i.e., Pattern IV, with $m_0 = +1$ and $a = 1.5$ or $a = 5$. Figures 17 and 18–19 plot $|\psi(x, y(t), t)|$ and $|\psi(0, y, t)|$, respectively, to show the sound wave propagation

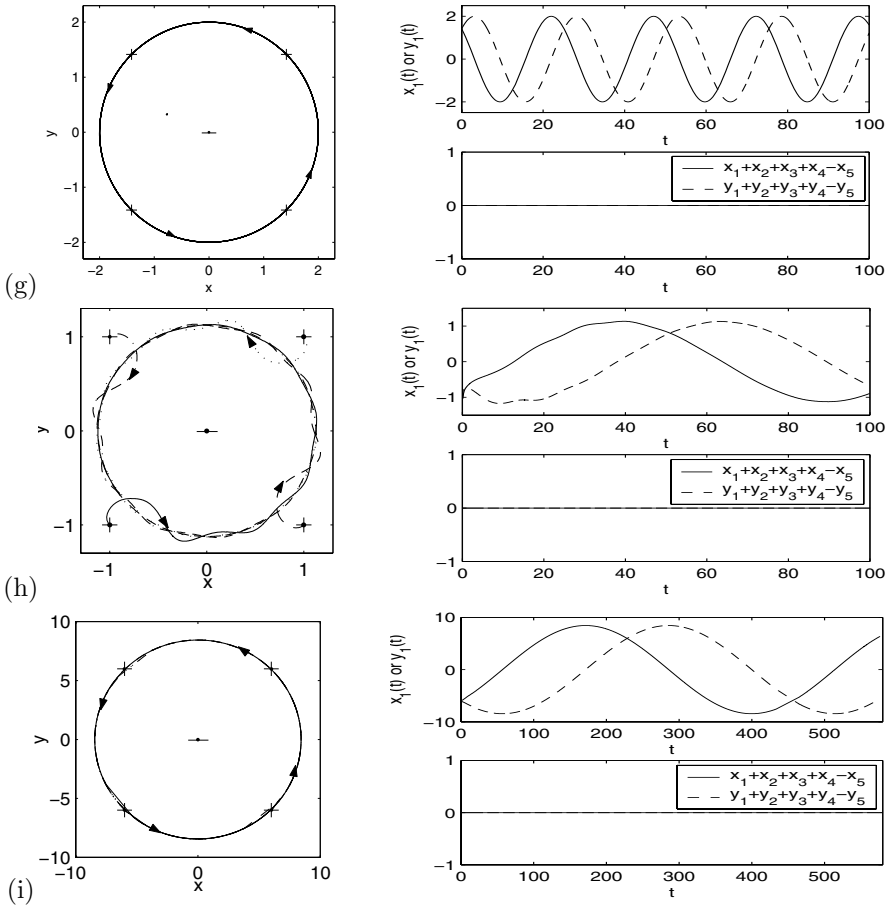


FIG. 15 (*cont.*). *Case* 3: $N = 5$ *with* (h) $a = 1$ *and* (i) $a = 6$.

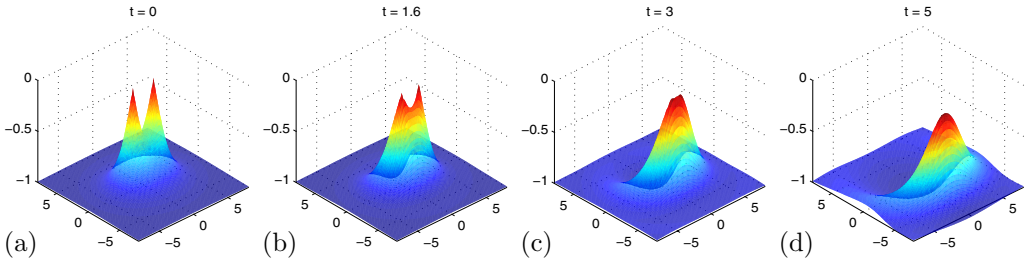


FIG. 16. *Surface plots of* $-|\psi|$ *at different times for the NLSE when the initial condition is chosen as Pattern* IV (2.8) *with* $m_0 = +1$. I. $a = 1.5 = O(r_1^0)$.

during the dynamics. Figure 20 shows the time evolution of the two vortex centers with different $d_0 = 2a$. In addition, Figure 21 shows some dynamical laws for the interaction.

From Figures 16–21, we can draw the following conclusions for the interaction of two opposite vortices in the NLSE when the initial condition is chosen as Pattern IV:

(i) The signed mass center of the two vortex centers is not conserved, at least when either the initial distance between the two vortices is not large or time t is small

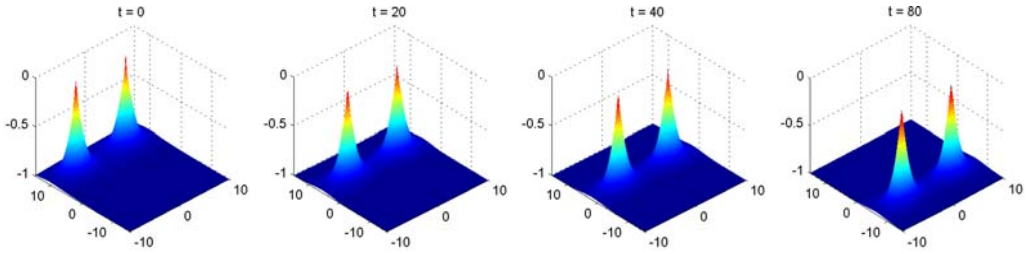


FIG. 16 (*cont.*). II. $a = 5 \gg r_1^0$.

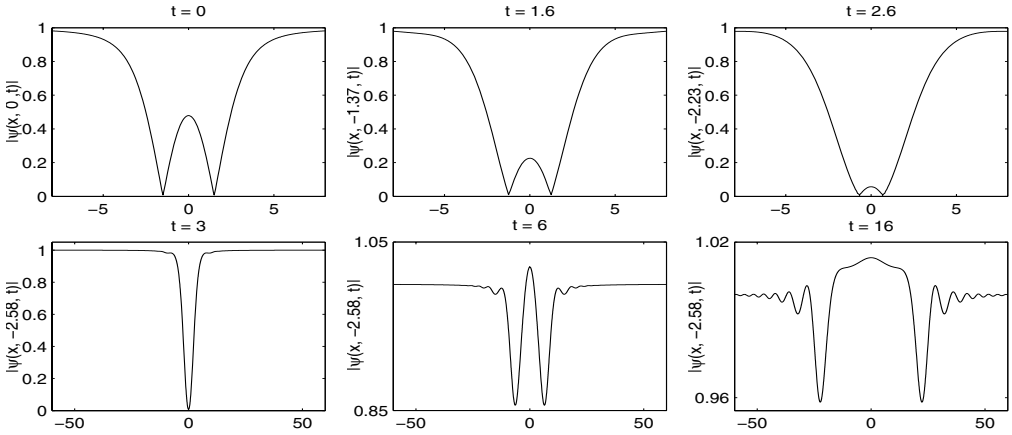


FIG. 17. *Plots of* $|\psi(x, y(t), t)|$ *at different times for the NLSE when the initial condition is chosen as Pattern* IV (2.8) *with* $m_0 = +1$ *and* $a = 1.5$*, showing sound wave propagation and radiation with the values of* $y(t)$ *given in the labels. Here* $y = y(t)$ *is the line passing through the two vortex centers at time* t *before they merge with each other around* $t_c \approx 3.0$.

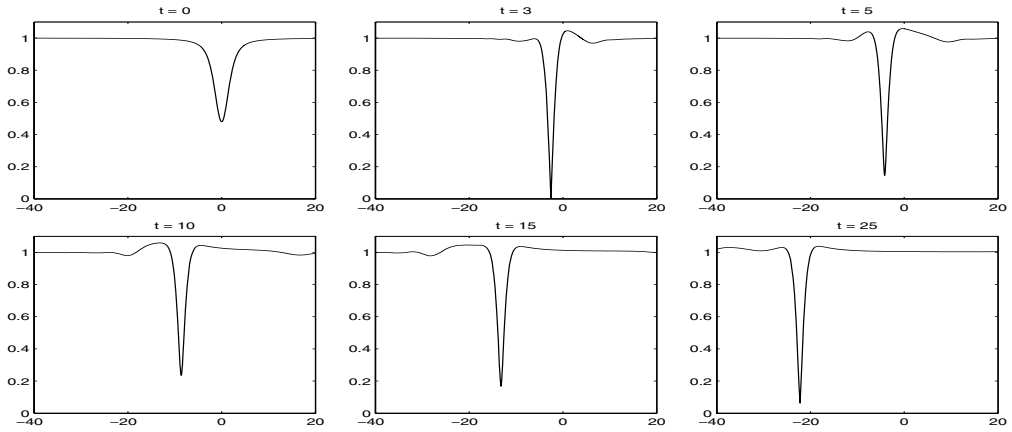


FIG. 18. *Plots of* $|\psi(0, y, t)|$ *at different times for the NLSE when the initial condition is chosen as Pattern* IV (2.8) *with* $m_0 = +1$ *and* $a = 1.5$*, showing solitary-like wave propagation.*

(cf. Figure 20(b),(c),(d)), which suggests that the conservation law in (2.4) is invalid when the initial distance between the two vortex centers at time $t = 0$ is not large.

(ii) There is a critical distance $d_{\rm cr}$ satisfying that, for $d_0 = |\mathbf{x}_1^0 - \mathbf{x}_2^0| < d_{\rm cr}$, the

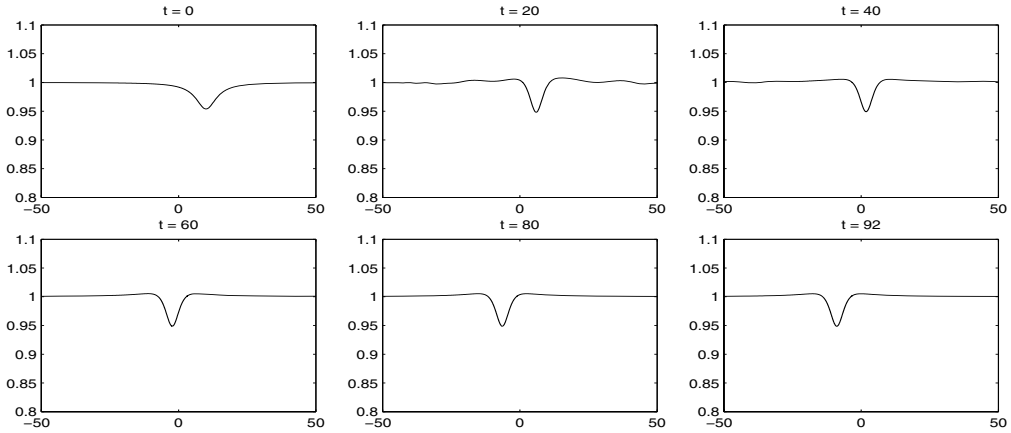


FIG. 19. *Plots of* $|\psi(0, y, t)|$ *at different times for the NLSE when the initial condition is chosen as Pattern* IV (2.8) *with* $m_0 = +1$ *and* $a = 5 \gg r_1^0$ *for solitary wave propagation.*

two vortices approach each other while drifting sideways and then collide and are annihilated at time $t = t_c$ (cf. Figures 16, 20(b)), and for $d_0 = |\mathbf{x}_1^0 - \mathbf{x}_2^0| > d_{cr}$, they move almost in a parallel course, perpendicular to the line joining them (cf. Figures 19, 20(c),(d)). Our numerical simulations suggest that $d_{\rm cr} \approx 2r_1^0 = 2 \times 1.75 = 3.5$, i.e., double the size of the core size r_1^0, which is almost triple the size of the theoretical prediction $d_{\rm cr} \approx \sqrt{2}$ derived in [30].

(iii) When $d_0 < d_{\rm cr} = 2r_1^0$, before collision, our numerical simulation reveals that two sound waves moving towards each other are generated along the line joining the centers of the two vortices (cf. Figure 17), which cause the collision, while no radiation is observed; after the collision, some outgoing radiation is observed along with a solitary-like sound wave also being observed in the y-axis (cf. Figure 18). In addition, a discontinuity or shock wave in the hydrodynamic velocity is observed just after the collision. Furthermore, for the initial setup in Pattern IV, the two vortices collide at the point $(0, -d_2)$ with $d_2 > 0$ when $t = t_c$. When the initial distance d_0 increases, both t_c and d_2 increase, and our numerical results suggest the following relation between them:

$$t_c \approx \frac{1}{7.0790}\, d_0^{2.0954}, \qquad d_1 \approx \frac{1}{1.9300}\, d_0^{1.0365}, \quad \text{with} \quad d_1 = \sqrt{d_0^2 + d_2^2}.$$

(iv) When $d_0 \gg d_{\rm cr} = 2r_1^0$, the two vortices drift almost on two parallel lines, perpendicular to the line joining them with a constant speed. Our numerical results confirm the speed (2.29) when $d_0 = 2a$ is large (cf. Figure 21(a)). In addition, a solitary wave is observed during the dynamics (cf. Figure 19).

(v) Again, in Pattern IV the solutions of the reduced dynamic laws agree qualitatively with our numerical results of the NLSE when $a \gg r_1^0$ and they are completely invalid when a is small (cf. Figure 20). When $a > r_1^0$, in general, for a fixed initial distance, the error increases when time increases; for a given time, the error decreases when the initial distance increases (cf. Figure 21(b)).

4.4. Interactions of vortices with nonsymmetric setups. Figures 22–24 show time evolution of the vortex centers when the initial data in (1.7) are chosen as the three cases in Pattern V, respectively.

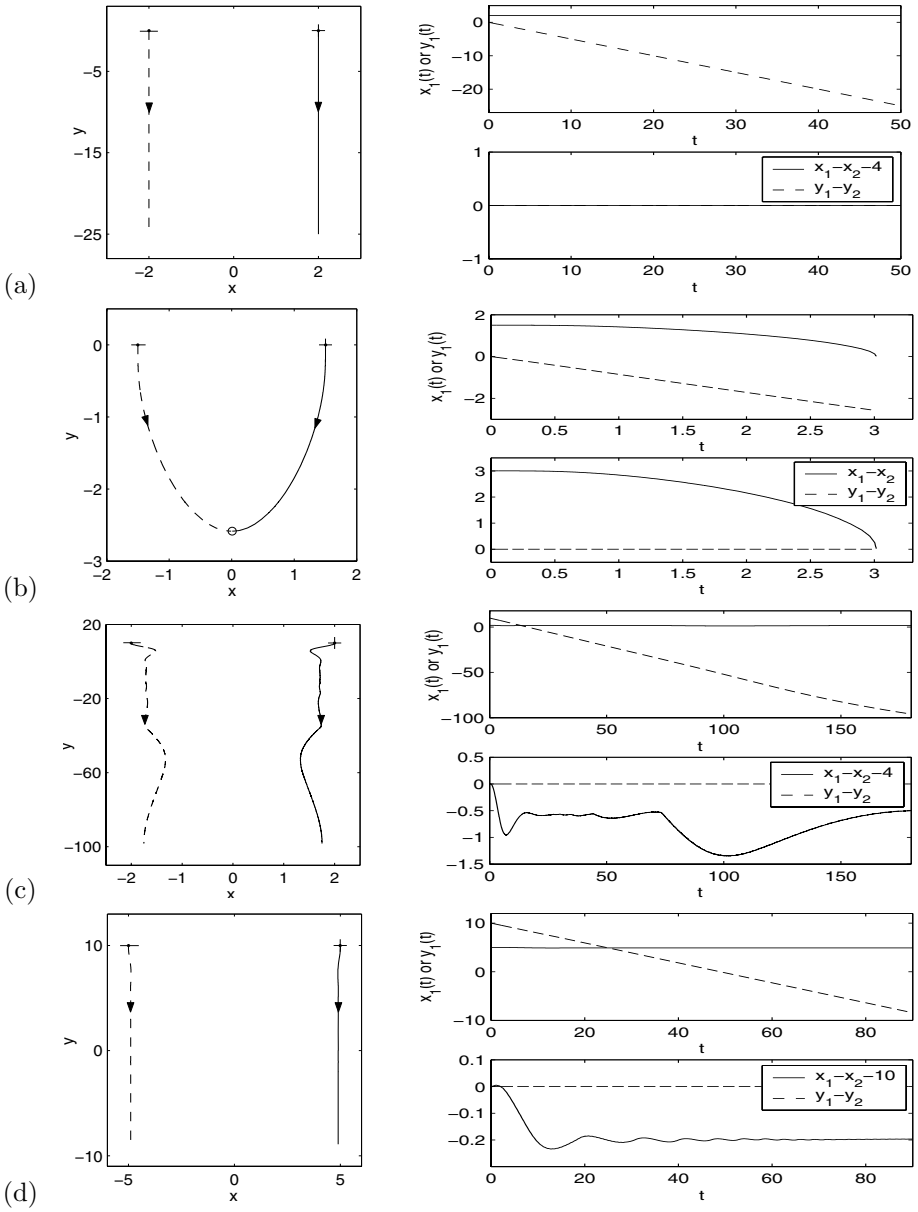


FIG. 20. *Time evolution of vortex centers for the NLSE when the initial condition is chosen as Pattern* IV. (a) *is from the reduced dynamic laws* (2.26)*, and* (b), (c)*, and* (d) *show direct simulation results of the NLSE with* $m_0 = +1$ *and* $a = 1.5 < r_1^0$*,* $a = 2 > r_1^0$*, and* $a = 5 \gg r_1^0$*, respectively.*

Based on Figures 22–24 and our additional numerical experiments, we can draw the following conclusions for three vortices with nonsymmetric initial setups:

(i) When they have the same index, they rotate and never collide (cf. Figure 22). On the contrary, when they have opposite indices, there exists a critical distance $d_{\rm cr}$, when their initial distances are less than $d_{\rm cr}$, they collide at finite time (cf. Figures 23(a), 24(a)), and after collision, only one vortex is left; on the other hand, when their

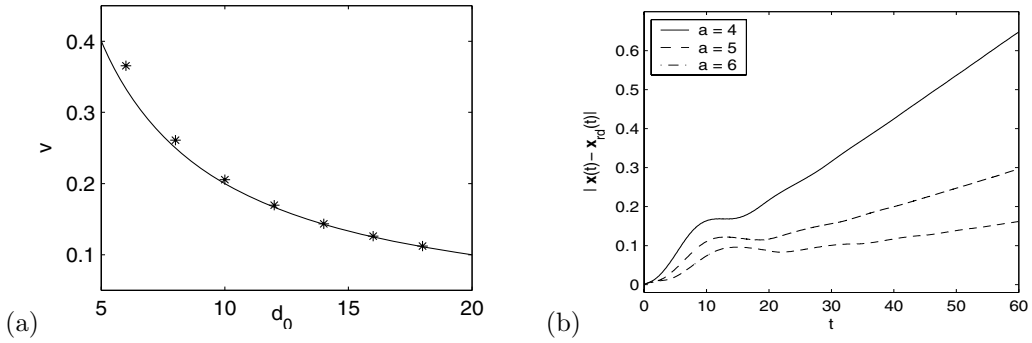


FIG. 21. *Dynamic laws for two opposite vortices, i.e., Pattern* IV *with* $N = 2$*, in the NLSE.* (a) *Speed* v *of the parallel motion.* (b) *Errors of the vortex centers between the solution* (2.26) *of the reduced dynamic laws* (*denoted as* $\mathbf{x}_{\rm rd}(t)$) *and our directly simulating results of the NLSE* (*denoted as* $\mathbf{x}(t)$) *for different initial distance* $d_0 = 2a$.

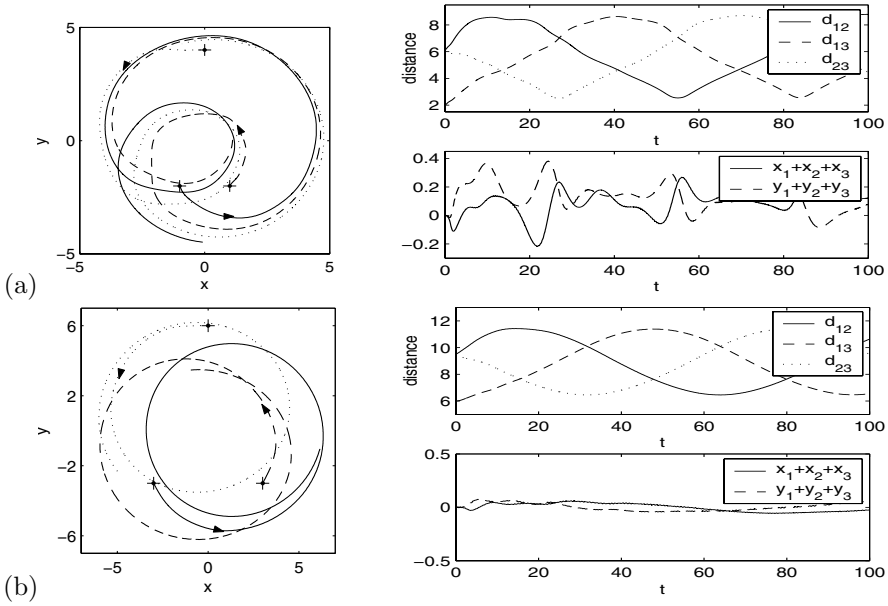


FIG. 22. *Time evolution of vortex centers for the NLSE when the initial condition is chosen as Case* 1 *in Pattern* V *with different* a *and* b. (a) $a = 1$, $b = 4$. (b) $a = 3$, $b = 6$.

initial distances are larger than $d_{\rm cr}$, two of them move in a parallel course and never collide (cf. Figures 23(b), 24(b)).

(ii) The signed mass centers of the vortex centers are not conserved (cf. Figures 22–24) during the dynamics, and these suggest that the reduced dynamics law (1.11)–(1.12) has considerable discrepancy in some regimes. Again, one may argue that a possible cause is the fact that the reduced dynamic law is the adiabatic approximation in the leading order when the N vortices are well separated, and thus the next-order effect becomes important in the underlying vortex motion of the original NLSE when the N vortices are not well separated. Also in our numerical results, the larger the distance between the vortex centers, the better the conservation of the signed mass center (cf. Figures 22–24). This again suggests that the reduced dynam-

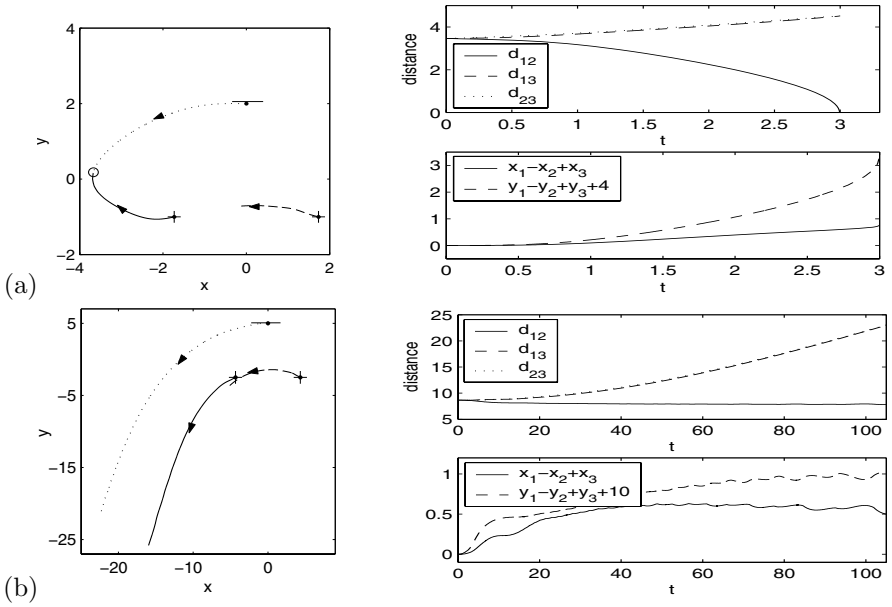


FIG. 23. *Time evolution of vortex centers for the NLSE when the initial condition is chosen as Case* 2 *in Pattern* V *with different* a. (a) $a = 2$. (b) $a = 5$.

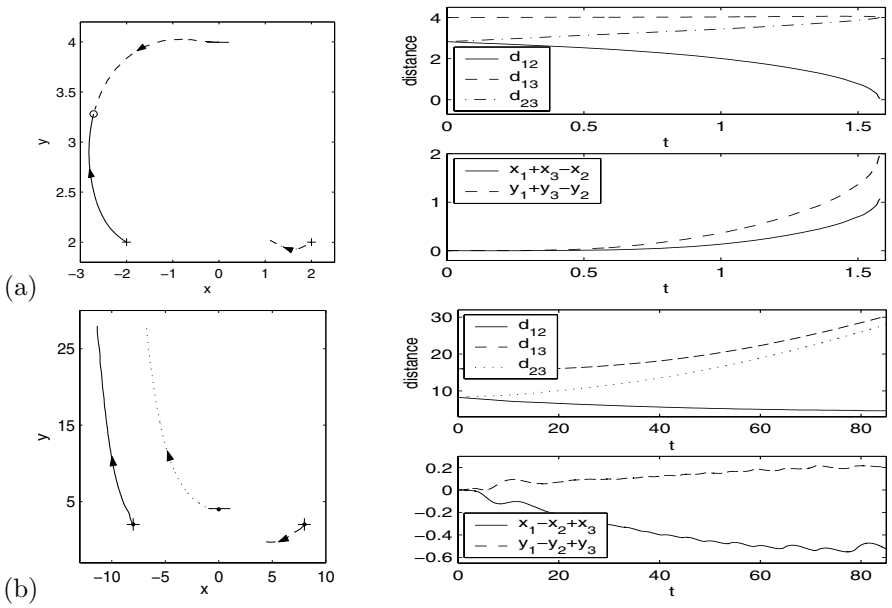


FIG. 24. *Time evolution of vortex centers for the NLSE when the initial condition is chosen as Case* 3 *in Pattern* V *with different* a *and* b. (a) $a = 2$, $b = 4$. (b) $a = 8$, $b = 4$.

ics law (1.11)–(1.12) is still a reasonable approximation to the vortex motion of the original NLSE in a nonsymmetric initial setup when the N vortices are well separated.

5. Vortex dynamics in the CGLE or in the GLSE with an external potential. In many applications of the GLSE, the physical situation is often more

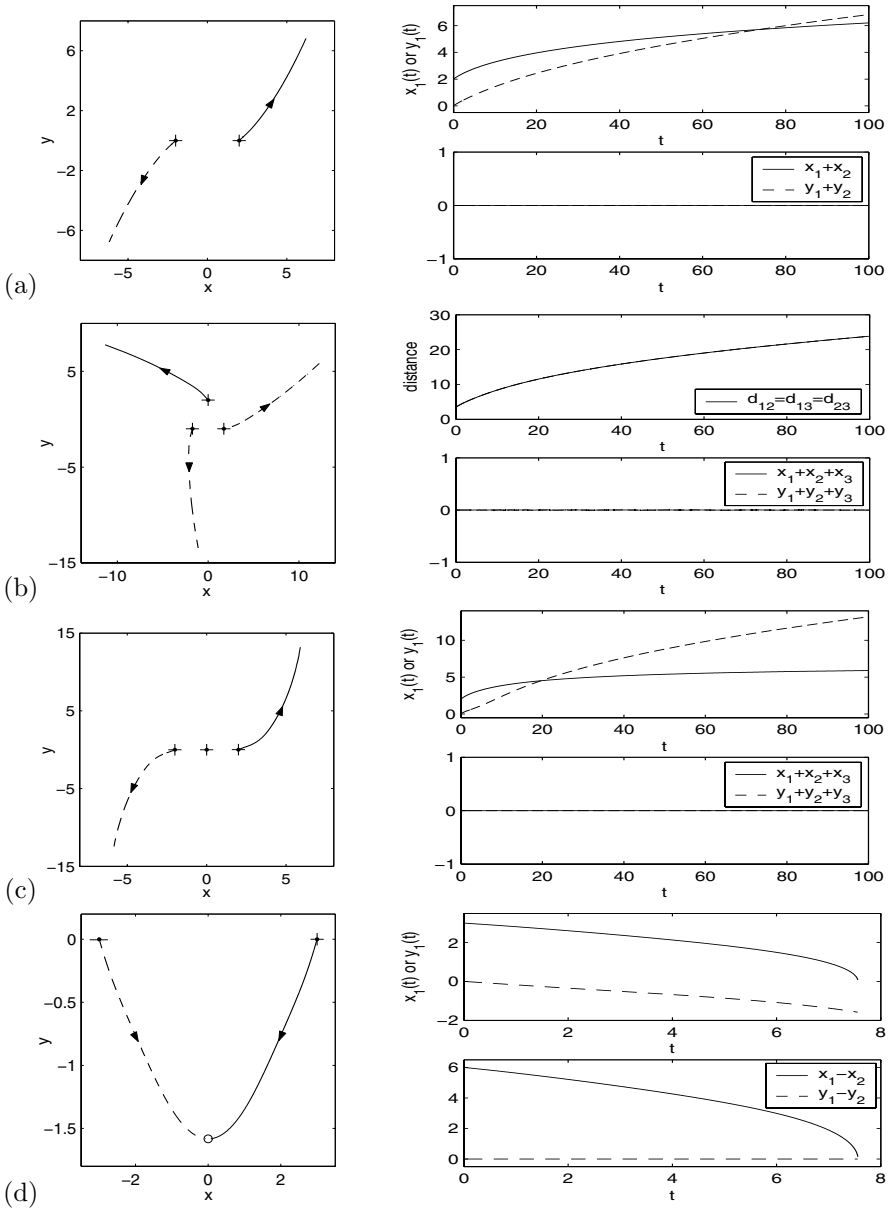


FIG. 25. *Time evolution of vortex centers for the CGLE when the initial condition is chosen as Patterns* I–IV *with different a.* (a) *Pattern* I *with* $N = 2$ *and* $a = 2$. (b) *Pattern* I *with* $N = 3$ *and* $a = 1.5$. (c) *Pattern* II *with* $N = 3$ *and* $a = 2$. (d) *Pattern* IV *with* $a = 3$.

complicated than the GLE and NLSE cases considered in the earlier sections. As an illustration, in this section we report numerical results of vortex interaction in the CGLE and vortex motion in the GLSE with an inhomogeneous external potential.

5.1. Numerical results for vortex dynamics in the CGLE. We take $\alpha = \beta = 1$, $\varepsilon = 1$, and $V(\mathbf{x}) \equiv 1$ in (1.1). Figure 25 shows the various time evolutions of the vortex centers when the initial data in (1.7) are chosen as Patterns I, II, and IV.

Based on Figure 25 and our additional numerical experiments, we can draw the following conclusions for vortex dynamics in the CGLE:

(i) Vortices with the same index undergo repulsive interactions and never collide. The trajectories are combinations of those from the GLE and the NLSE (cf. Figure 25(a), (b), and (c)).

(ii) Two vortices with opposite indices collide after some time t_c (cf. Figure 25(d)) and the collision position is $(0, -d_2)$. The collision time and position depend on the initial distance between the two vortices $d_0 = 2a$. Our numerical results suggest the following relation between them:

$$t_c \approx \frac{1}{8.9837}\, d_0^{2.0655}, \qquad d_1 \approx \frac{1}{4.7781}\, d_0^{1.0184}, \quad \text{with} \quad d_1 = \sqrt{d_0^2 + d_2^2}.$$

In addition, based on the numerical results in Figure 25 it is reasonable to make the following conjecture about the reduced dynamic laws for the interaction of N well-separated vortices with winding number $m_j = +1$ or -1:

$$\mathbf{v}_j(t) := \frac{d\mathbf{x}_j(t)}{dt} = 2 \sum_{l=1,\, l\neq j}^{N} m_l \frac{\mathbf{Q}_j\,(\mathbf{x}_j(t) - \mathbf{x}_l(t))}{|\mathbf{x}_j(t) - \mathbf{x}_l(t)|^2}, \qquad t \geq 0, \tag{5.1}$$

$$\mathbf{x}_j(0) = \mathbf{x}_j^0, \qquad 1 \leq j \leq N, \tag{5.2}$$

where $\mathbf{Q}_j$ is given as

$$\mathbf{Q}_j = \begin{pmatrix} m_j\,\kappa_1 & -\kappa_2 \\ \kappa_2 & m_j\,\kappa_1 \end{pmatrix} = m_j \kappa_1\, I + \kappa_2\, J, \qquad j = 1, 2, \ldots, N,$$

with κ_1 and κ_2 being constants determined from α, β in (1.1) and the initial setup (1.7). Formal derivation of the above reduced dynamics laws (5.1) for the CGLE can be followed from those in [27] for the GLE and NLSE. Again, the nonlinear ODEs (5.1) can be solved analytically as those in section 2.3 for the GLE and section 2.4 for the NLSE when the initial conditions in (5.2) are given by Patterns I–IV in section 2.2. The details are omitted here. For comparison, Figure 26 shows numerical solutions of (5.1) for different initial setups. This figure clearly confirms our conjecture (5.1) about the reduced dynamic laws of the CGLE for the interaction of N well-separated vortices with winding number $m_j = +1$ or -1.

5.2. Vortex motion under an inhomogeneous external potential. The particular external potential we take is of the form

$$V(\mathbf{x}) = \frac{\frac{1}{2} + \gamma_x x^2 + \gamma_y y^2}{1 + \gamma_x x^2 + \gamma_y y^2} = 1 - \frac{1}{2\,(1 + \gamma_x x^2 + \gamma_y y^2)}, \qquad \mathbf{x} \in \mathbb{R}^2, \tag{5.3}$$

where γ_x and γ_y are two positive constants. It is easy to see that $V(\mathbf{x})$ attains its minimum value $1/2$ at the origin $(0, 0)$. Here we study numerically the dynamics of a vortex in the following two cases:

Case I. Isotropic external potential, e.g., $\gamma_x = \gamma_y = 1$ in (5.3).

Case II. Anisotropic external potential, e.g., $\gamma_x = 1$ and $\gamma_y = 5$ in (5.3).

For the GLE, i.e., $\alpha = 1$ and $\beta = 0$ in (1.1), the velocity of the induced motion due to the inhomogeneous impurities was obtained in [17]:

$$\mathbf{v}(t) := \frac{d\mathbf{x}(t)}{dt} = -\nabla \ln V(\mathbf{x}(t)), \quad t \geq 0, \qquad \text{with} \quad \mathbf{x}(0) = \mathbf{x}^0. \tag{5.4}$$

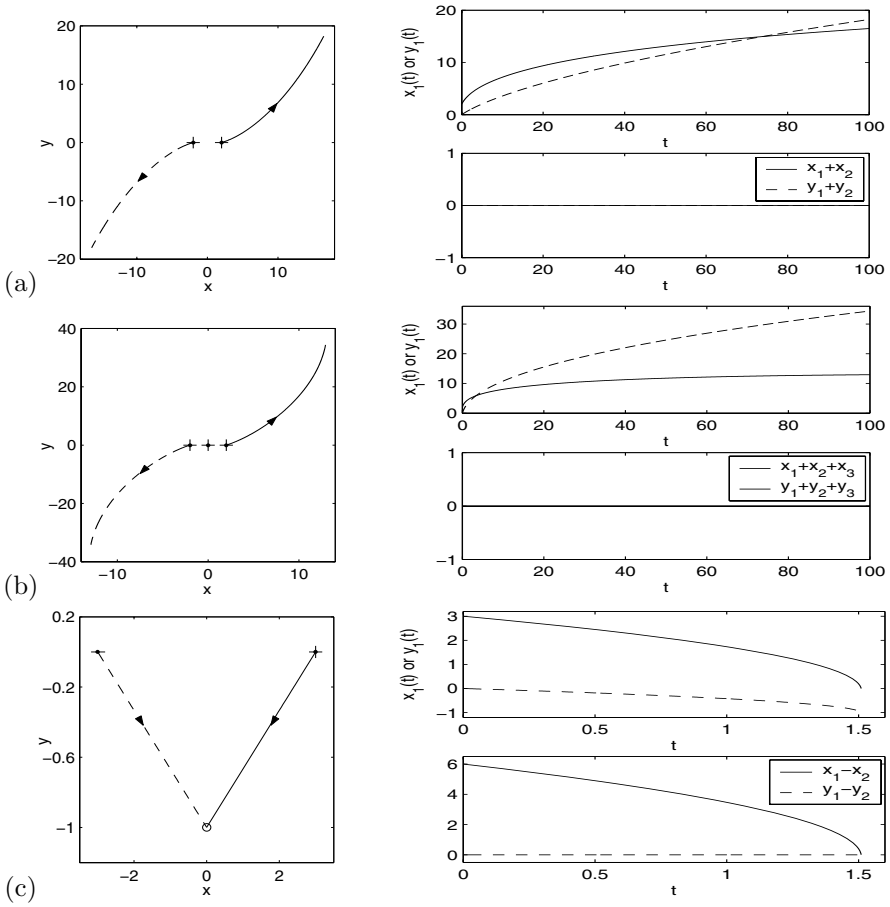


FIG. 26. *Numerical solutions of the reduced dynamical laws* (5.1) *for the CGLE with* $\kappa_1 = 3$ *and* $\kappa_2 = 1$ *for different initial setups.* (a) *Pattern* I *with* $N = 2$ *and* $a = 2$. (b) *Pattern* II *with* $N = 3$ *and* $a = 2$. (c) *Pattern* IV *with* $a = 3$.

This implies that here, the vortex would move to the minimizer of the external potential $V(\mathbf{x})$. Furthermore, if the external potential is isotropic, the trajectory is a segment connecting $\mathbf{x}^0$ and the minimization point of $V(\mathbf{x})$, while for the NLSE and CGLE, the dynamic laws with impurities remain to be established.

The initial condition in (1.2) is chosen as $\psi(\mathbf{x}, 0) = \phi_1\left(\mathbf{x} - \mathbf{x}^0\right)$ for $\mathbf{x} \in \mathbb{R}^2$, where $\phi_1 = \phi_1(\mathbf{x})$ is the vortex state solution (1.4) with winding number $m = +1$ and $\mathbf{x}^0$ is a given point. Figure 27 displays the time evolution of the vortex center in the GLE with $\mathbf{x}^0 = (1, 2)^T$ for different ε, and Figures 28 and 29 show similar results for the CGLE and NLSE, respectively.

From Figures 27–29, we can draw the following conclusions. First, for the GLE and CGLE, the vortex center moves monotonically to the position where the external potential $V(\mathbf{x})$ attains its minimum value (cf. Figures 27 and 28). The speed of the motion depends on the values of the parameter ε. The trajectory of the vortex center depends on the external potential $V(\mathbf{x})$, which agrees with the analytical results for the GLE in [16, 17, 21]. After the vortex reaches the minimum point of the external potential, it always stays at that point, which illustrates the pinning effect. Second,

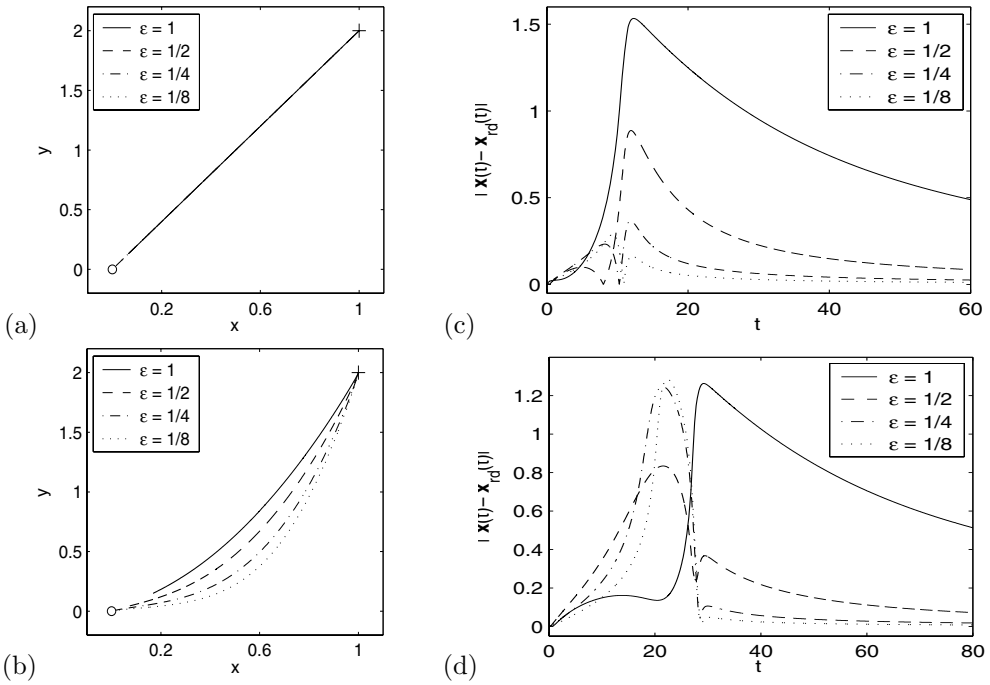


FIG. 27. *Time evolution of the vortex center under an inhomogeneous external driving potential in the GLE.* (a), (b): *Case* I *and* II, *respectively. Trajectory for different* ε. (c), (d): *Errors between the numerical results of the GLE* (*denoted as* $\mathbf{x}(t)$) *and the solution of the reduced dynamic laws* (5.4) (*denoted as* $\mathbf{x}_{\rm rd}(t)$).

for the NLSE, the vortex center moves rotationally clockwise when $m = +1$ and counterclockwise when $m = -1$, to the minimum position of the external potential (cf. Figure 29). The smaller ε, the longer the vortex center stays on a circle. Additional experiments were carried out for Case II. Similar motion patterns were observed, so the results are omitted here.

Based on the numerical results in Figures 28–29, it is reasonable to make the following conjectures about the vortex motion in the NLSE and CGLE: For the NLSE under an inhomogeneous potential, the velocity of the induced motion satisfies (cf. the right-hand side of Figure 29)

$$\mathbf{v}(t) := \frac{d\mathbf{x}(t)}{dt} = -m\kappa\,\mathbf{J}\nabla \ln V(\mathbf{x}(t)), \quad t \geq 0, \qquad \text{with} \quad \mathbf{x}(0) = \mathbf{x}^0, \tag{5.5}$$

where m is the winding number of the vortex, κ is a constant, and $\mathbf{J}$ is the symplectic matrix given in (1.13), while for the CGLE, it can be given by (cf. Figure 28(c),(d))

$$\mathbf{v}(t) := \frac{d\mathbf{x}(t)}{dt} = -\mathbf{Q}\nabla \ln V(\mathbf{x}(t)), \quad t \geq 0, \qquad \text{with} \quad \mathbf{x}(0) = \mathbf{x}^0, \tag{5.6}$$

where the matrix $\mathbf{Q} = m\,\kappa\mathbf{J} + \mathbf{I}$ with κ a constant, and $\mathbf{J}$ and $\mathbf{I}$ are the symplectic matrix in (1.13) and identity matrix, respectively. Their rigorous mathematical justification is not yet available.

6. Conclusion. We have studied the dynamics and interaction of quantized vortices in the Ginzburg–Landau–Schrödinger equation (GLSE) for modeling superconductivity and superfluidity both analytically and numerically. Along the analytical

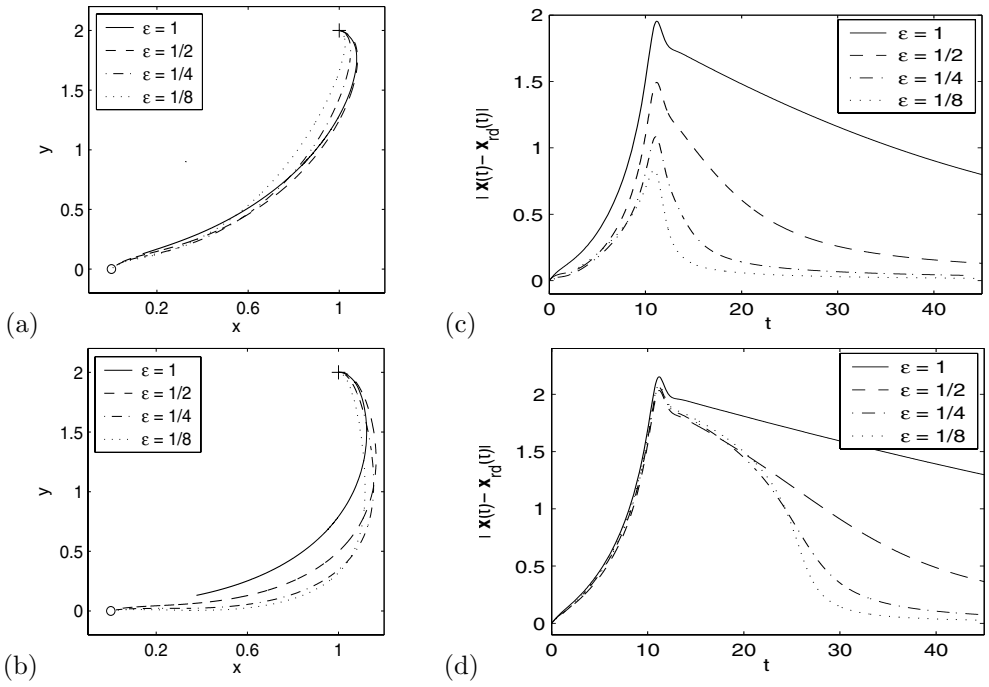


FIG. 28. *Time evolution of the vortex center under an inhomogeneous external driving potential in the CGLE.* (a), (b): *Case* I *and* II, *respectively. Trajectory for different* ε. (c), (d): *Errors between the numerical results of the CGLE* (*denoted as* $\mathbf{x}(t)$) *and the solution of the reduced dynamic laws* (5.6) *with* $\kappa = 1$ (*denoted as* $\mathbf{x}_{\rm rd}(t)$).

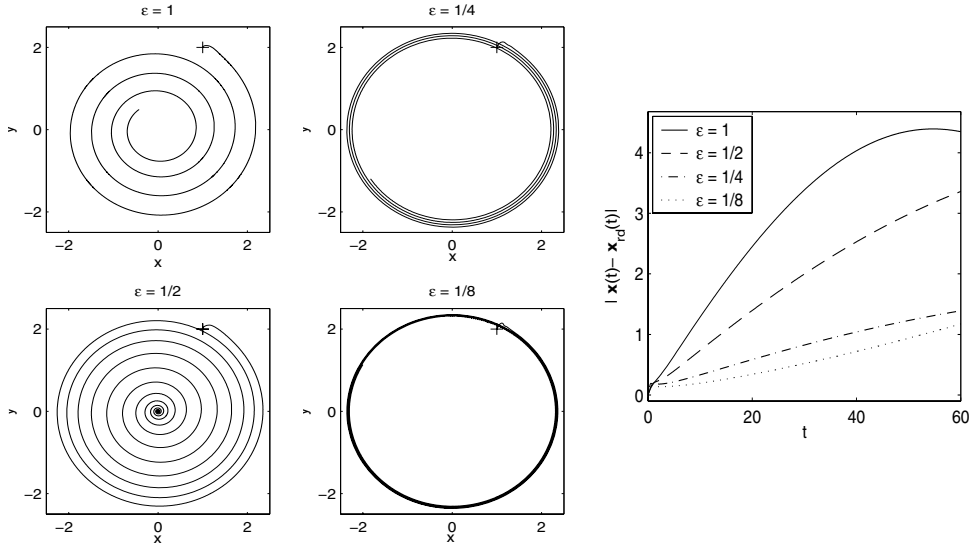


FIG. 29. *Time evolution of the vortex center under an inhomogeneous external driving potential in the NLSE with different* ε *for Case* I. *Left and middle: Trajectory for different* ε. *Right: Errors between the numerical results of the NLSE* (*denoted as* $\mathbf{x}(t)$) *and the solution of the reduced dynamic laws* (5.5) *with* $\kappa = 3$ (*denoted as* $\mathbf{x}_{\rm rd}(t)$).

front, we proved the conservation of the mass center and the signed mass center of N vortex centers governed by the reduced dynamic laws for the Ginzburg–Landau equation (GLE) and nonlinear Schrödinger equation (NLSE), respectively. We also solved analytically the nonlinear ordinary differential equations (ODEs) governing the reduced dynamic laws of the GLE and NLSE for some initial data with symmetrically placed vortices. On the numerical side, by applying an efficient, accurate, and unconditionally stable numerical method for the GLSE with nonzero far-field conditions in two dimensions, we numerically examined issues such as the interaction of vortices and the motion of a vortex under an inhomogeneous external potential in the GLSE. Comparisons between the solutions of the reduced dynamic laws and direct simulation results of the GLSE were provided. Some conclusive findings were obtained, and discussions on numerical and theoretical results were provided for further understanding of vortex interactions in the GLSE. In addition, the vortex motion under an inhomogeneous external potential in the GLSE was investigated numerically for the first time and some conjectures for the motion were made based on our computational findings. In the future, we will extend our efficient and accurate numerical method to the study of the dynamics and the interaction of vortex line states in three dimensions and in bounded domains for the GLSE.

REFERENCES

[1] I. Aranson and L. Kramer, *The world of the complex Ginzburg-Landau equation*, Rev. Modern Phys., 74 (2002), pp. 99–133.

[2] W. Bao, *Numerical methods for the nonlinear Schrödinger equation with nonzero far-field conditions*, Methods Appl. Anal., 11 (2004), pp. 367–387.

[3] W. Bao, Q. Du, and Y. Zhang, *Dynamics of rotating Bose–Einstein condensates and its efficient and accurate numerical computation*, SIAM J. Appl. Math., 66 (2006), pp. 758–786.

[4] W. Bao and Y. Zhang, *Dynamics of the ground state and central vortex states in Bose-Einstein condensation*, Math. Models Methods Appl. Sci., 15 (2005), pp. 1863–1896.

[5] W. Bao and Y. Zhang, *Dynamics of the center of mass in rotating Bose-Einstein condensates*, Appl. Numer. Math., 57 (2007), pp. 697–709.

[6] P. Bauman, C. Chen, D. Phillips, and P. Sternberg, *Vortex annihilation in nonlinear heat flow for Ginzburg-Landau systems*, European J. Appl. Math., 6 (1995), pp. 115–126.

[7] S. J. Chapman and G. Richardson, *Motion of vortices in type* II *superconductors*, SIAM J. Appl. Math., 55 (1995), pp. 1275–1296.

[8] J. E. Colliander and R. L. Jerrard, *Vortex dynamics for the Ginzburg–Landau–Schrödinger equation*, Internat. Math. Res. Notices, 7 (1998), pp. 333–358.

[9] Q. Du, *Finite element methods for the time dependent Ginzburg-Landau model of superconductivity*, Comput. Math. Appl., 27 (1994), pp. 119–133.

[10] Q. Du, *Numerical approximations of the Ginzburg-Landau models for superconductivity*, J. Math. Phys., 46 (2005), 095109.

[11] Q. Du, M. D. Gunzburger, and J. S. Peterson, *Analysis and approximation of the Ginzburg–Landau model of superconductivity*, SIAM Rev., 34 (1992), pp. 54–81.

[12] Q. Du, M. Gunzburger, and J. Peterson, *Computational simulation of type*-II *superconductivity including pinning phenomena*, Phys. Rev. B, 51 (1995), pp. 16194–16203.

[13] Q. Du and W. Zhu, *Stability analysis and application of the exponential time differencing schemes*, J. Comput. Math., 22 (2004), pp. 200–209.

[14] W. E, *Dynamics of vortices in Ginzburg-Landau theories with applications to superconductivity*, Phys. D, 77 (1994), pp. 383–404.

[15] R. Jerrard and H. M. Soner, *Dynamics of Ginzburg-Landau vortices*, Arch. Rational Mech. Anal., 142 (1998), pp. 99–125.

[16] H. Y. Jian, *The dynamical law of Ginzburg-Landau vortices with a pinning effect*, Appl. Math. Lett., 13 (2000), pp. 91–94.

[17] H. Y. Jian and B. H. Song, *Vortex dynamics of Ginzburg-Landau equations in inhomogeneous superconductors*, J. Differential Equations, 170 (2001), pp. 123–141.

[18] O. Lange and B. J. Schroers, *Unstable manifolds and Schrödinger dynamics of Ginzburg-Landau vortices*, Nonlinearity, 15 (2002), pp. 1471–1488.
[19] F.-H. Lin, *Some dynamical properties of Ginzburg-Landau vortices*, Comm. Pure Appl. Math., 49 (1996), pp. 323–359.
[20] F.-H. Lin, *Complex Ginzburg-Landau equations and dynamics of vortices, filaments, and codimension-2 submanifolds*, Comm. Pure Appl. Math., 51 (1998), pp. 385–441.
[21] F.-H. Lin and Q. Du, *Ginzburg–Landau vortices: Dynamics, pinning, and hysteresis*, SIAM J. Math. Anal., 28 (1997), pp. 1265–1293.
[22] F.-H. Lin and J. X. Xin, *On the dynamical law of the Ginzburg-Landau vortices on the plane*, Comm. Pure Appl. Math., 52 (1999), pp. 1189–1212.
[23] F.-H. Lin and J. X. Xin, *On the incompressible fluid limit and the vortex motion law of the nonlinear Schrödinger equation*, Comm. Math. Phys., 200 (1999), pp. 249–274.
[24] F.-H. Lin and J. X. Xin, *A unified approach to vortex motion laws of complex scalar field equations*, Math. Res. Lett., 5 (1998), pp. 455–460.
[25] P. Mironescu, *Les minimiseurs locaux pour l'équation de Ginzburg-Landau sont à symétrie radiale*, C. R. Acad. Sci. Paris Sér. I Math., 323 (1996), pp. 593–598.
[26] P. Mironescu, *On the stability of radial solutions of the Ginzburg-Landau equation*, J. Funct. Anal., 130 (1995), pp. 334–344.
[27] J. C. Neu, *Vortices in complex scalar fields*, Phys. D, 43 (1990), pp. 385–406.
[28] J. C. Neu, *Vortex dynamics of the nonlinear wave equation*, Phys. D, 43 (1990), pp. 407–420.
[29] Y. N. Ovchinnikov and I. M. Sigal, *Ginzburg-Landau equation* I. *Static vortices*, in Partial Differential Equations and Their Applications, CRM Proc. Lecture Notes 12, AMS, Providence, RI, 1997, pp. 199–220.
[30] Y. N. Ovchinnikov and I. M. Sigal, *The Ginzburg-Landau equation* III. *Vortex dynamics*, Nonlinearity, 11 (1998), pp. 1277–1294.
[31] Y. N. Ovchinnikov and I. M. Sigal, *Long-time behavior of Ginzburg-Landau vortices*, Nonlinearity, 11 (1998), pp. 1295–1309.
[32] Y. N. Ovchinnikov and I. M. Sigal, *Asymptotic behavior of solutions of Ginzburg-Landau and related equations*, Rev. Math. Phys., 12 (2000), pp. 287–299.
[33] Y. N. Ovchinnikov and I. M. Sigal, *Symmetry-breaking solutions of the Ginzburg-Landau equation*, J. Exp. Theor. Phys., 99 (2004), pp. 1090–1107.
[34] E. Sandier, *The symmetry of minimizing harmonic maps from a two-dimensional domain to the sphere*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 549–559.
[35] M. I. Weinstein and J. Xin, *Dynamics stability of vortex solutions of Ginzburg-Landau and nonlinear Schrödinger equations*, Comm. Math. Phys., 180 (1996), pp. 389–428.
[36] Y. Zhang, W. Bao, and Q. Du, *Numerical simulation of vortex dynamics in Ginzburg–Landau–Schrödinger equation*, European J. Appl. Math., to appear.

SIAM J. APPL. MATH.
Vol. 67, No. 6, pp. 1776–1807

SMALL- AND WAITING-TIME BEHAVIOR OF THE THIN-FILM EQUATION*

JAMES F. BLOWEY†, JOHN R. KING‡, AND STEPHEN LANGDON§

Abstract. We consider the small-time behavior of interfaces of zero contact angle solutions to the thin-film equation. For a certain class of initial data, through asymptotic analyses, we deduce a wide variety of behavior for the free boundary point. These are supported by extensive numerical simulations.

Key words. thin-film, waiting-time, interface, nonlinear degenerate parabolic

AMS subject classifications. 35R35, 35K35, 35K55, 35K65, 65M60

DOI. 10.1137/060667682

1. Introduction.

This paper is concerned with the small-time behavior of interfaces of zero contact angle solutions to the "thin-film" equation

$$\frac{\partial h}{\partial t} = -\frac{\partial}{\partial x}\left(h^n \frac{\partial^3 h}{\partial x^3}\right), \tag{1.1a}$$

$$\text{with} \quad h = h_0(x) \quad \text{at } t = 0 \quad \text{and} \tag{1.1b}$$

$$h = \frac{\partial h}{\partial x} = h^n \frac{\partial^3 h}{\partial x^3} = 0 \quad \text{at } x = s(t), \tag{1.1c}$$

where $h \geq 0$ represents the thickness of a fluid film and $x = s(t)$ denotes the right-hand interface (with $h \equiv 0$ for $x > s(t)$); since we are concerned with the local behavior at such an interface we need not specify conditions at any left-hand moving boundary. The first boundary condition of (1.1c) defines the moving boundary (as the point at which the film thickness reaches zero), the second ensures a zero contact angle, and the third represents conservation of mass.

In the last few years the range $0 < n \leq 3$ has been considered in the literature from a modeling point of view. With $n = 3$, (1.1a–c) models the lubrication approximation of a surface tension-driven thin viscous film spreading on a solid horizontal surface, with a no-slip condition at the solid/liquid/air interface [5, 6, 10, 11, 12, 14, 34]. However, the no-slip condition implies an infinite force at the interface [19, 27]. To avoid this, more realistic models allowing slip have been proposed (see, e.g., [4, 22, 26]) for which it has been shown that the qualitative behavior of solutions in the vicinity of the interface corresponds to that of the solution of (1.1a–c) with $n \in (0, 3)$; this applies to questions of spreading or nonspreading as well as to questions of locally preserved positivity and local film rupture [17]. We also note that an application of

*Received by the editors August 18, 2006; accepted for publication (in revised form) June 6, 2007; published electronically October 5, 2007.
http://www.siam.org/journals/siap/67-6/66768.html

†Department of Mathematical Sciences, University of Durham, Durham DH1 3LE, UK (j.f.blowey@durham.ac.uk). The work of this author was partially supported by the EPSRC, UK through grant GR/M30951.

‡School of Mathematical Sciences, University of Nottingham University Park, Nottingham NG7 2RD, UK (John.King@nottingham.ac.uk).

§Department of Mathematics, University of Reading, Whiteknights, P.O. Box 220, Berkshire RG6 6AX, UK (s.langdon@reading.ac.uk). The work of this author was partially supported by the EPSRC, UK through grant GR/M30951 and by a Leverhulme Trust Early Career Fellowship.

(1.1a) with $2 < n < 3$ to power-law shear-thickening fluids is derived in [30]. With $n \in (0, 3)$ it is also well known (see, e.g., [7, 8]) that (1.1a–c) admits solutions with a finite speed of propagation property; i.e., $s(t)$ represents a moving boundary, which moves at finite speed.

In this paper we thus consider only values of n in the moving front regime $0 < n < 3$, and we assume further that the film is thick enough that Van der Waals forces play no part. When considering solutions to (1.1a–c), the primary physical question is often to do with the movement of the free boundary. Where $h = 0$ there is no diffusion in (1.1a), and this can lead to waiting-time behavior, where the interface remains stationary for a period before moving; alternatively the interface may either advance or retreat immediately. A determination of the regimes in which such behavior can occur has considerable implications regarding the possibility of film rupture in the presence of a very thin prewetting layer; see, e.g., [31].

There has been much recent effort in the literature to answer outstanding questions about the initial movement of the interface. Theoretical results in [4, 5] have shown that the interface cannot retreat if $n \geq 3/2$, but that film rupture may occur for $n < 1/2$ (see also [13, 14]). Moreover, numerical evidence [4, 10, 12] suggests that for small values of n solutions which are initially strictly positive may vanish at some point x_0 after a finite time t_0, with the solution becoming zero on a set of positive measure shortly after the finite time singularity, a phenomenon called "dead core" in other fields. The existence of a *critical exponent* (a value of $n_* > 0$ for which solutions stay positive for $n > n_*$ and where finite-time singularities are possible for $n \leq n_*$) has been conjectured in [11], where it is remarked that numerical simulations suggest $1 < n_* < 3.5$. Our results below support and clarify these conjectures; in particular, here we provide the first concrete solutions to (1.1a–c) displaying retreat.

As explained in [31], subsequent to any waiting time the local behavior of solutions to (1.1a–c) takes the form

$$h \sim \left(\frac{n^3 \dot{s}}{3(3-n)(2n-3)} (s-x)^3 \right)^{\frac{1}{n}} \quad \text{as } x \to s^- \quad \text{for } \frac{3}{2} < n < 3, \tag{1.2}$$

$$h \sim \left(\frac{3}{4} \dot{s} (s-x)^3 \ln \left[\frac{1}{(s-x)} \right] \right)^{\frac{2}{3}} \quad \text{as } x \to s^- \quad \text{for } n = \frac{3}{2}, \tag{1.3}$$

$$h \sim B(t)(s-x)^2 \quad \text{as } x \to s^- \quad \text{for } n < \frac{3}{2}. \tag{1.4}$$

With $0 < n < 3$, in (1.2) we require that $\dot{s} > 0$, whereas in (1.4) the interface velocity $\dot{s}$ may take either sign, with $B(t)$ determined as part of the solution. One of the key motivations for the current analysis is to provide criteria under which $\dot{s} < 0$ holds for sufficiently small time; since $\dot{s} > 0$ typically holds for large times, for example for the Cauchy problem with initial data of finite mass, a large-time analysis provides no insight into such matters.

For definiteness, we shall consider the case

$$h_0(x) \sim A_0(x_0 - x)^\alpha + C_0(x_0 - x)^\beta \quad \text{as } x \to x_0^-, \tag{1.5}$$

where A_0, α, and β are positive constants with $\beta > \alpha$; C_0 is a constant; and $x_0 = s(0)$.

Extensive studies of the small-time behavior have already been done for the corresponding second-order problem, the porous-medium equation:

$$\frac{\partial h}{\partial t} = \frac{\partial}{\partial x} \left(h^n \frac{\partial h}{\partial x} \right) \tag{1.6}$$

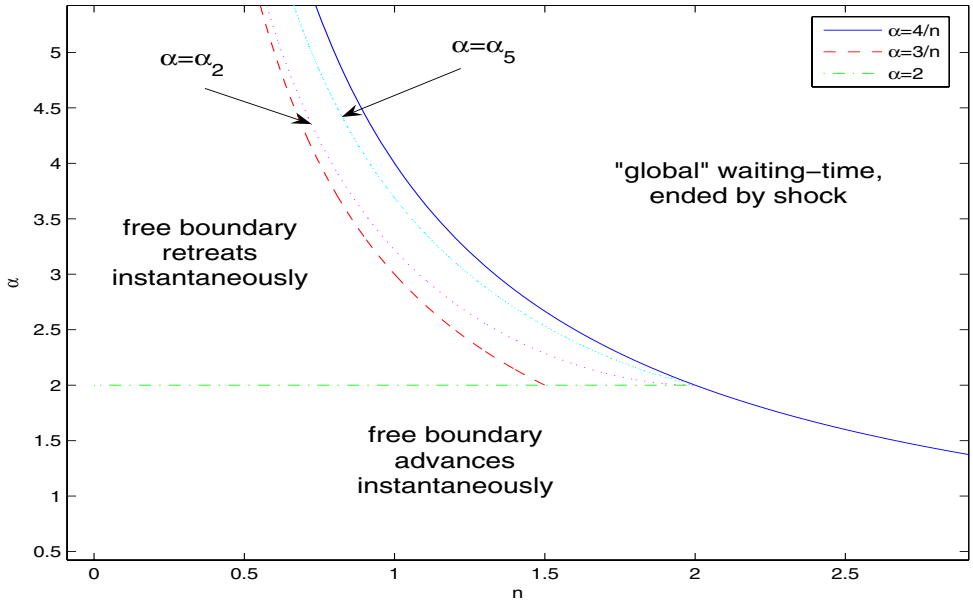


FIG. 1.1. *A summary of the possible small-time behaviors with respect to n and α. By "shock" we mean that a steep front suddenly overruns the interface. In the region* $\max(2, 3/n) < \alpha < 4/n$ *a diverse range of waiting-time scenarios are seen: specifically (see sections* 4.4.1 *and* 4.6.1*) the interface waits, but the local profile changes instantaneously from that of the initial data and can exhibit monotonic (if* $\alpha_5 < \alpha < 4/n$*) or oscillatory (if* $\alpha_2 < \alpha < \alpha_5$*) decay to the local solution, or limit-cycle behavior (if* $\max(2, 3/n) < \alpha < \alpha_2$*).*

with $n > 0$. We present our results in this context. The variety of possible small-time behaviors for (1.1a–c) is summarized in Figure 1.1, and can be characterized as follows:

(i) For α greater than some critical value, the interface "waits" for some finite time t_w, whereby

$$s(t) = x_0 \quad \text{for } 0 \le t \le t_w,$$

after which time it moves. (See also [18, 21, 23] for rigorous studies of such waiting-time phenomena for (1.1a).) For $\alpha = 4/n$ an upper bound on t_w can be deduced from the local behavior of the solution (cf. [33] for the corresponding case (1.6)); more generally, information about t_w can be obtained from the full (global) solution (cf. [32] for (1.6)).

(ii) For α below the critical value, the interface will move at once, with (in view of (1.2)–(1.3))

$$s(t) > x_0 \quad \text{for } t > 0$$

for $3/2 \le n < 3$ (cf. [25] for (1.6)). For $n < 3/2$, however, $\dot{s} < 0$ is also possible, so a further classification is required according to whether $\dot{s} > 0$ or $\dot{s} < 0$ for small $t > 0$. This does not arise in the corresponding analysis of (1.6), since $\dot{s} \ge 0$ necessarily holds.

In addition, the higher order of (1.1a) leads, as we shall see, to a much more diverse range of waiting-time scenarios than that which occurs for (1.6), as shown in Figure 1.1.

The definitions of α_2 and α_5 are rather complicated; for details we refer to section 4.4.1, section 4.6.1, and Appendix A.

In seeking a physical explanation for these results, we remark that larger n implies weaker slip, and large α a shallow initial "contact angle." Broadly speaking, the larger the value of n/α, the stronger is the tendency of solutions to stay positive. The current phenomena are associated with perfectly wetting (zero-contact-angle) boundary conditions and should not be confused with those associated with finite static contact angles. In the latter (i.e., partially wetting) case, for viscous fluids with an initial condition characterized by a contact angle sufficiently greater (respectively, less) than equilibrium, the droplet tends to spread (contract) with no waiting. For intermediate contact angles, waiting-time behavior associated with contact-angle hysteresis can occur. Although such behavior has some similarities with that described below (in particular, retreating contact lines are associated with initial data that are "smaller" than advancing ones), there are also important differences, notably that waiting-time behavior is in general associated with the "smallest" initial data.

We are not aware of any experimental evidence to support our conjectures, but in light of our results such experiments might be timely. For a discussion of the physical length scales pertinent to the slip-dominated ($n = 2$) model, see, for example, [20], and also references therein regarding such strong slip conditions. (We note that this paper also includes an additional term, not present in the thin-film equation, that is relevant for slip lengths even longer than those for which (1.1a–c) applies with $n = 2$.) Instead, we support our asymptotic conjectures with numerical results. Without loss of generality we assume that $s(0) > 0$ and, for numerical purposes, we first approximate (1.1a–c) by replacing (1.1c) by

$$\frac{\partial h}{\partial x} = h^n \frac{\partial^3 h}{\partial x^3} = 0 \quad \text{for } x = 0, l, \tag{1.7c}$$

where $l \gg s(0)$, and restrict (1.1a) to hold on $(0, l)$. Existence of solution concepts for (1.1a,b), (1.7c) may be found in [5, 9, 14] and the references cited therein.

As described in [2], we discretize (1.1a,b), (1.7c) using finite elements in space and finite differences in time, using uniform spatial and temporal discretization parameters δx and δt, respectively; see section 2 for details. We expect that this method will be able to compute the zero contact angle solution for the following reasons:

1. In [5], the existence of solutions to (1.1a,b), (1.7c) is proved for $0 < n < 3$, where $h(\cdot, t)$ may be $C^1([0, l])$ for almost every $t > 0$ (the zero contact angle solution), or alternatively $h(\cdot, t)$ may have nonexpansive support.
2. In [2] it was proved that the numerical solution converges, as $\delta x, \delta t \to 0$, to a weak solution of (1.1a,b), (1.7c) (in the sense of [5, 9, 14]). The only remaining question is whether this is the zero contact angle solution or a solution with nonexpansive support.
3. In a sequence of experiments, taking $\delta t = O(\delta x^{\frac{1}{2}})$, the numerical method computes a solution with nonexpansive support.
4. In a sequence of experiments, taking $\delta t = O(\delta x^2)$, the numerical method can compute solutions where $|\dot{s}(0)| = \infty$ (zero contact angle solutions).
5. In [2] a self-similar source type solution was successfully computed with $\delta t = O(\delta x^2)$. Moreover, taking a nonsmooth stationary solution as initial data, i.e., $h_0(x) = \alpha \max\{\gamma^2 - x^2, 0\}$ and $0 < \gamma < l$, the numerical method computed a smooth solution for $0 < n < 3$, and it was concluded that $h(x, t) \equiv h_0(x)$ for $n > 3$.

Hence in our experiments, in order to be sure that we are approximating the zero contact angle solution we always choose $\delta t = O(\delta x^2)$. We report that the numerical solution always appears to be smooth.

An outline of the paper is as follows. We begin in section 2 by describing our numerical scheme in more detail. We then proceed in sections 3 and 4 with a formal asymptotic analysis, supported by numerical experiments, for the two cases $\alpha \geq 4/n$ and $\alpha < 4/n$, respectively. Videos demonstrating more graphically how some of the numerical solutions of these sections evolve over time can be found online at http://www.personal.rdg.ac.uk/∼sms03sl/4thorder/4thorder.html. Finally, in section 5 we present some conclusions.

2. Numerical approximation. Following the approach of [2], and as described in section 1, we restrict (1.1a) to a finite space interval $(0, l)$, introduce a potential w, and rewrite it as the system of equations

$$\frac{\partial h}{\partial t} = \frac{\partial}{\partial x}\left(h^n \frac{\partial w}{\partial x}\right) \quad \text{in } (0,l)\times(0,T), \tag{2.1a}$$

$$-\frac{\partial^2 h}{\partial x^2} = w \quad \text{in } (0,l)\times(0,T). \tag{2.1b}$$

A nonnegativity constraint is imposed on (2.1b) via a variational inequality in the weak form, and then we discretize (2.1a,b) using the finite element method. Now, given positive integers N and M, denote by $\delta t := T/M$ and $\delta x := l/N$ the temporal and spatial discretization parameters, $t_k := k\delta t$, $k = 1, \ldots, M$, and $x_j = j\delta x$, $j = 0, \ldots, N$; then the discretization may be written in the following way.

For $k = 1, \ldots, M$ and $j = 1, \ldots, N-1$ find $\{H_j^{k+1}, W_j^{k+1}\}$ such that

$$\frac{H_j^{k+1} - H_j^k}{\delta t} + \frac{1}{\delta x^2}\left[\int_{x_{j-1}}^{x_j} \left(\frac{(x - x_{j-1})}{\delta x} H_j^k + \frac{(x_j - x)}{\delta x} H_{j-1}^k\right)^n \mathrm{d}x\right]\left(\frac{W_j^{k+1} - W_{j-1}^{k+1}}{\delta x}\right)$$

$$+ \frac{1}{\delta x^2}\left[\int_{x_j}^{x_{j+1}} \left(\frac{(x - x_j)}{\delta x} H_{j+1}^k + \frac{(x_{j+1} - x)}{\delta x} H_j^k\right)^n \mathrm{d}x\right]\left(\frac{W_j^{k+1} - W_{j+1}^{k+1}}{\delta x}\right) = 0, \tag{2.2a}$$

$$\left[\frac{-H_{j+1}^{k+1} + 2H_j^{k+1} - H_{j-1}^{k+1}}{\delta x^2} - W_j^{k+1}\right] H_j^{k+1} = 0, \tag{2.2b}$$

$$\frac{-H_{j+1}^{k+1} + 2H_j^{k+1} - H_{j-1}^{k+1}}{\delta x^2} - W_j^{k+1} \geq 0, \tag{2.2c}$$

$$H_j^{k+1} \geq 0, \tag{2.2d}$$

where $H_j^k \approx h(x_j, t_k)$, $W_j^k \approx w(x_j, t_k)$, $H_j^0 = h_0(x_j)$; similar equations/inequalities appropriate for boundary data (1.7c) hold for $j = 0, N$ when $k = 1, \ldots, M$. This nonlinear system is solved using a Gauss–Seidel algorithm in multigrid mode; for details we refer to [3]. We found this approach to have several advantages over some other algorithms previously proposed in the literature, such as the Uzawa-type algorithm [2, (3.7a–c)], [24]. Specifically, we find the following:

1. If $H_{j-1}^k = H_j^k = H_{j+1}^k = 0$, then it follows from (2.2a) that $H_j^{k+1} = H_j^k = 0$, so that the free boundary advances at most one mesh point from time level k

to time level $k+1$. The advantage of using the nonsymmetric Gauss–Seidel smoother is that this constraint is easier to impose on the numerical method than with a symmetric smoother.

2. Working within a multigrid framework significantly increases the rate of convergence. This allows us to reduce the tolerance for the stopping criterion of the iterative scheme (the maximum absolute difference in successive iterates is smaller than tol) to $tol = 10^{-12}$, as compared with $tol = 10^{-8}$ in [2], and therefore to solve the nonlinear system more accurately, thereby helping to avoid spurious behavior.
3. Nonnegativity of the computed numerical solution is guaranteed, and so defining the position x_c^k of the numerical free boundary at time t_k to be

$$x_c^k := \{x_j > 0 : H_m^k \le \epsilon \text{ for all } m \ge j,\ H_{j-1}^k > \epsilon\},$$

we take $\epsilon = 0$, which tracks the free boundary more accurately than with $\epsilon > 0$; this compares with $\epsilon = 10^{-6}$ in [2]. We remark that because the numerical free boundary is defined on a discrete set of points, its movement appears to "stutter" in the figures below. Although the interface always advances or retreats with a stepping motion, oscillations are seen only in certain cases. Moreover, it is sometimes the case that the oscillations in H_j^k begin and grow before the contact line moves; hence they do not appear to be caused by this "stuttering."

In the numerical experiments of sections 3 and 4 we solve (1.1a,b), (1.7c) with $l = 1$ and

$$h_0(x) = 5 \max\left\{\left(\tfrac{9}{16} - x^2\right)^{\alpha}, 0\right\}; \tag{2.3}$$

the key properties are that the maximum value of h_0 is $O(1)$ and the thin film is symmetrically distributed about 0 with $x_0 = 3/4$. These experiments were performed for a sequence of space steps δx, where $\delta t = C_{\alpha,n}\delta x^2$ and convergence of (2.2a)–(2.2d) to a weak solution of (1.1a,b), (1.7c), (2.3) was assured (see [2]). For reasons of space we refer to [15] for further figures and numerical results, including results from many more experiments with values of n and α closer to the edges of the parameter regimes.

3. Formal asymptotic analysis and numerical results for $\alpha \ge 4/n$. For $\alpha \ge 4/n$, the formal asymptotic analysis of this section suggests that we might expect waiting-time behavior. In particular, for $\alpha > 4/n$ (section 3.1) we anticipate "global" waiting-time behavior, by which we mean that the asymptotic expansion tells us to expect a waiting time but gives no clue as to the local behavior; in this case the interface starts to move due to shock formation, with the gradient becoming unbounded near the free boundary. For $\alpha = 4/n$ (section 3.2) this global breakdown can occur for the full range $0 < n < 3$, but for $2 < n < 3$ "local" waiting-time behavior is also possible; by this we mean that the dominant term in the asymptotic expansion switches at the end of the waiting time.

3.1. $\alpha > 4/n$: Global waiting-time behavior. Provided that $h_0(x)$ is analytic away from the interfaces, then the small-time expansion

$$h \sim h_0 - \frac{\mathrm{d}}{\mathrm{d}x}\left(h_0^3 \frac{\mathrm{d}^3 h_0}{\mathrm{d}x^3}\right) t \quad \text{as } t \to 0^+ \tag{3.1}$$

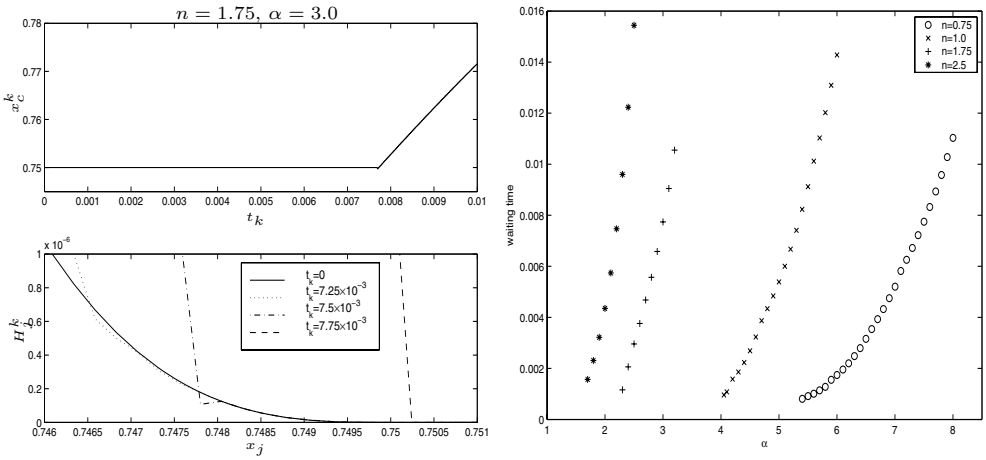


FIG. 3.1. *Waiting-time behavior, various* n, α.

holds, at least away from the interface. From (1.5) we have

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(h_0^n \frac{\mathrm{d}^3 h_0}{\mathrm{d}x^3}\right) \sim \alpha(\alpha-1)(\alpha-2)((n+1)\alpha-3)A_0^{n+1}(x_0-x)^{(n+1)\alpha-4} + \left(n\alpha(\alpha-1)(\alpha-2)+\beta(\beta-1)(\beta-2)\right)(n\alpha+\beta-3)A_0^n C_0 (x_0-x)^{n\alpha+\beta-4} \quad \text{as } x \to x_0^-; \tag{3.2}$$

for $\alpha > 4/n$ we have $(n+1)\alpha - 4 > \alpha$, and we may expect the local behavior

$$h \sim A_0(x_0-x)^\alpha \quad \text{as } x \to x_0^- \tag{3.3}$$

to hold up to some finite time $t = t_w > 0$, implying a waiting-time scenario in which the local behavior at the interface does not change for some nonzero waiting time.

To test this conjecture, we ran numerical experiments for a large range of n and $\alpha > 4/n$, considering in particular $n = 0.75$, 1.0, 1.75, and 2.5, so as to cover all of the different regimes important in the case $\alpha < 4/n$ (see section 4).

In the upper left panel of Figure 3.1 we plot x_c^k against t_k for $n = 1.75$ and $\alpha = 3.0 > 4/n = 2.29$. The numerical free boundary remains stationary for a period before advancing. We also plot profiles of H_j^k in the vicinity of the interface at times just before and just after x_c^k begins to move (lower left panel). Shock-type behavior at the end of the waiting time can be observed (cf. [32] for the second-order case).

Similar waiting-time behavior, with shock type behavior at the end of the waiting time, was observed for all n, α combinations tested in this range. Approximate waiting times are plotted against α in the right half of Figure 3.1, for $n = 0.75, 1.0, 1.75, 2.5$ and for various $\alpha > 4/n$. For fixed n, the waiting time increases as α increases.

3.2. $\alpha = 4/n$: Local waiting-time behavior for $2 < n < 3$. In the critical case, the leading term in (1.5) suggests the separable local behavior

$$h \sim \Lambda(t)(x_0-x)^{4/n} \quad \text{as } x \to x_0^-, \tag{3.4}$$

with (1.1a–c) implying

$$\dot{\Lambda} = -\tfrac{4}{n}\left(\tfrac{4}{n}-1\right)\left(\tfrac{4}{n}-2\right)\left(\tfrac{4}{n}+1\right)\Lambda^{n+1}.$$

Hence if $n \neq 2$ (recalling that we consider in this paper only $0 < n < 3$), then

$$\Lambda = A_0 \left(1 + \frac{8(4-n)(2-n)(n+4)A_0^n t}{n^3}\right)^{-\frac{1}{n}}. \tag{3.5}$$

This local solution also represents waiting-time behavior; (3.5) blows up in finite time if $2 < n < 3$, so the waiting time t_w then satisfies

$$t_w \leq t_c \equiv \frac{n^3}{8(4-n)(n-2)(n+4)A_0^n};$$

$\Lambda(t)$ decreases with time for $0 < n < 2$, but we nevertheless expect (3.4) to remain valid only up to some finite t_w, after which the front begins to move due to shock, as described in section 3.1. Thus for $2 < n < 3$ local waiting-time behavior ($t_w = t_c$) is possible, analogous to that for the porous-medium equation [33], while global breakdown ($t_w < t_c$ for $2 < n < 3$) can occur for the full range $0 < n < 3$ (cf. [32]).

4. Formal asymptotic analysis and numerical results for $\alpha < 4/n$. We begin in section 4.1 by deriving some local similarity solutions. Based on these and the local behavior indicated in (1.2)–(1.4), we conjecture in sections 4.2–4.6 some parameter regimes for the small-time behavior when $\alpha < 4/n$, in which case (3.3) fails for any $t > 0$. This does not mean that for $\alpha < 4/n$ there is no waiting-time behavior; on the contrary, unlike for the corresponding second-order problem (1.6), a diverse range of waiting-time scenarios can occur in this case. In addition to these waiting-time scenarios, the front may also advance or retreat instantaneously. Many of our conjectures are supported by extensive numerical verifications, detailed below; we leave open their rigorous confirmation.

4.1. Local similarity solutions. In view of (1.5), a natural conjecture for the small-time behavior for $\alpha < 4/n$ (balancing the terms in the expansion so that they are of the same size) is the self-similar form

$$h \sim t^{\frac{\alpha}{4-n\alpha}} f\left((x-x_0)/t^{\frac{1}{4-n\alpha}}\right), \quad s(t) \sim x_0 + \eta_0 t^{\frac{1}{4-n\alpha}}, \tag{4.1}$$

where with $\eta := (x-x_0)/t^{1/(4-n\alpha)}$, $f(\eta)$ satisfies the boundary-value problem

$$\frac{1}{4-n\alpha}\left(\alpha f - \eta \frac{\mathrm{d}f}{\mathrm{d}\eta}\right) = -\frac{\mathrm{d}}{\mathrm{d}\eta}\left(f^n \frac{\mathrm{d}^3 f}{\mathrm{d}\eta^3}\right), \tag{4.2}$$

$$\text{(4.3a) as } \eta \to -\infty,\ f \sim A_0(-\eta)^\alpha - \alpha(\alpha-1)(\alpha-2)((n+1)\alpha-3)A_0^{n+1}(-\eta)^{(n+1)\alpha-4},$$

$$\text{at } \eta = \eta_0, \quad f = \frac{\mathrm{d}f}{\mathrm{d}\eta} = f^n \frac{\mathrm{d}^3 f}{\mathrm{d}\eta^3} = 0. \tag{4.3b}$$

Here $s(t)$ is the position of the interface at time t, and η_0 is a free constant determined by the boundary-value problem.

The behavior as $\eta \to -\infty$ in (4.3a) thereby matches via (3.1) with the leading terms in (1.5) and (3.2). The constant A_0 can be removed via the change of variables

$$f = A_0^{\frac{4}{4-n\alpha}} \hat{f}, \quad \eta = A_0^{\frac{n}{4-n\alpha}} \hat{\eta},$$

suggesting in particular the delicacy of the limit $\alpha \to (4/n)^-$, and the transformation

$$f = |\eta|^{\frac{4}{n}} g(\xi), \quad \xi = \ln|\eta|$$

enables (4.2) to be reduced to a fourth-order autonomous problem. Nevertheless, the complexities of the resulting four-dimensional phase space mean that a global analysis of (4.2) (akin to that in [33] for the second-order problem) is not practicable here. Instead we base our conjectures in large part on a number of closed-form solutions to (4.2), which we now note. We assume (4.2)–(4.3a,b) to have a unique nonnegative solution.

(I) Separable solution

$$f(\eta) = \left(\frac{n^3}{8(4-n)(2-n)(n+4)} (-\eta)^4 \right)^{\frac{1}{n}} \tag{4.4}$$

is an explicit solution to (4.2) for $0 < n < 2$, providing a possible local behavior as $\eta \to 0^-$ for solutions with $\eta_0 = 0$; the circumstances under which (4.4) may be applicable are clarified in Appendix A.

(II) Steady-state solution

$$f(\eta) = A_0(-\eta)^2, \quad \eta_0 = 0, \tag{4.5}$$

gives the solution to (4.2)–(4.3a,b) when $\alpha = 2$.

(III) Traveling wave solution

$$f(\eta) = A_0(\eta_0 - \eta)^{\frac{3}{n}}, \quad \eta_0 = \frac{3(3-n)(2n-3)A_0^n}{n^3}, \tag{4.6}$$

is the solution to (4.2)–(4.3a,b) when $\alpha = 3/n$, $n \neq 3/2$; here $\eta_0 > 0$ if $3/2 < n < 3$, and $\eta_0 < 0$ if $0 < n < 3/2$.

To complete our catalogue of pertinent closed-form solutions we note that for $n = 1$ the polynomial solution (cf. [28])

$$\begin{aligned} h &= \frac{A_0^3}{4(1+30C_0^2t/A_0)C_0^2} \left((1 + C_0(x_0 - x)/A_0)^2 - (1 + 30C_0^2t/A_0)^{\frac{2}{5}} \right)^2, \\ s &= x_0 - A_0 \left((1 + 30C_0^2t/A_0)^{\frac{1}{5}} - 1 \right) / C_0, \end{aligned} \tag{4.7}$$

corresponds to

$$h_0 = A_0(x_0 - x)^2 + C_0(x_0 - x)^3 + C_0^2(x_0 - x)^4/(4A_0), \tag{4.8}$$

so that $\alpha = 2$, $\beta = 3$ in (1.5); hence s initially decreases if $C_0 > 0$ but increases if $C_0 < 0$, with

$$s(t) \sim x_0 - 15C_0t \quad \text{as } t \to 0^+, \tag{4.9}$$

this dependence on the sign of C_0 being perhaps counter intuitive, which is far from unusual in such high-order diffusion problems.

4.2. Small-time behavior for $2 < n < 3$. In this regime, the solution (4.4) is not available to describe the local behavior of $f(\eta)$ at the interface; (4.6) has the expected local behavior (1.2), while (4.5) corresponds to $\alpha > 4/n$ and therefore lies in the waiting-time regime discussed in section 3.1. We thus anticipate that for any $\alpha < 4/n$ the support of h expands immediately according to (4.1) with $\eta_0 > 0$ and, in (4.2)–(4.3a,b),

$$f(\eta) \sim \left(\frac{n^3\eta_0}{3(3-n)(2n-3)(4-n\alpha)} (\eta_0 - \eta)^3 \right)^{\frac{1}{n}} \quad \text{as } \eta \to \eta_0^-, \tag{4.10}$$

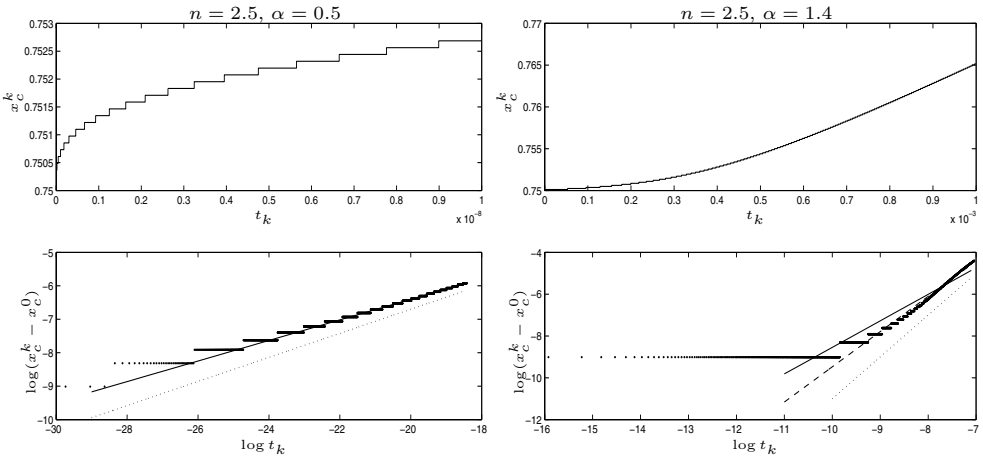


FIG. 4.1. *Numerical results for* $n = 2.5$, $\alpha = 0.5$, $T = 10^{-8}$ *(left), and for* $n = 2.5$, $\alpha = 1.4$, $T = 10^{-3}$ *(right). In the top panels the advancing free boundary is shown. In the bottom panels* $\log t_k$ *is plotted against* $\log(x_c^k - x_c^0)$ *as a discrete set of points, with the solid line following from a least squares fitting, the straight dotted line from asymptotic theory, and the dashed line in the lower right section from a least squares fitting with the early data removed.*

which follows from (1.2). The interface advances with unbounded initial velocity for $\alpha < 3/n$, with finite positive initial velocity if $\alpha = 3/n$ (with $f(\eta)$ given by (4.6)), and with velocity tending to zero as $t \to 0^+$ for $\alpha > 3/n$. The behavior in this regime is very much analogous to that exhibited by the porous-medium equation (cf. [25]).

To test this conjecture we ran numerical experiments for $n = 2.5$, for which $3/n = 1.2$ and $4/n = 1.6$, and for $\alpha \in [0.5, 1.5]$. Our results support the conjecture. In each case x_c^k advances, with the speed of the advance decreasing as α increases from 0.5 to 1.5. We plot x_c^k against t_k for $n = 2.5$ and for $\alpha = 0.5 < 3/n$ and $\alpha = 1.4 > 3/n$ in the upper half of Figure 4.1. Note the different time scales on the two plots.

In the lower half of Figure 4.1 we test the hypothesis that for small times

$$x_c^k = x_c^0 + At_k^\gamma, \tag{4.11}$$

for some constants $A > 0$ and γ, by plotting $\log(x_c^k - x_c^0)$ against $\log t_k$ (as a discrete set of points—these appear to "stutter" since the numerical free boundary advances by one discrete mesh point at a time). If the hypothesis is correct, we expect a straight line with slope γ. To estimate the value of γ we take a least squares fit. For presentational purposes we plot the best fitting least squares line as a solid line, and for comparison we also plot a dotted line with slope $(4 - n\alpha)^{-1}$, the expected value of γ (recall (4.1)).

For $\alpha = 0.5$ the log-log plot is fairly straight, and the estimated value of $\gamma = 0.31$ is close to the expected value of 0.36. For $\alpha = 1.4$ the best fitting least squares line gives an estimate of $\gamma = 1.27$, which is not close to the expected value of 2.00 and is a poor fit to the data. In this case the immediate yet slow advance of the free boundary means that, for t_k small, x_c^k overestimates the exact position of the free boundary. This is demonstrated by the fact that the lowest horizontal line of dots on the log-log plot, corresponding to the first step in the advance of x_c^k, matches very poorly with the rest of the data. In the lower right plot of Figure 4.1 we thus also show as a dashed line the best fitting least squares approximation to the data with the first step

TABLE 4.1
Estimated and expected values of γ for $n = 2.5$, various α.

α	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
$\frac{1}{4-n\alpha}$	0.36	0.40	0.44	0.50	0.57	0.67	0.80	1.00	1.33	2.00	4.00
γ	0.31	0.35	0.39	0.46	0.53	0.61	0.76	0.91	1.12	1.69	3.43

in the advance of x_c^k excluded (equivalently, taking $t_k \gtrapprox 5 \times 10^{-5}$ rather than $t_k > 0$ on the log-log plot). This dashed line, with a slope of 1.69, matches the slope of the data and the expected value of γ much more closely than our original estimate.

The expected and estimated values of γ for each value of α tested are shown in Table 4.1. For $\alpha \leq 1.3$ we estimate γ using all of the data, but for $\alpha = 1.4$ and $\alpha = 1.5$ we exclude the first step in the advance of x_c^k, as discussed above. The numerical results give a value of γ slightly lower than the expected value, but the difference is small, and the trend of γ increasing with α is clear. Our estimate for γ is more accurate for values of α away from the edges of the parameter regime.

4.3. Small-time behavior for $n = 2$. The behavior for $\alpha < 2$ is as described in section 4.2. However, for $\alpha = 2$ the solution (4.1) is not applicable and, partly because $\alpha = 2$ will also play an important role in what follows, additional comments regarding the resulting waiting-time scenario are instructive. The small-time solution (4.5) suggests that there is initially no change in local behavior, while (3.2) becomes

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(h_0^n \frac{\mathrm{d}^3 h_0}{\mathrm{d}x^3}\right) \sim (\beta+1)\beta(\beta-1)(\beta-2)A_0^2 C_0 (x_0-x)^\beta t; \tag{4.12}$$

both suggest seeking a local solution of the form

$$h \sim A_0(x_0-x)^2 + H(x,t) \quad \text{as } x \to x_0^-; \tag{4.13}$$

we note that H need not be positive on $x < x_0$ because it represents a correction term to the (quadratic) leading order behavior. Linearizing in H yields

$$\frac{\partial H}{\partial t} = -A_0^2 \frac{\partial}{\partial x}\left((x_0-x)^4 \frac{\partial^3 H}{\partial x^3}\right),$$

and so, given (1.5) in which $\beta > 2$ is required, the correction term takes the separable form

$$H = C_0(x_0-x)^\beta \exp\left(-(\beta+1)\beta(\beta-1)(\beta-2)A_0^2 t\right),$$

consistent with (3.1), (4.12). The perturbation to the quadratic term thus decays exponentially, and we expect (4.13) to persist up to some finite waiting time, after which the interface will start to move due to shock formation (as in other global waiting-time cases described here; see [32] for the second-order analogue).

4.4. Small-time behavior for $3/2 < n < 2$. In this case (4.6) again has the expected interface behavior (1.2), while (4.5) is nongeneric in the sense that it is smoother than (1.2) at the interface; the solution of (4.2)–(4.3a,b) for $\alpha = 2$ is therefore an exceptional connection in phase space and can be expected to play a role in separating distinct regimes, as we now suggest. The other noteworthy change to occur as n drops below two is that the local behavior (4.4) can come into play.

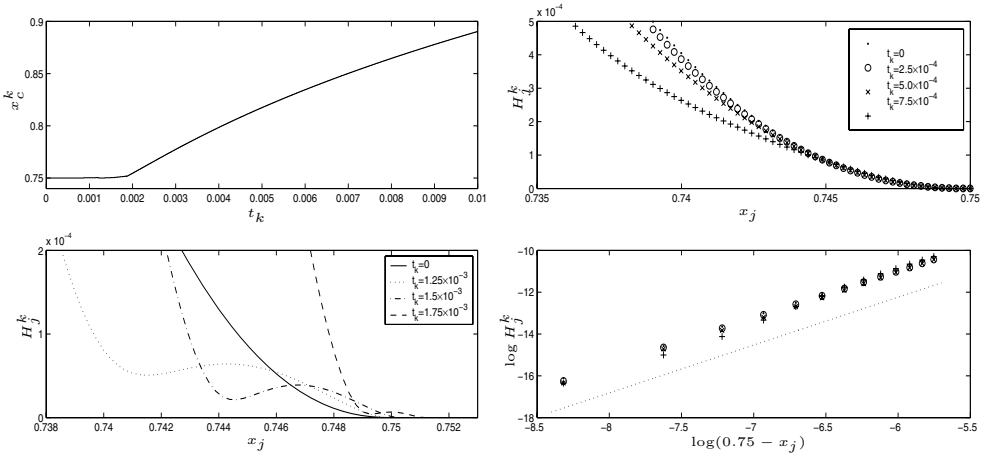


FIG. 4.2. *Numerical results for $n = 1.75$, $\alpha = 2.24$: waiting-time behavior (upper left plot); profiles of H_j^k near the interface while the free boundary is stationary (upper right plot), and as the free boundary advances (lower left plot); $\log H_j^k$ against $\log(0.75 - x_j)$ in the vicinity of the free boundary, with a dotted line from asymptotic theory (lower right plot—same legend as upper right).*

4.4.1. $2 < \alpha < 4/n$. The solution to (4.2)–(4.3a,b) has the local behavior which decays as $(-\eta)^{4/n}$ as $\eta \to 0^-$ and exhibits a finite waiting time; for $\alpha_2 < \alpha < 4/n$, where α_i, $i = 1, 2, 5$ are defined in Appendix A, $f(\eta)$ has local behavior (4.4), so the solution decreases such that

$$h \sim \left(\frac{n^3 (x_0 - x)^4}{8(4-n)(2-n)(n+4)t} \right)^{1/n} \quad \text{as } x \to x_0^-,\ 0 < t < t_w, \tag{4.14}$$

for the duration of the period of waiting. Moreover, for $\alpha_5 < \alpha < 4/n$ we expect nonoscillatory decay, whereas for $\alpha_2 < \alpha < \alpha_5$ we expect damped oscillations to occur. See Appendix A for details. For $2 < \alpha < \alpha_2$ the behavior is slightly more subtle, with a limit cycle (see (A.5) of Appendix A) arising in the local description for $0 < t < t_w$; the limiting behavior as $\alpha \to 2$ is addressed in Appendix B, providing additional support for conjectures about the (rather subtle) asymptotic behavior.

We present numerical results for $n = 1.75$ (giving $3/n = 1.7143$, $\alpha_2(n) = 2.0768$, $\alpha_5(n) = 2.2$, $4/n = 2.2857$), and for $\alpha = 2.24$, 2.10 and 2.04, thus covering each of the three parameter regimes described above. In the upper left plots of Figures 4.2, 4.3, and 4.4, we plot x_c^k against t_k for $n = 1.75$ and $\alpha = 2.24$, 2.10, and 2.04, respectively. In each case x_c^k remains stationary for a period before advancing, with the length of the waiting period appearing to decrease as α decreases. We also plot in each figure profiles of H_j^k near the interface at various times while the free boundary is stationary (upper right plot) and just as the free boundary is beginning to advance (lower left plot). In each case as $x \to x_0^-$ the profile of H_j^k appears to remain unchanged for a short waiting period. In the lower right plot of each figure we plot $\log H_j^k$ against $\log(0.75 - x_j)$ in the vicinity of the free boundary at the same times and using the same legend as in the upper right plot of each figure, plotting also a dotted line with slope $4/n$ for comparison.

For $\alpha = 2.24$ (Figure 4.2) the (nearly) straight lines with slopes 2.26 for $t_k = 2.5 \times 10^{-4}$ and 2.29 for $t_k = 5.0 \times 10^{-4}$ (estimated as before) compare well with the value of $4/n = 2.29$ proposed in the conjecture. For $t_k = 7.5 \times 10^{-4}$ the log-log plot

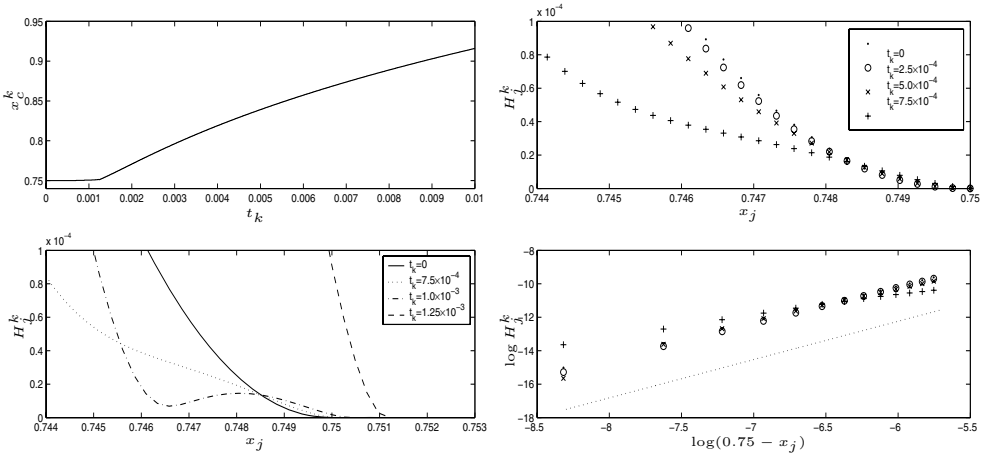


FIG. 4.3. *Numerical results for $n = 1.75$, $\alpha = 2.10$: waiting-time behavior (upper left plot); profiles of H_j^k near the interface while the free boundary is stationary (upper right plot), and as the free boundary advances (lower left plot); $\log H_j^k$ against $\log(0.75 - x_j)$ in the vicinity of the free boundary, with a dotted line from asymptotic theory (lower right plot—same legend as upper right).*

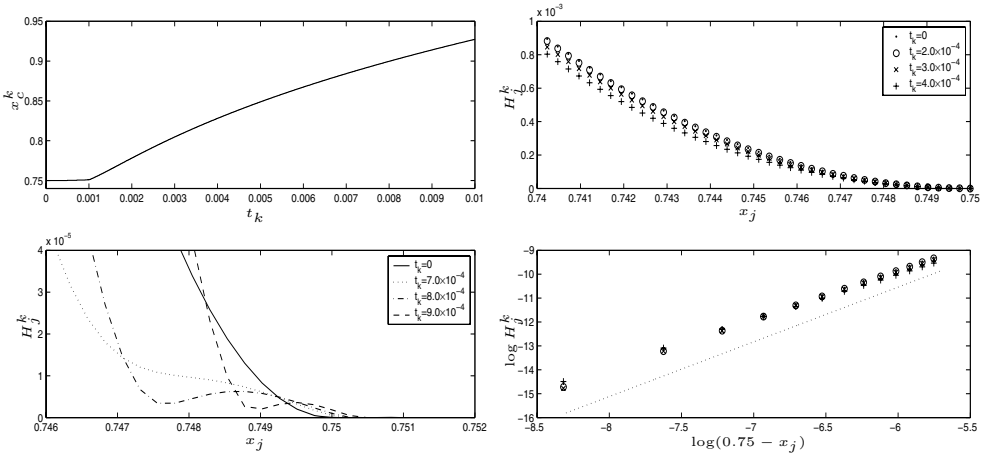


FIG. 4.4. *Numerical results for $n = 1.75$, $\alpha = 2.04$: waiting-time behavior (upper left plot); profiles of H_j^k near the interface while the free boundary is stationary (upper right plot), and as the free boundary advances (lower left plot); $\log H_j^k$ against $\log(0.75 - x_j)$ in the vicinity of the free boundary, with a dotted line from asymptotic theory (lower right plot—same legend as upper right).*

is no longer straight, and the best fitting least squares line has a slope of 2.47; by this time the profile of H_j^k has begun to change.

For $\alpha = 2.10$ (Figure 4.3), the (nearly) straight lines have slopes 2.17 for $t_k = 2.5 \times 10^{-4}$ and 2.18 for $t_k = 5.0 \times 10^{-4}$. These values are slightly lower than for $\alpha = 2.24$, but still compare fairly well with the value of $4/n = 2.29$ proposed in the conjecture. For $t_k = 7.5 \times 10^{-4}$ again the log-log plot is no longer straight, and the best fitting least squares line has a slope of 1.27; by this time the profile of H_j^k has again begun to change.

For $\alpha = 2.04$ (Figure 4.4) each log-log plot is again (nearly) a straight line;

however, the slopes of 2.09 for $t_k = 2.0 \times 10^{-4}$ and 2.08 for $t_k = 3.0 \times 10^{-4}$ are somewhat smaller than the value of $4/n = 2.29$. As t_k increases from zero, the slope of the log-log plot increases from 2.04 up to a maximum of 2.09 before decreasing. For $t_k = 4.0 \times 10^{-4}$ the line has a slope of 1.91; by this time the profile of H_j^k has begun to change noticeably.

4.4.2. $\alpha = 2$. This is the most delicate case, with the small-time behavior depending on the correction term in (1.5), with $\beta > 2$. In (3.2) we have

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(h_0^n \frac{\mathrm{d}^3 h_0}{\mathrm{d}x^3}\right) \sim \beta(\beta-1)(\beta-2)(2n+\beta-3)A_0^n C_0 (x_0-x)^{2n+\beta-4}, \tag{4.15}$$

and (4.13) yields

$$\frac{\partial H}{\partial t} = -A_0^n \frac{\partial}{\partial x}\left((x_0-x)^{2n}\frac{\partial^3 H}{\partial x^3}\right), \tag{4.16}$$

implying, in view of (4.15), the small-time behavior

$$H = A_0^{\frac{n\beta}{4-2n}} C_0 t^{\frac{\beta}{4-2n}} \Phi(\xi), \qquad \xi = (x-x_0)/\left(A_0^{\frac{n}{4-2n}} t^{\frac{1}{4-2n}}\right),$$

being a similarity reduction of (4.16) in which $\Phi(\xi; n, \beta)$ is required to satisfy the matching conditions

$$\text{as } \xi \to -\infty, \quad \Phi \sim (-\xi)^\beta - \beta(\beta-1)(\beta-2)(2n+\beta-3)(-\xi)^{2n+\beta-4},$$
$$\text{at } \xi = 0^-, \quad \Phi = (-\xi)^{2n}\frac{\mathrm{d}^3\Phi}{\mathrm{d}\xi^3} = 0,$$

from which it follows that

$$\Phi \sim \kappa(\beta, n)(-\xi) \quad \text{as } \xi \to 0^- \tag{4.17}$$

for some constant κ (which could in principle take either sign, reliable intuition about the signs of such quantities being hard to come by in high-order diffusion problems). In fact, for $\beta = 1 + 2(2-n)N$ for integer N (such that $\beta > 2$), $\Phi(\xi)$ takes the form

$$\Phi = (-\xi)\sum_{m=0}^{N} a_m (-\xi)^{2(2-n)m}$$

with $a_N = 1$ and where a_0 alternates in sign with increasing N. More significantly, for $\beta = 2(1 + (2-n)N)$, we have

$$\Phi = (-\xi)^2 \sum_{m=0}^{N} a_m (-\xi)^{2(2-n)m},$$

so that

$$\kappa(2(1+(2-n)N), n) = 0 \tag{4.18}$$

gives explicitly the values of β,

$$\beta_N = 2(1+(2-n)N), \quad N = 1, 2, 3, \ldots, \tag{4.19}$$

at which κ changes sign. Thus κ changes sign infinitely often as $\beta \to \infty$.

In view of (4.13) and (4.17) there is a further, narrower, inner region with

$$x = x_0 + t^{\frac{\beta-1}{4-2n}}\zeta, \quad h \sim t^{\frac{\beta-1}{2-n}}\Psi(\zeta), \tag{4.20}$$

the dominant balance as $t \to 0^+$ being given by

$$\frac{\mathrm{d}}{\mathrm{d}\zeta}\left(\Psi^n \frac{\mathrm{d}^3\Psi}{\mathrm{d}\zeta^3}\right) = 0,$$

implying

$$\frac{\mathrm{d}^3\Psi}{\mathrm{d}\zeta^3} = 0. \tag{4.21}$$

For $C_0\kappa > 0$ we thus have instantaneous advance of the interface (with velocity zero at $t = 0^+$) with

$$\Psi = A_0(\zeta_0 - \zeta)^2, \quad \zeta_0 = A_0^{\frac{n(\beta+1)-4}{4-2n}} C_0\kappa, \quad s \sim x_0 + \zeta_0 t^{\frac{\beta-1}{4-2n}}, \tag{4.22}$$

in order to match with (4.17). A yet narrower inner region, with

$$\zeta = \zeta_0 + O(t^{(\beta-2)/(2n-3)}),$$

is then present near the interface, with scalings

$$x = s(t) + t^{\frac{\beta-5+2n}{2(2-n)(2n-3)}} z, \quad h \sim t^{\frac{\beta-5+2n}{(2-n)(2n-3)}}\phi(z), \tag{4.23}$$

whereby, matching with (4.22),

$$\frac{\beta-1}{4-2n}\zeta_0 = \phi^{n-1}\frac{\mathrm{d}^3\phi}{\mathrm{d}z^3}, \tag{4.24a}$$

$$\text{as } z \to -\infty, \quad \phi \sim A_0(-z)^2, \tag{4.24b}$$

$$\text{at } z = 0^-, \quad \phi = \frac{\mathrm{d}\phi}{\mathrm{d}z} = 0. \tag{4.24c}$$

This completes the description of the case $C_0\kappa > 0$.

The problem (4.24a–c) has no solution for $\zeta_0 < 0$, corresponding to the fact that interfaces cannot recede when $n \geq 3/2$; a different scenario is therefore needed when $C_0\kappa < 0$ in which $\zeta = \zeta_0$ in (4.22) no longer coincides with the interface; in other words, a quantity $\sigma(t)$, with

$$\sigma \sim x_0 + \zeta_0 t^{\frac{\beta-1}{4-2n}} \quad \text{as } t \to 0^+, \tag{4.25}$$

replaces $s(t)$ in (4.23) (with $s(t) = x_0$ now holding for $t \leq t_w$), and (4.24a–c) becomes

$$\frac{\beta-1}{4-2n}\zeta_0(\phi - \phi_\infty) = \phi^n\frac{\mathrm{d}^3\phi}{\mathrm{d}z^3}, \tag{4.26a}$$

$$\text{as } z \to -\infty, \quad \phi \sim A_0(-z)^2, \tag{4.26b}$$

$$\text{as } z \to \infty, \quad \phi \sim \phi_\infty, \tag{4.26c}$$

a boundary condition count indicating that, since $\zeta_0 < 0$, (4.26a–c) suffices to determine $\phi(z)$, up to translates in z, and ϕ_∞. The scaling properties of (4.26a–c) imply

that ϕ and ϕ_∞ are proportional to $(\zeta_0^2/A_0^3)^{1/(2n-3)}$, with z scaling as $(|\zeta_0|/A_0^n)^{1/(2n-3)}$. In $\sigma < x < x_0$, whereby

$$x_0 - x = O\left(t^{\frac{\beta-1}{4-2n}}\right), \quad h = O\left(t^{\frac{\beta-5+2n}{(2-n)(2n-3)}}\right),$$

we have to leading order that $\partial h/\partial t = 0$ with, in view of (4.25)–(4.26a–c), the matching condition

$$h \sim ((x_0 - x)/|\zeta_0|)^{\widehat{\alpha}(\beta)} \quad \text{on } t = \sigma^{-1}(x),$$

where

$$\widehat{\alpha}(\beta) := \frac{2(\beta - 5 + 2n)}{(\beta - 1)(2n - 3)},$$

implying that

$$h \sim ((x_0 - x)/|\zeta_0|)^{\widehat{\alpha}(\beta)} \quad \text{for } \sigma < x < x_0. \tag{4.27}$$

The exponent $\widehat{\alpha}(\beta)$ in (4.27) is monotonic increasing in β (given that $\beta > 2$) and satisfies

$$\widehat{\alpha}(2) = 2, \quad \widehat{\alpha}\left(1 + \frac{2n}{3}\right) = \frac{4}{n}, \quad \widehat{\alpha}(\infty) = \frac{2}{2n-3}.$$

It follows for $\beta > 1 + 2n/3$ that $\widehat{\alpha}$ lies in the regime of section 3.1, so that (4.27) describes the behavior near the interface up to the waiting time; for $2 < \beta < 1 + 2n/3$, however, $\widehat{\alpha}$ lies in the regime of section 4.4.1, so that (4.27) in turn breaks down sufficiently close to the interface and (4.14) is attained locally via a small-time similarity solution of the form (4.1)–(4.3a,b), with α replaced by $\widehat{\alpha}$. Such behavior represents a novel waiting-time phenomenon for degenerate parabolic equations (there being no corresponding scenario for the porous-medium equation), but there are similarities with, for example, Hele–Shaw flows with suction, whereby the free surface profile can instantly change to a new configuration (cf. (4.27)), which then persists (see [29]).

Analysis of cases with $\kappa = 0$ requires specification of an additional term in the local (1.5), and remarkably fine structure arises in consequence. Thus (cf. (4.18)) for

$$h_0(x) \sim A_0(x_0 - x)^2 + C_0(x_0 - x)^{2(1+(2-n)N)} + D_0(x_0 - x)^\gamma, \tag{4.28}$$

where $\gamma > 2(1 + (2 - n)N)$, we expect for each N a sequence of critical values of γ which represent further refined dividing lines between solutions that expand at once and those that wait; for those borderline values of γ, a further term in the expansion of (4.28) must be incorporated and so on. The first set of these dividing lines can be identified concisely via the one-degree-of-freedom (i.e., overspecified) family of local solutions (obtained by constructing an algebraic expansion for h about the leading-order term in (4.29))

$$h \sim a(t)(x_0 - x)^2 - \frac{\dot{a}(t)}{12(2-n)(5-2n)(3-n)a^n(t)}(x_0 - x)^{6-2n} + O((x_0 - x)^{10-4n}), \tag{4.29}$$

corresponding to (4.28) with $N = 1$ and

$$a(0) = A_0, \qquad \dot{a}(0) = -12(2-n)(5-2n)(3-n)A_0^n C_0,$$

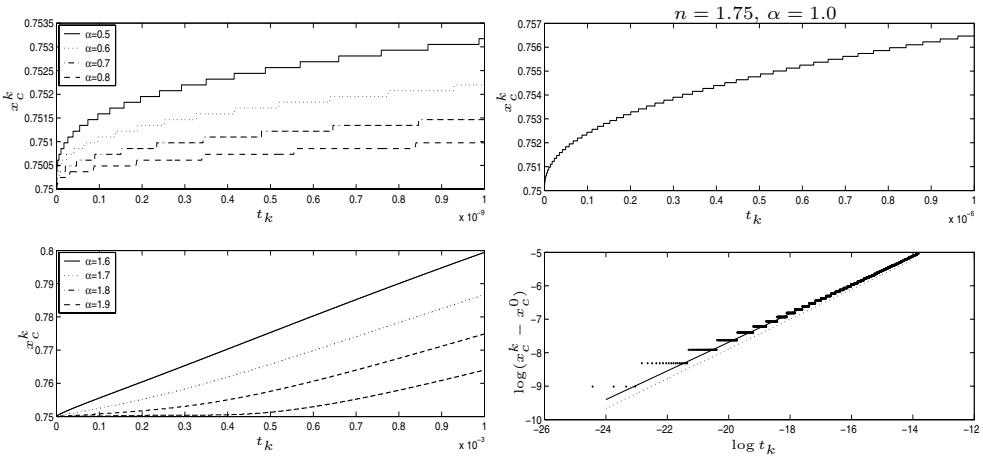


FIG. 4.5. *Numerical results for* $n = 1.75$, $\alpha < 2$. *On the left we show the numerical free boundary advancing for* $\alpha = 0.5, 0.6, 0.7, 0.8$, $T = 10^{-9}$ *(upper left plot) and for* $\alpha = 1.6, 1.7, 1.8, 1.9$, $T = 10^{-3}$ *(lower left plot); on the right we present results for* $\alpha = 1.0$, $T = 10^{-6}$, *with the numerical free boundary plotted against time in the upper right plot, and with* $\log t_k$ *plotted against* $\log(x_c^k - x_c^0)$ *as a discrete set of points in the lower right plot, with the solid line following from a least squares fitting and the straight dotted line from asymptotic theory.*

and identifying the first critical value of γ for $N = 1$ to be $7 - 4n$; higher values of N correspond to $\dot{a}(0) = 0$ in this local expansion. Because it is overspecified, the local expansion of (4.29) pertains only when the local form of the initial data is consistent with the powers of $x_0 - x$ therein and, as already implied, it represents a borderline between solutions of the form (1.2) and (4.14).

4.4.3. $\alpha < 2$. Here the interface advances immediately, with $f(\eta)$ having local behavior (4.10) and with unbounded initial velocity for $\alpha < 3/n$, finite for $\alpha = 3/n$, and tending to zero for $3/n < \alpha < 2$. The last of these ranges disappears as n drops below $3/2$, providing one indication of the need to address this regime separately.

To test this we ran numerical experiments for $n = 1.75$ (giving $3/n = 1.7143$, $4/n = 2.2857$), with $\alpha \in [0.5, 1.9]$. Our results again support the conjecture. In each case x_c^k advances, with the speed of the advance decreasing as α increases. This is shown in Figure 4.5, in which x_c^k is plotted against t_k for $n = 1.75$ with $\alpha = 0.5$, 0.6, 0.7, and 0.8 (upper left plot) and $\alpha = 1.6$, 1.7, 1.8, and 1.9 (lower left plot). Note the different time scales on the two axes.

In the upper right plot of Figure 4.5 we show the numerical free boundary advancing for $n = 1.75$ and $\alpha = 1.0 < 3/n$. As before, we test the hypothesis (4.11) by plotting $\log t_k$ against $\log(x_c^k - x_c^0)$ (in the lower right plot), and we estimate $\gamma = 0.43$. For comparative purposes we plot a dotted line with slope $(4 - n\alpha)^{-1} = 0.44$ on the same graph. The expected and estimated values of γ are shown in Table 4.2. For $\alpha = 1.9$ we exclude the first step in the advance of x_c^k, as discussed in section 4.2. The estimates for γ are very close to the expected values, with this being especially true for values of α away from the edges of the parameter regime.

4.5. Small-time behavior for $n = 3/2$. For $\alpha < 2$, the support of h expands immediately with unbounded velocity; in view of (1.3), the local behavior of $f(\eta)$ then

TABLE 4.2
Estimated and expected values of γ for various α, $n = 1.75$.

α	0.5	0.7	0.9	1.1	1.3	1.5	1.7	1.9
$(4-n\alpha)^{-1}$	0.32	0.36	0.41	0.48	0.58	0.73	0.98	1.48
γ	0.30	0.34	0.40	0.47	0.58	0.73	0.98	1.53

takes the form

$$f(\eta) \sim \left(\frac{3\eta_0}{2(8-3\alpha)}(\eta_0-\eta)^3 \ln\left(\frac{1}{(\eta_0-\eta)}\right)\right)^{\frac{2}{3}} \quad \text{as } \eta \to \eta_0^-.$$

For $2 < \alpha < 8/3$ the behavior is again as described in section 4.4.1. The case $\alpha = 2$ is particularly delicate, with the initial exponents $\alpha = 2$ and $\alpha = 3/n$ coinciding for $n = 3/2$. Much of the analysis in section 4.4.2 nevertheless still pertains—in particular (4.17)–(4.19) hold—so that $\beta_N = 2 + N$, as does (4.20)–(4.22). However, the scalings (4.23) are evidently inapplicable for $n = 3/2$, and the appropriate scalings are instead (for $C_0\kappa > 0$)

$$h = (s-x)^2\Phi(\xi,t), \quad \xi = t^{\beta-2}\ln(1/(s-x)), \quad s \sim x_0 + \zeta_0 t^{\beta-1}, \tag{4.30}$$

so that the spatial scaling is exponentially small in t, which yield as the dominant balance

$$(\beta-1)\zeta_0 = 2\Phi_0^{1/2}\frac{\mathrm{d}\Phi_0}{\mathrm{d}\xi},$$

and hence

$$\Phi_0 = \left(A_0^{3/2} + \frac{3}{4}(\beta-1)\zeta_0\xi\right)^{2/3}, \tag{4.31}$$

where we have matched with (4.22). For $C_0\kappa > 0$, this has the required local behavior (1.3) and completes the small-time analysis. For $C_0\kappa < 0$, we again introduce

$$\sigma \sim x_0 + \zeta_0 t^{\beta-1},$$

which specifies the interior layer location, and replace the scalings in (4.30) by

$$h = (\sigma - x)^2\Phi(\xi,t), \quad \xi = t^{\beta-2}\ln(1/(\sigma-x)),$$

to recover (4.31) with ζ_0 (given by (4.22)) negative. Hence ϕ_0 becomes zero at $\xi = \xi_c$, where

$$\xi_c = \frac{4A_0^{3/2}}{3(\beta-1)|\zeta_0|}.$$

There is now a further asymptotic region in which

$$x = \sigma(t) + \rho(t)\mathrm{e}^{-\xi_c/t^{\beta-2}}z, \quad h \sim |\dot\sigma|^{2/3}\rho^2\mathrm{e}^{2\xi_c/t^{\beta-2}}\phi, \tag{4.32}$$

where the scaling on h is chosen to obtain the appropriate leading order balance, namely (cf. (4.26a–c))

$$-(\phi-\phi_\infty) = \phi^{3/2}\frac{\mathrm{d}^3\phi}{\mathrm{d}z^3},$$

$$\text{as } z \to -\infty, \quad \phi \sim \left(\frac{3}{4}(-z)^3\ln(-z)\right)^{2/3},$$

$$\text{as } z \to \infty, \quad \phi \sim \phi_\infty,$$

where ϕ_∞ is again to be determined as part of the solution and we have matched with (4.31). The preexponential factor $\rho(t)$ is expected to be algebraic in t; its calculation would require correction terms in the various expansions to be evaluated and we shall not pursue such matters further. Applying arguments similar to those in section 4.4.2, we obtain from (4.32) that

$$\ln h \sim -2\xi_c|\zeta_0|^{\frac{\beta-2}{\beta-1}}/(x_0-x)^{\frac{\beta-2}{\beta-1}} \quad \text{for } \sigma < x < x_0, \tag{4.33}$$

so the height of the film left behind by the retreating interior layer at $x = \sigma$ is exponentially small (and thus in particular implies waiting-time behavior at $x = x_0$). This reflects the status of $n = 3/2$ as a critical case; as we shall shortly see, for $n < 3/2$ the interface $x = s$ itself retreats in the corresponding regime (in other words, the film thickness left behind $x = \sigma$ drops from being algebraically small for $3/2 < n < 2$, as in (4.27), through exponentially small for $n = 3/2$ (see (4.33)) to zero for $n < 3/2$).

4.6. Small-time behavior for $n < 3/2$. We have already alluded to the qualitatively new feature, implicit in the local behavior (1.4), which can occur in this regime, namely that the interface can retreat. Such behavior is most simply demonstrated by the case $\alpha = 3/n$ in which the small-time similarity solution is given by (4.6), with the interface retreating at a finite rate; in this regime (4.6) is nongeneric, being smoother than the expected local behavior (1.4). For $\alpha > 3/n$, we anticipate waiting-time behavior, as in section 4.4.1; see also Appendix A. (In addition, an analysis similar to that of Appendix B can be performed in the limit $\alpha \to (3/n)^+$.) The result for $\alpha = 3/n$ and $\alpha = 2$ suggests that for $\alpha < 2$ the interface expands at an unbounded rate, with

$$f(\eta) \sim \beta(\eta_0 - \eta)^2 \quad \text{as } \eta \to \eta_0^- \tag{4.34}$$

for some constant η_0, while for $2 < \alpha < 3/n$ contraction occurs at an unbounded rate, with $\eta_0 < 0$ in (4.34). The critical case $\alpha = 2$ is again of particular interest, in particular since waiting-time behavior can in principle occur in this case also (but only for extremely special initial data; cf. (4.29)). The analysis for $\alpha = 2$ is more straightforward than that above, since (4.22)–(4.24a–c) now apply right up to the interface for $C_0\kappa < 0$ (retreat) as well as for $C_0\kappa > 0$ (advance). Such behavior can be illustrated by the explicit solution (4.7)–(4.9) for $n = 1$, wherein $\beta = 3$ and $\kappa = -15$.

For $\alpha = 2$ we have that

$$s(t) \sim x_0 + \zeta_0 t^{\frac{\beta-1}{4-2n}} \quad \text{as } t \to 0^+,$$

which exhibits the same interface time-dependence as (4.1) with $\alpha = (4(\beta-2)+2n)/((\beta-1)n)$; since $\beta > 2$, it follows that this α lies in the range $2 < \alpha < 4/n$, and it implies that the same $s(t)$ can result from quite different initial data (in this case for $2 < \alpha < 3/n$, for which the interface retreats). This reflects the high order of (1.1a), whereby the local behavior at the interface contains a further degree of freedom in addition to $s(t)$ and contrasts with the situation for the second-order case (1.6) (cf. [1]).

4.6.1. $3/n < \alpha < 4/n$. In this regime we anticipate waiting-time behavior, as described in section 4.4.1. For $\alpha_5 < \alpha < 4/n$ we expect monotonic decay onto the $4/n$ solution, for $\alpha_2 < \alpha < \alpha_5$ we expect oscillatory decay, and for $3/n < \alpha < \alpha_2$ we expect a limit cycle to arise in the local description for $0 < t < t_w$ (see (A.5)).

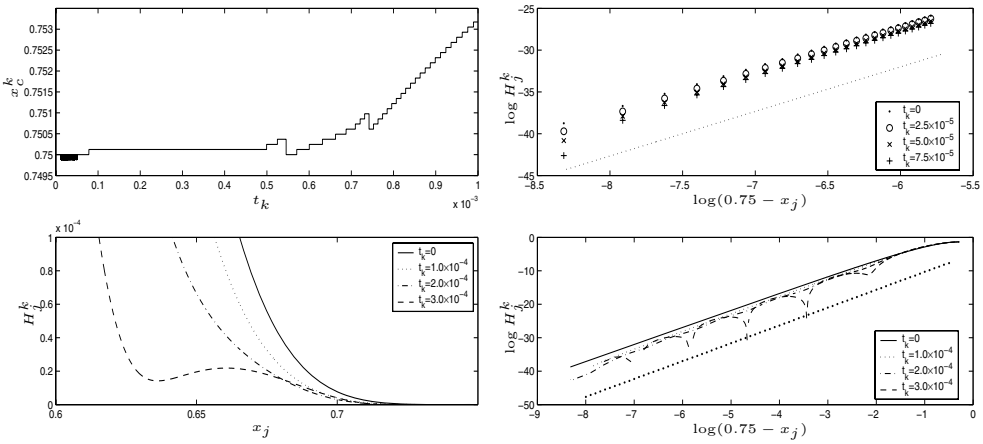


FIG. 4.6. *Numerical results for* $n = 0.75$, $\alpha = 5.1$: *waiting-time behavior (upper left plot)*; *profile of* H_j^k *near the interface at various times (lower left plot)*; $\log H_j^k$ *against* $\log(x_c^k - x_j)$ *in the vicinity of the free boundary (upper right plot), and over the whole range* $x_j \in [0, x_c^k)$ *(lower right plot), with a dotted line of gradient* $4/n$ *(from asymptotic theory) in each case.*

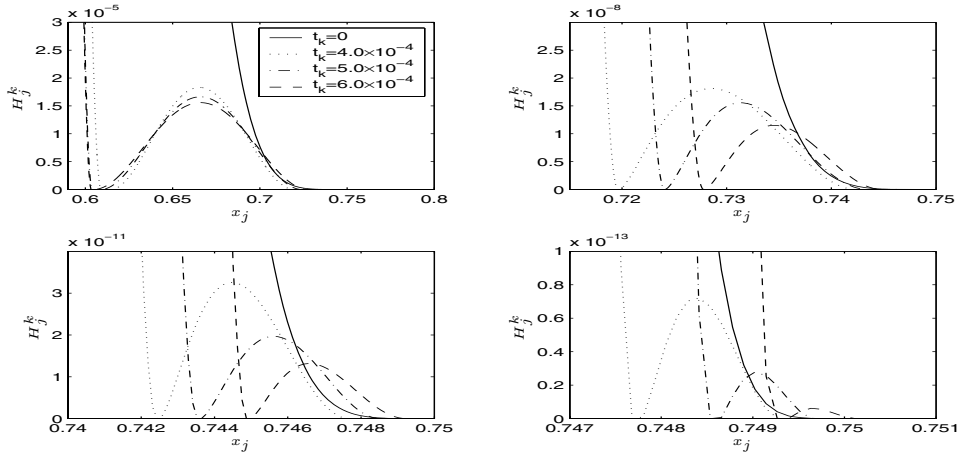


FIG. 4.7. *Numerical results for* $n = 0.75$, $\alpha = 5.1$; *profile of* H_j^k *near the interface at various times (same legend for each plot).*

By monotonic decay we mean decay like $\eta^{-\gamma}$, where γ is real, and by oscillatory decay we mean that the solution decays like $\eta^{-\gamma - \mathrm{i}\mu}$, where γ, μ are real. We present numerical results below for $n = 0.75$ (giving $3/n = 4$, $\alpha_2(n) = 4.2061$, $\alpha_5(n) = 4.9$, $4/n = 5.3333$), with $\alpha = 5.1$ (Figures 4.6 and 4.7), $\alpha = 4.5$ (Figures 4.8 and 4.9) and $\alpha = 4.1$ (Figures 4.10 and 4.11), thus covering each of the three parameter regimes described above (see also Appendix A). We also present results for $n = 1.0$ (for which $\alpha_2(n) = 3.2195$) with $\alpha = 3.1$ (Figures 4.12 and 4.13), with this second example in the range $3/n < \alpha < \alpha_2$ reflecting the extremely delicate nature of the results in this regime.

In the upper left plots of Figures 4.6, 4.8, 4.10, and 4.12 we plot x_c^k against t_k for each example. In Figures 4.6 and 4.8, x_c^k remains stationary for a period before

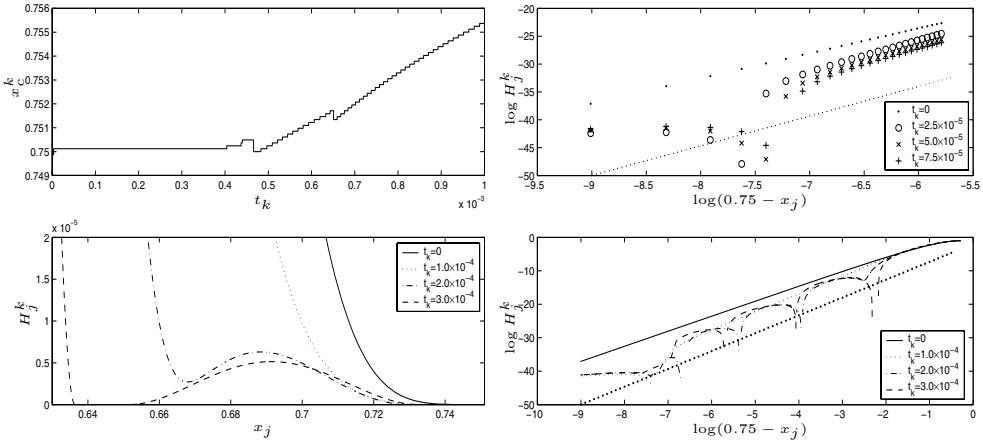


FIG. 4.8. *Numerical results for* $n = 0.75$, $\alpha = 4.5$: *waiting-time behavior (upper left plot)*; *profile of* H_j^k *near the interface at various times (lower left plot)*; $\log H_j^k$ *against* $\log(x_c^k - x_j)$ *in the vicinity of the free boundary (upper right plot), and over the whole range* $x_j \in [0, x_c^k)$ *(lower right plot), with a dotted line of gradient* $4/n$ *(from asymptotic theory) in each case.*

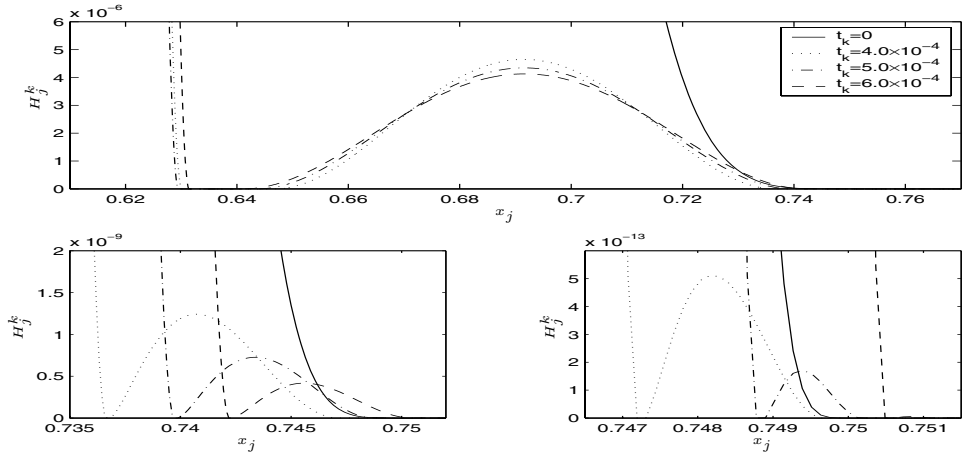


FIG. 4.9. *Numerical results for* $n = 0.75$, $\alpha = 4.5$; *profile of* H_j^k *near the interface at various times (same legend for each plot).*

advancing, with a shorter waiting period in Figure 4.8. In Figures 4.10 and 4.12, x_c^k appears to immediately retreat, wait, and then advance. However, this retreat is over a very short distance, and over a longer time scale the boundary appears to wait; note the different scales on the two plots in the upper left corner of Figure 4.12.

In the lower left corner of Figures 4.6, 4.8, 4.10, and 4.12 we show profiles of H_j^k in the vicinity of the free boundary at various times before the free boundary has begun to advance. In each case the value of H_j^k drops faster further behind the free boundary, leading to the formation of humps near the boundary. In order to demonstrate the existence of more than one such hump, we show profiles of H_j^k on smaller and smaller scales nearer and nearer to the free boundary in Figures 4.7, 4.9, 4.11, and 4.13. Note the different scales on the horizontal and vertical axes of each

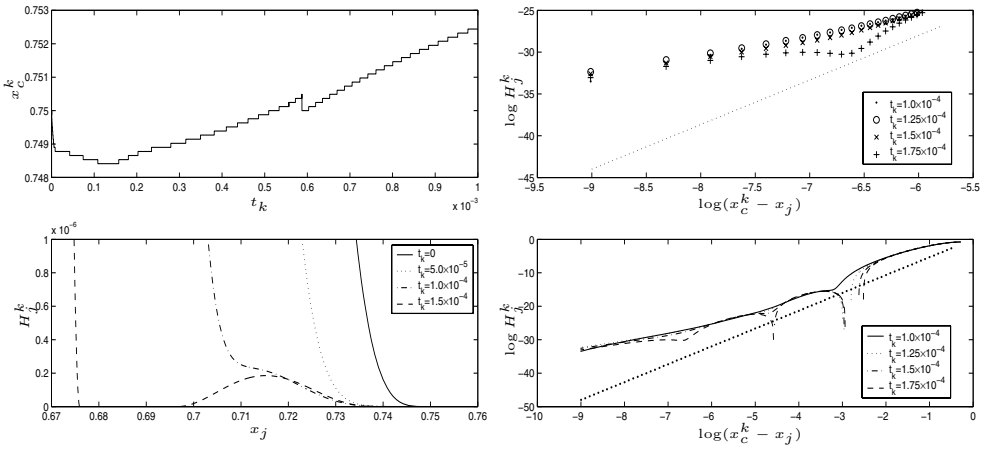


FIG. 4.10. *Numerical results for* $n = 0.75$, $\alpha = 4.1$: *motion of the numerical free boundary (upper left plot); profile of* H_j^k *near the interface at various times (lower left plot);* $\log H_j^k$ *against* $\log(x_c^k - x_j)$ *in the vicinity of the free boundary (upper right plot), and over the whole range* $x_j \in [0, x_c^k)$ *(lower right plot), with a dotted line of gradient* $4/n$ *(from asymptotic theory) in each case.*

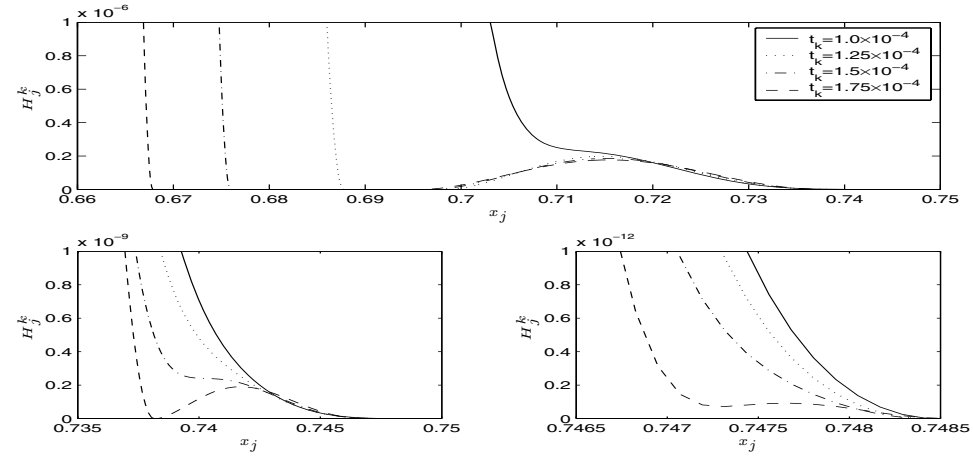


FIG. 4.11. *Numerical results for* $n = 0.75$, $\alpha = 4.1$; *profile of* H_j^k *near the interface at various times (same legend for each plot).*

plot in Figures 4.7, 4.9, 4.11, and 4.13. Due to the limitations of the numerical method and the scale of the plots it is possible that some of these results may be spurious, but the repetition of the evidence found on the larger scales does provide a degree of support for the conjectures.

The issue of whether these types of profiles lead to film break up (i.e., satellite droplets separated by dead cores in which h is identically zero) is an interesting one warranting further study. In the current context we note first that there remain open questions regarding the range of n for which such rupture can occur, which it would not be appropriate to explore here; second that the small-time similarity solutions cannot exhibit such break up (each satellite drop must conserve mass, which is inconsistent with their self-similar form); and finally, for $n < 1/2$ they could contain touch down

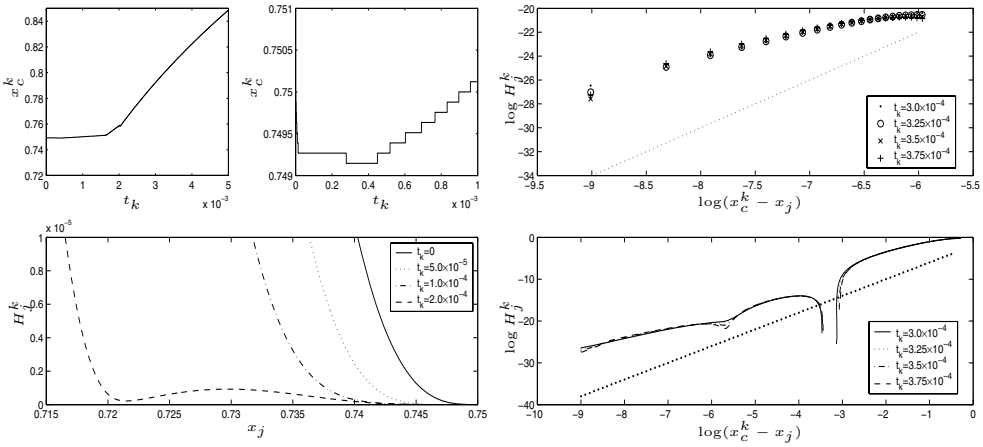


FIG. 4.12. *Numerical results for $n = 1.0$, $\alpha = 3.1$: motion of the numerical free boundary (upper left plots: note the different scales on each figure); profile of H_j^k near the interface at various times (lower left plot); $\log H_j^k$ against $\log(x_c^k - x_j)$ in the vicinity of the free boundary (upper right plot), and over the whole range $x_j \in [0, x_c^k)$ (lower right plot), with a dotted line of gradient $4/n$ (from asymptotic theory) in each case.*

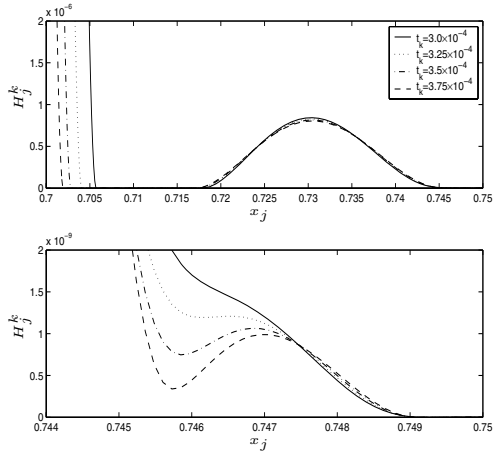


FIG. 4.13. *Numerical results for $n = 1.0$, $\alpha = 3.1$; profile of H_j^k near the interface at various times (same legend for each plot).*

points (at which $h = 0$), as analyzed in [31].

In the upper right corner of Figures 4.6, 4.8, 4.10 and 4.12 we plot $\log H_j^k$ against $\log(x_c^k - x_j)$ in the vicinity of the free boundary at various times before the free boundary has begun to advance. For comparison we also plot a dotted line with slope $4/n$ in each figure. In Figure 4.6, the best fitting least squares line for $t_k = 0$ has a slope of 5.10, rising to 5.26 for $t_k = 2.5 \times 10^{-5}$ and 5.45 for $t_k = 5.0 \times 10^{-5}$. For $t_k = 7.5 \times 10^{-5}$ the log-log plot is no longer very straight. We remark that in this case, with $n = 0.75$ and $\alpha = 5.1$, H_j^k is very close to zero quite far behind the free boundary, hence the numerical results are extremely delicate. In Figure 4.8, the log-log plot is not straight immediately in the vicinity of the free boundary for any $t_k > 0$, although

it is fairly straight further away from the free boundary. In Figures 4.10 and 4.12, the log-log plots are not very straight, and the best fitting least squares lines have slopes significantly lower than $4/n$.

In the lower right corners of Figures 4.6, 4.8, 4.10 and 4.12 we plot $\log H_j^k$ against $\log(x_c^k - x_j)$ over the whole range $x_j \in [0, x_c^k)$ at various times before the free boundary has begun to advance, plotting again a dotted line with slope $4/n$ for comparison. In Figure 4.6, as t_k increases, the log-log plot becomes less and less straight, and for $t_k = 3.0 \times 10^{-3}$ the periodic behavior of the solution near the interface can clearly be seen. The log-log plots for $t_k = 4.0 \times 10^{-3}$ and $t_k = 5.0 \times 10^{-3}$ are very similar to that for $t_k = 3.0 \times 10^{-3}$ but are not shown here. In Figure 4.8, the formation of humps further and further from the free boundary becomes apparent. For $t_k = 1.0 \times 10^{-4}$, the slope of the log-log plot away from the free boundary is close to $4/n$. For each of Figures 4.10 and 4.12, as t_k increases, the formation of extra humps in the vicinity of the free boundary becomes apparent.

4.6.2. $2 < \alpha < 3/n$. To test the conjecture that the free boundary retreats instantaneously with unbounded velocity, we ran experiments with $n = 1.0$ and $\alpha \in [2.1, 2.9]$. We plot x_c^k against t_k in the left panel of Figure 4.14 for $\alpha = 2.2$, 2.4, 2.6, and 2.8. The results support the conjecture. In each case the free boundary retreats, waits, and then advances, although the subsequent advance can only be seen in the figure for $\alpha = 2.2$. The initial velocity of x_c^k appears to decrease as α increases, although as α increases, the length of the period for which the free boundary retreats also increases, so that the maximum distance retreated occurs for $\alpha = 2.9$.

As before we test the hypothesis $x_c^k = x_c^0 - At_k^\gamma$ for some constants $A > 0$ and γ by plotting $\log(x_c^0 - x_c^k)$ against $\log t_k$. Again, if the hypothesis is correct, we expect a straight line with slope γ, and to estimate γ we take a least squares fit over the range for which the log-log plot is approximately straight. In the right half of Figure 4.14 we plot x_c^k against t_k (upper section) and this log-log plot (lower section) for $n = 1.0$ and $\alpha = 2.5$. The log-log plot is approximately straight, and the best fitting least

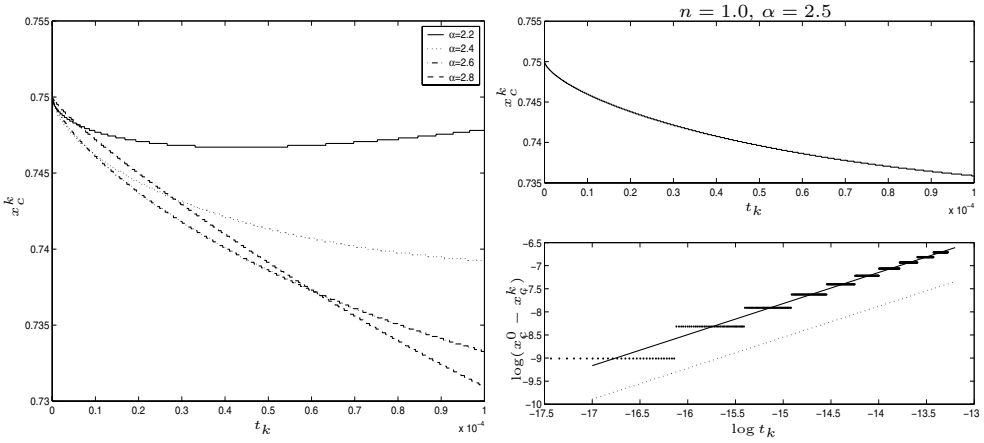


FIG. 4.14. *Numerical results for $n = 1.0$, $2 < \alpha < 3/n$. In the left panel we plot the retreating free boundary for various α. In the right panels we show results for $n = 1.0$, $\alpha = 0.5$: in the upper right section we show the retreating free boundary; in the lower right section we plot $\log t_k$ against $\log(x_c^0 - x_c^k)$ as a discrete set of points, with the solid line following from a least squares fitting and the straight dotted line from asymptotic theory.*

TABLE 4.3
Estimated and expected values of γ for $n = 1.0$, various α.

α	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9
$(4-n\alpha)^{-1}$	0.53	0.56	0.59	0.63	0.67	0.71	0.77	0.83	0.91
γ	0.44	0.53	0.58	0.63	0.67	0.72	0.77	0.81	0.85

squares line is plotted as a solid line on the same figure. For comparison we also plot a dotted line with slope $(4-n\alpha)^{-1} = 0.67$. The estimated value of $\gamma = 0.67$ matches this exactly to two decimal places. The expected and estimated values of γ for each value of α tested are shown in Table 4.3. The trend of γ increasing with α is clear, and away from the edges of the parameter regime the estimated value of γ is very close to the expected value.

4.6.3. $\alpha < 2$. To test the conjecture that the free boundary advances instantaneously, with an unbounded velocity, we ran experiments with $n = 1.0$ and $\alpha \in [0.5, 1.9]$. We plot x_c^k against t_k for $\alpha = 0.6$, 0.7, 0.8, and 0.9 (upper left plot), and for $\alpha = 1.5$, 1.6, 1.7, and 1.8 (lower left plot) in Figure 4.15. Note the different time scales on the two plots. The results support the conjecture, and the initial velocity of x_c^k decreases as α increases.

We again test the hypothesis (4.11), plotting x_c^k against t_k (upper right section of Figure 4.15) and $\log(x_c^k - x_c^0)$ against $\log t_k$ (lower right section of Figure 4.15) for $n = 1.0$ and $\alpha = 1.2$. The log-log plot is approximately straight, and the best fitting least squares line, plotted as a solid line on the same figure, has a slope of 0.34. For comparison we also plot a dotted line with slope $(4-n\alpha)^{-1} = 0.36$ on the same figure. The expected and estimated values of γ are shown in Table 4.4. The trend of γ increasing with α is clear, and for values of α away from the borderline value $\alpha = 2$ the estimated value of γ is very close to the expected value.

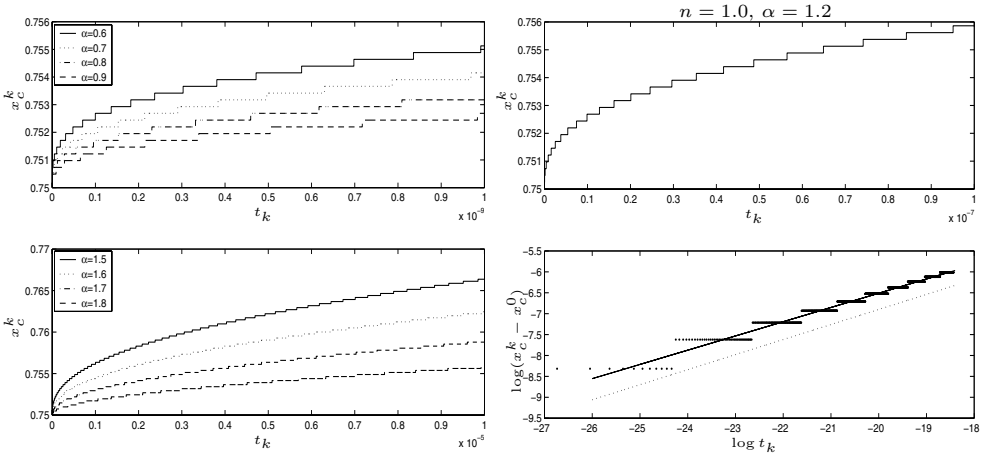


FIG. 4.15. $n = 1.0$, $\alpha < 2$. *In the left half of the figure we plot the advancing free boundary for various α (note the different time scales on the two plots). In the right half of the figure we show results for $n = 1.0$, $\alpha = 1.2$: in the upper right section we show the advancing free boundary; in the lower right section we plot $\log t_k$ against $\log(x_c^k - x_c^0)$ as a discrete set of points, with the solid line following from a least squares fitting and the straight dotted line from asymptotic theory.*

TABLE 4.4
Estimated and expected values of γ for $n = 1.0$, various $\alpha < 2$.

α	0.5	0.7	0.9	1.1	1.3	1.5	1.7	1.9
$(4 - n\alpha)^{-1}$	0.29	0.30	0.32	0.34	0.37	0.40	0.43	0.48
γ	0.28	0.29	0.31	0.33	0.36	0.40	0.43	0.53

5. Conclusions. As we have seen, the thin-film equation (1.1a) exhibits a much broader range of small-time phenomena than its second-order analogue, (1.6). Thus, while the behavior of the former with $2 < n < 3$ corresponds very closely to that of the latter with any $n > 0$, for $n < 2$ equation (1.1a) exhibits a range of α in which the interface waits but the local profile changes instantaneously from that of the initial data (this combination does not occur for (1.6)) and can exhibit monotonic or oscillatory decay to the local solution (4.14) or limit-cycle behavior of the form (A.5). Moreover, for $n < 3/2$ fronts can either advance or retreat, and our small-time classification gives rather precise criteria on the initial data in this regard. The very delicate interlacing of initial profiles leading to immediate expansion or to a finite waiting time, as outlined in section 4.4.2, for example, also deserves highlighting.

In Table 5.1 we demonstrate how these results apply to the cases $n = 1$ (which describes thin films in a Hele–Shaw cell [16] and the strong-slip limit of the Greenspan [22] slip regularization) and $n = 2$ (which corresponds to the strong-slip limit of the usual (Navier) slip-regularization; see [26], for example). See also [34] for further relevant background. It is noteworthy that the case $n = 2$ is a critical one in a number of respects (some of which are implicit in Figure 1.1).

The high-order problem (1.1) is a demanding one from the numerical point of view; this wealth of distinct behaviors occurring over short length and time scales necessitates particularly refined, careful, and detailed computational studies if the relevant phenomena are to be captured adequately, and we have sought to implement the required program of extensive numerical investigations. Taking into account the delicacy of some of the asymptotic results and the limitations of the numerical scheme, numerical results are shown only for those parameter regimes wide enough that suitable "intermediate" values of n and α can be used.

A number of generalizations immediately suggest themselves. In higher dimensions, the small-time behavior of an initially smooth interface will be locally one-dimensional, so most of the conclusions carry over. For $n \geq 3$, the smoothest solutions have fixed interfaces, and here waiting-time phenomena relate (for $3 \leq n < 4$) to a delay in the contact angle becoming finite; we shall not elaborate on such matters

TABLE 5.1
Small- and waiting-time behavior for $n = 1$ and $n = 2$.

n	Range of α	Behavior
1	$4/n = 4 < \alpha$	Global waiting-time, ended by shock.
	$\alpha_5 \approx 3.7 < \alpha < 4$	Interface waits but local profile changes instantaneously from that of initial data and can exhibit monotonic decay to local solution.
	$\alpha_2 \approx 3.2 < \alpha < 3.7$	As above, but with oscillatory decay.
	$3/n = 3 < \alpha < 3.2$	As above, but with limit cycle behavior.
	$2 < \alpha < 3$	Interface retreats instantaneously.
	$\alpha > 2$	Interface advances instantaneously.
2	$4/n = 2 \leq \alpha$	Global waiting time, ended by shock.
	$\alpha < 2$	Free boundary advances instantaneously.

here, noting only that the approaches we have described above apply equally well in such contexts. As a final instance, we note that for $n < 3$ a finite contact angle condition can be imposed in place of the second of (1.1c) and a similar investigation performed; again, we shall not report the results of such a study here.

Appendix A. Applicability of the local solution (4.4). In this appendix we use boundary condition counting arguments to assess the applicability of (4.4) as a local solution to (4.2) for $0 < n < 2$. Writing

$$f \sim \left(\frac{n^3}{8(4-n)(2-n)(n+4)} (-\eta)^4 \right)^{\frac{1}{n}} + F \tag{A.1}$$

and linearizing yields

$$\frac{1}{4-n\alpha}(\alpha F - \eta F_\eta) = -\frac{n^3}{8(4-n)(2-n)(n+4)} \frac{\mathrm{d}}{\mathrm{d}\eta} \left(\eta^4 \frac{\mathrm{d}^3 F}{\mathrm{d}\eta^3} \right) - \frac{n}{n+4} \frac{\mathrm{d}}{\mathrm{d}\eta} (\eta F),$$

with solutions

$$F = K(-\eta)^p, \tag{A.2}$$

where the possible p are the roots of the quartic

$$\frac{n^3 p(p-1)(p-2)(p+1)}{8(4-n)(2-n)(n+4)} + \frac{n(p+1)}{n+4} + \frac{\alpha - p}{4-n\alpha} = 0. \tag{A.3}$$

The expansion of (A.1) with F given by (A.2) is self-consistent if $\mathrm{Re}(p) > 4/n$, so the relations between α and n such that two roots of (A.3) have $\mathrm{Re}(p) = 4/n$ are crucial; these relations can be shown to be

$$\alpha_1(n) = \frac{\alpha_{-b} + \alpha_\Delta}{\alpha_d}, \qquad \alpha_2(n) = \frac{\alpha_{-b} - \alpha_\Delta}{\alpha_d},$$

where

$$\begin{aligned} \alpha_d &= 2\left(n^2 - n - 8\right)\left(7\,n^3 - 84\,n^2 + 400\,n - 640\right) n, \\ \alpha_{-b} &= 47n^5 - 674n^4 + 3384n^3 - 3520n^2 - 17408n + 36864, \\ \alpha_\Delta &= (n+4)(2-n)(8-n)\sqrt{9216 - 5632n + 896n^2 + 112n^3 - 31n^4}, \end{aligned}$$

so that

$$\alpha_1 \sim 2 + \frac{11}{10}(2-n) + O((2-n)^2) \text{ as } n \to 2^-, \quad \alpha_1 \sim \frac{3}{n} + \frac{5}{24} + O(n) \text{ as } n \to 0^+,$$
$$\alpha_2 \sim 2 + \frac{31}{22}(2-n)^2 + O((2-n)^3) \text{ as } n \to 2^-, \quad \alpha_2 \sim \frac{21}{5n} - \frac{1}{120} + O(n) \text{ as } n \to 0^+.$$

It is also instructive to note the curves in (α, n) space on which roots of (A.3) become complex, namely the repeated roots case in which (A.3) and

$$\frac{\mathrm{d}}{\mathrm{d}p}\left[\frac{n^3 p(p-1)(p-2)(p+1)}{8(4-n)(2-n)(n+4)} + \frac{n(p+1)}{n+4} + \frac{\alpha - p}{4-n\alpha} \right] = 0 \tag{A.4}$$

are satisfied simultaneously. These curves are shown in Figure A.1.

We define $\alpha_5 = \alpha_5(n)$ to be the repeated root case ($\mathrm{v_b}$) shown in Figure A.1. The various curves in (α, n) space relevant to our discussion are all shown in Figure A.2.

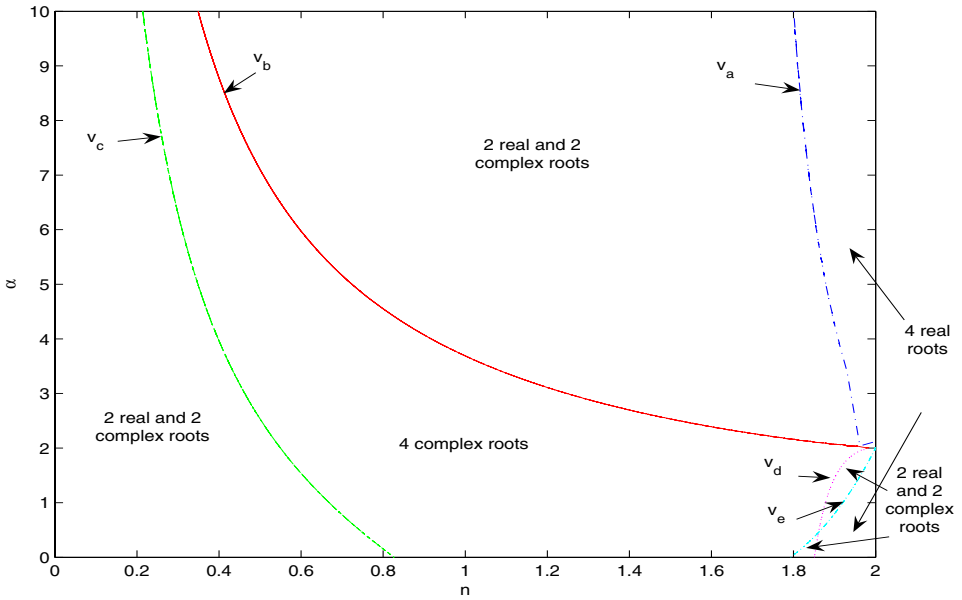


FIG. A.1. *The solutions to* (A.3) *and* (A.4). *To the right of* (v_a) *and of the rightmost of* (v_d) *and* (v_e) *there are four real roots: between* (v_a) *and* (v_b), *between* (v_d) *and* (v_e), *and to the left of* (v_c) *there are two real and two complex roots; between* (v_c), (v_b), *and the leftmost of* (v_d) *and* (v_e) *there are four complex roots.*

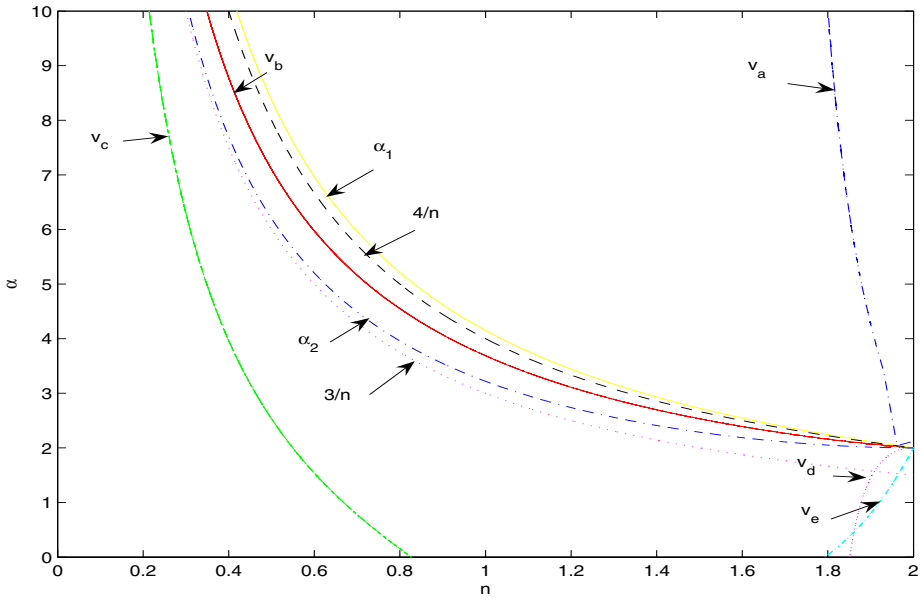


FIG. A.2. (α, n) *space, showing* $\alpha = \alpha_1$, $\alpha = 4/n$, $\alpha = \alpha_2$, $\alpha = 3/n$, *and the solutions to* (A.3) *and* (A.4). *Three of the roots become unbounded as* $\alpha = 4/n$ *is approached, with the fourth having* $p \sim 4/n$. *Above* $\alpha = \alpha_1$ *and below* $\alpha = \alpha_2$, *none of the roots have* $\mathrm{Re}(p) > 4/n$; *between* $\alpha = \alpha_2$ *and* (v_b), *two of the four complex roots have* $\mathrm{Re}(p) > 4/n$ *and the other two* $\mathrm{Re}(p) < 4/n$; *between* (v_b) *and* $\alpha = 4/n$ *both real roots satisfy* $p > 4/n$ *and both complex ones* $\mathrm{Re}(p) < 4/n$; *between* $\alpha = 4/n$ *and* $\alpha = \alpha_1$ *the real roots have* $p < 4/n$ *and the complex ones* $\mathrm{Re}(p) > 4/n$.

In $\alpha_2 < \alpha < \frac{4}{n}$, two roots of (A.3) have $\mathrm{Re}(p) > \frac{4}{n}$, and the local expansion (A.1) is correctly specified (the two degrees of freedom being the K's in (A.2) corresponding to those two roots). As α drops below α_2 we anticipate that a Hopf bifurcation occurs in (4.2)–(4.3a,b) with the local behavior as $\eta \to 0^-$ taking for $\alpha > \max(2, 3/n)$ the limit-cycle form

$$f \sim (-\eta)^{\frac{4}{n}} \Omega(-\ln(-\eta)), \tag{A.5}$$

where $\Omega(\xi)$ is periodic of period P, say, in ξ. Since on $\alpha = \alpha_2$

$$\mathrm{Im}(p) = \pm \frac{1}{\sqrt{2}n} \sqrt{96 - 24n - n^2 - \sqrt{9216 - 5632n + 896n^2 + 112n^3 - 31n^4}},$$

we anticipate that

$$P \sim \frac{2\sqrt{2}\pi n}{\sqrt{96 - 24n - n^2 - \sqrt{9216 - 5632n + 896n^2 + 112n^3 - 31n^4}}} \quad \text{as } \alpha \to \alpha_2^-.$$

Note that P will depend on α and n but, in view of the scaling properties of (4.2)–(4.3a,b), not on A_0. For $\alpha_5 < \alpha < \frac{4}{n}$, the decay to (4.4) is nonoscillatory, while for $\alpha_2 < \alpha < \alpha_5$ damped oscillations occur.

Appendix B. $3/2 < n < 2$, $\alpha \to 2$. We are concerned here with the behavior of (4.2)–(4.3a,b) for α close to two. Writing $\alpha = 2 + \epsilon$, $0 < |\epsilon| \ll 1$, we have for $\eta = O(1)$ that

$$f \sim A_0(-\eta)^2 + \epsilon f_1(\eta) \tag{B.1}$$

with

$$\frac{1}{2(2-n)} \left(2f_1 - \eta \frac{\mathrm{d}f_1}{\mathrm{d}\eta} + A_0(-\eta)^2 \right) = -A_0^n \frac{\mathrm{d}}{\mathrm{d}\eta} \left((-\eta)^{2n} \frac{\mathrm{d}^3 f_1}{\mathrm{d}\eta^3} \right), \tag{B.2a}$$

$$\text{as } \eta \to -\infty, \quad f_1 \sim A_0(-\eta)^2 \ln(-\eta) - 2(2n-1)A_0^{n+1}(-\eta)^{2(n-1)}, \tag{B.2b}$$

$$\text{as } \eta \to 0^-, \quad f_1 = (-\eta)^{2n} \frac{\mathrm{d}^3 f_1}{\mathrm{d}\eta^3} = 0. \tag{B.2c}$$

It follows from (B.2a–c) that

$$f_1 \sim -\mu(n) A_0^{\frac{4-n}{2(2-n)}} (-\eta) \quad \text{as } \eta \to 0^+; \tag{B.3}$$

we believe the constant μ, which is determined as part of the solution to (B.2a–c), to be positive; the dependence on A_0 in (B.3) follows from rescaling f_1 by $A_0^{2/(2-n)}$ and η by $A_0^{n/(2(2-n))}$ in (B.2a–c).

The expansion (B.1) breaks down for small η with inner scalings $\eta = |\epsilon|\xi$, $f = |\epsilon|^2 g(\xi)$, and $\frac{\mathrm{d}}{\mathrm{d}\xi}(g_0^n \frac{\mathrm{d}^3 g_0}{\mathrm{d}\xi^3}) = 0$. For $\epsilon < 0$ we thus have

$$g_0 = A_0(\xi_0 - \xi)^2, \quad \xi_0 = \frac{1}{2}\mu A_0^{\frac{n}{2(2-n)}}, \tag{B.4}$$

with $\eta_0 \sim |\epsilon|\xi_0$ and with inner-inner scalings $\eta = \eta_0 + |\epsilon|^{1/(2n-3)}\zeta$, $f = |\epsilon|^{2/(2n-3)} h(\zeta)$,

whereby

$$\frac{1}{2(2-n)}\xi_0 = h_0^{n-1}\frac{\mathrm{d}^3 h_0}{\mathrm{d}\zeta^3},$$
$$\text{as } \zeta \to -\infty, \quad h_0 \sim A_0(-\zeta)^2,$$
$$\text{as } \zeta \to 0^-, \quad h_0 \sim \left(\frac{n^3\xi_0}{6(3-n)(2n-3)(2-n)}(-\zeta)^3\right)^{\frac{1}{n}},$$

producing the desired local behavior (1.2). However, for $\epsilon > 0$ the expression (B.4) is replaced by

$$g_0 = A_0(-\xi_0 - \xi)^2, \quad \xi_0 = \frac{1}{2}\mu A_0^{\frac{n}{2(2-n)}}, \tag{B.5}$$

and (recalling that the interface cannot contract for $n > 3/2$) the inner-inner scalings read $\xi = -\xi_c(\epsilon) + \epsilon^{2(2-n)/(2n-3)}\zeta$, $g = \epsilon^{4(2-n)/(2n-3)}h(\zeta)$, where $\xi_c(0) = \xi_0$ and

$$-\frac{1}{2(2-n)}\xi_0(h_0 - H_\infty) = h_0^n\frac{\mathrm{d}^3 h_0}{\mathrm{d}\zeta^3}, \tag{B.6a}$$
$$\text{as } \zeta \to -\infty, \quad h_0 \sim A_0(-\zeta)^2, \tag{B.6b}$$
$$\text{as } \zeta \to \infty, \quad h_0 \to H_\infty, \tag{B.6c}$$

which determines both h_0 (up to translations in ζ) and H_∞, the decay of h_0 to H_∞ being nonoscillatory. In $-\xi_0 < \xi < 0$ we then have

$$h \sim H_\infty\frac{(-\xi)^2}{\xi_0^2}, \tag{B.7}$$

the left-hand side of (4.2) dominating. The scaling properties of (B.6a–c) show that h_0 and H_∞ scale with $A_0^{2/(2-n)}$ and ζ with $A_0^{n/(2(2-n))}$, so we may rewrite (B.7) as

$$h \sim \nu(n)A_0(-\xi)^2. \tag{B.8}$$

Now setting

$$f = \nu^{\frac{2}{2-n}}\epsilon^{\frac{8}{2n-3}}\widehat{f}, \qquad \eta = \nu^{\frac{n}{2(2-n)}}\epsilon^{\frac{2n}{2n-3}}\widehat{\eta}, \tag{B.9}$$

we have that

$$\widehat{f} \sim A_0(-\widehat{\eta})^2 + \epsilon f_1(\widehat{\eta}),$$

where f_1 satisfies (B.2a–c) with η replaced by $\widehat{\eta}$, implying that the above structure repeats itself on a sequence of finer and finer scales, consistent with the limit cycle behavior referred to in Appendix A. Thus if we denote the variables in the mth member of the sequence by $f^{(m)}$, $\eta^{(m)}$ with $f^{(0)} = f$ and $\eta^{(0)} = \eta$, we have from (B.9) that

$$f^{(m)} \sim \nu^{\frac{2}{2-n}}\epsilon^{\frac{8}{2n-3}}f^{(m-1)}, \qquad \eta^{(m)} \sim \nu^{\frac{n}{2(2-n)}}\epsilon^{\frac{2n}{2n-3}}\eta^{(m-1)},$$

implying that

$$f^{(m)}(\eta) \sim \nu^{\frac{2m}{2-n}}\epsilon^{\frac{8m}{2n-3}}f\left(\frac{\eta}{\nu^{\frac{mn}{2(2-n)}}\epsilon^{\frac{2mn}{2n-3}}}\right),$$

where leading-order expressions for f on the right-hand side are given through a single cycle of the oscillation by (B.1), (B.5), (B.6a–c), (B.8). This is consistent with (A.5) with

$$P \sim \frac{2n}{2n-3} \ln\left(\frac{1}{\epsilon}\right) - \frac{n}{2(2-n)} \ln \nu$$

being large; note that the region described by (B.6a–c) is particularly significant because it leads to the $(-\eta)^{4/n}$-type decay in (A.5), despite the solution behaving quadratically in other regions.

REFERENCES

[1] S. Angenent, *Solutions of the one-dimensional porous-medium equation are determined by their free boundary*, J. London Math. Soc., 42 (1990), pp. 339–353.

[2] J. W. Barrett, J. F. Blowey, and H. Garcke, *Finite element approximation of a fourth order nonlinear degenerate parabolic equation*, Numer. Math., 80 (1998), pp. 525–556.

[3] J. W. Barrett and S. Langdon, *A Multigrid Method for a Fourth Order Elliptic System*, in preparation.

[4] J. Becker and G. Grün, *The thin-film equation: Recent advances and some new perspectives*, J. Phys. Condens. Matter, 17 (2005), pp. S291–S307.

[5] E. Beretta, M. Bertsch, and R. Dal Passo, *Nonnegative solutions of a 4th-order nonlinear degenerate parabolic equation*, Arch. Ration. Mech. Anal., 129 (1995), pp. 175–200.

[6] F. Bernis, *Viscous flows, fourth order nonlinear degenerate parabolic equations and singular elliptic problems*, in Free Boundary Problems: Theory and Applications, Pitman Res. Notes Math. 323, A. L. J. I. Diaz, M. A. Herrero, and J. L. Vazquez, eds., Longman, Harlow, UK, 1995, pp. 40–56.

[7] F. Bernis, *Finite speed of propagation and continuity of the interface for thin viscous flows*, Adv. Differential Equations, 1 (1996), pp. 337–368.

[8] F. Bernis, *Finite speed of propagation for thin viscous flows when* $2 \le n < 3$, C. R. Acad. Sci. Paris Sér. I Math., 322 (1996), pp. 1169–1174.

[9] F. Bernis and A. Friedman, *Higher order nonlinear degenerate parabolic equations*, J. Differential Equations, 83 (1990), pp. 179–206.

[10] A. L. Bertozzi, *Symmetric singularity formation in lubrication-type equations for interface motion*, SIAM J. Appl. Math., 56 (1996), pp. 681–714.

[11] A. L. Bertozzi, *The mathematics of moving contact lines in thin liquid films*, Notices Amer. Math. Soc., 45 (1998), pp. 689–697.

[12] A. L. Bertozzi, M. P. Brenner, T. F. Dupont, and L. P. Kadanoff, *Singularities and similarities in interface flows*, in Trends and Perspectives in Applied Mathematics, Appl. Math. Sci. 100, L. Sirovich, ed., Springer-Verlag, New York, 1994, pp. 155–209.

[13] A. L. Bertozzi, G. Grün, and T. P. Witelski, *Dewetting films: Bifurcations and concentrations*, Nonlinearity, 14 (2001), pp. 1569–1592.

[14] A. L. Bertozzi and M. Pugh, *The lubrication approximation for thin viscous films: Regularity and long time behavior of weak solutions*, Comm. Pure Appl. Math., 49 (1996), pp. 85–123.

[15] J. F. Blowey, J. R. King, and S. Langdon, *Small- and Waiting-Time Behavior of the Thin-Film Equation*, Brunel University Department of Mathematical Sciences Technical Report TR/03/03, West London, UK, 2003.

[16] P. Constantin, T. F. Dupont, R. E. Goldstein, L. P. Kadanoff, M. J. Shelley, and S. Zhou, *Droplet breakup in a model of the Hele-Shaw cell*, Phys. Rev. E, 47 (1993), pp. 4169–4181.

[17] R. Dal Passo, H. Garcke, and G. Grün, *On a fourth-order degenerate parabolic equation: Global entropy estimates, existence, and qualitative behavior of solutions*, SIAM J. Math. Anal., 29 (1998), pp. 321–342.

[18] R. Dal Passo, L. Giacomelli, and G. Grün, *A waiting time phenomenon for thin film equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 30 (2001), pp. 437–463.

[19] E. B. Dussan and S. H. Davis, *On the motion of a fluid-fluid interface along solid surface*, J. Fluid Mech., 65 (1974), pp. 71–95.

[20] R. Fetzer, M. Rauscher, A. Münch, B. A. Wagner, and K. Jacobs, *Slip-controlled thin-film dynamics*, Europhys. Lett., 75 (2006), pp. 638–644.

[21] L. Giacomelli and G. Grün, *Lower bounds on waiting times for degenerate parabolic equations and systems*, Interfaces Free Bound., 8 (2006), pp. 111–129.

[22] H. P. Greenspan, *On the motion of a small viscous droplet that wets a surface*, J. Fluid Mech., 84 (1978), pp. 125–143.
[23] G. Grün, *Droplet spreading under weak slippage: The waiting time phenomenon*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 21 (2004), pp. 255–269.
[24] G. Grün and M. Rumpf, *Nonnegativity preserving convergent schemes for the thin film equation*, Numer. Math., 87 (2000), pp. 113–152.
[25] R. E. Grundy, *Local similarity solutions for the initial-value problem in non-linear diffusion*, IMA J. Appl. Math., 30 (1983), pp. 209–214.
[26] L. M. Hocking, *Sliding and spreading of thin two-dimensional drops*, Quart. J. Mech. Appl. Math., 34 (1981), pp. 37–55.
[27] E. Huh and L. E. Scriven, *Hydrodynamic model of steady movement of a solid/liquid/fluid contact line*, J. Colloid Interface Sci., 35 (1971), pp. 85–101.
[28] J. R. King, *Exact polynomial solutions to some nonlinear diffusion equations*, Phys. D, 64 (1993), pp. 35–65.
[29] J. R. King, *Development of singularities in some moving boundary problems*, European J. Appl. Math., 6 (1995), pp. 491–507.
[30] J. R. King, *Two generalisations of the thin film equation*, Math. Comp. Modelling, 34 (2001), pp. 737–756.
[31] J. R. King and M. Bowen, *Moving boundary problems and non-uniqueness for the thin film equation*, European J. Appl. Math., 12 (2001), pp. 321–356.
[32] A. A. Lacey, *Initial motion of the free-boundary for a non-linear diffusion equation*, IMA J. Appl. Math., 31 (1983), pp. 113–119.
[33] A. A. Lacey, J. R. Ockendon, and A. B. Tayler, *"Waiting-time" solutions of a nonlinear diffusion equation*, SIAM J. Appl. Math., 42 (1982), pp. 1252–1264.
[34] T. G. Myers, *Thin films with high surface tension*, SIAM Rev., 40 (1998), pp. 441–462.